

The Language of Evidence: Robustness and Bias in Code-Switched RAG

Medha Hira

Carnegie Mellon University
mhira@andrew.cmu.edu

Joel Mire

Carnegie Mellon University
jmire@andrew.cmu.edu

Maarten Sap

Carnegie Mellon University
msap2@andrew.cmu.edu

Abstract

Many people code-switch, mixing multiple languages within a single conversation, text or both in everyday communication and increasingly in interactions with AI systems. While prior work has examined how code-switching affects LLM outputs, its impact on RAG systems remains largely unexplored. We present a systematic study of multilingual retrieval under code-switched queries. First, we introduce a retrieval evaluation framework jointly measuring ranking accuracy, language alignment, and bilingual coverage, revealing language preferences that standard metrics obscure. In particular, the strongest relevance-performing retrievers tend to exhibit an English preference. Second, we propose Language-Targeted Contrastive Training to mitigate this retrieval bias, but find a fundamental trade-off: improving alignment reduces exposure to relevant content across languages. Third, controlled RAG experiments show that generation quality depends on retrieval language, with the preferred evidence language varying across query languages. Together, our findings expose tensions between retrieval accuracy, language coverage, and fairness in multilingual RAG, arguing for evaluation protocols that account for query-language coverage.

1 Introduction

The phenomenon of code-switching, i.e., alternating between languages within a conversation or even a single utterance (Poplack, 2013; Doğruöz et al., 2021), is common in multilingual communication and poses important challenges for natural language processing systems (Solorio et al., 2014; Zhang et al., 2023). As LLMs are increasingly used for multilingual interactions, understanding their ability to handle code-switched inputs has emerged as an important research challenge (Zhang

et al., 2023; Oh et al., 2026a). However, despite recent advances in the multilingual LLMs (Üstün et al., 2024; Lai et al., 2024), existing multilingual benchmarks predominantly evaluate monolingual or fully translated inputs, leaving mixed-language settings comparatively underexplored (Hu et al., 2020; Liang et al., 2020). This lack of code-switching in benchmarks prevents the proper assessment of performance and failures for code-switching users (Zhang et al., 2023).

This limitation is especially consequential in *retrieval-augmented generation (RAG)* systems (Lewis et al., 2020), where retrieved evidence directly shapes downstream generation. As illustrated in Figure 1, mixed-language queries can disrupt cross-lingual embedding alignment (Artetxe and Schwenk, 2019) and skew semantic representations toward high-resource languages such as English (Pires et al., 2019; Conneau et al., 2020), causing some retrievers to preferentially surface English documents even when there are equally relevant alternatives (Joshi et al., 2020). This *English anchoring* can reduce coverage of non-English evidence. However, language preferences vary across retrievers, and how the language of retrieved evidence affects downstream generation quality in multilingual RAG, and whether improving language alignment comes at the expense of retrieval quality, remains poorly understood. These questions are central to building robust multilingual RAG systems.

In this work, we ask a central question: *how should retrieval systems balance relevance and language coverage to support robust RAG under code-switched queries?* To answer this, we present a study of multilingual retrieval under code-switched queries, guided by three research questions:

RQ1: How do multilingual retrievers differ in effectiveness and language bias? We eval-

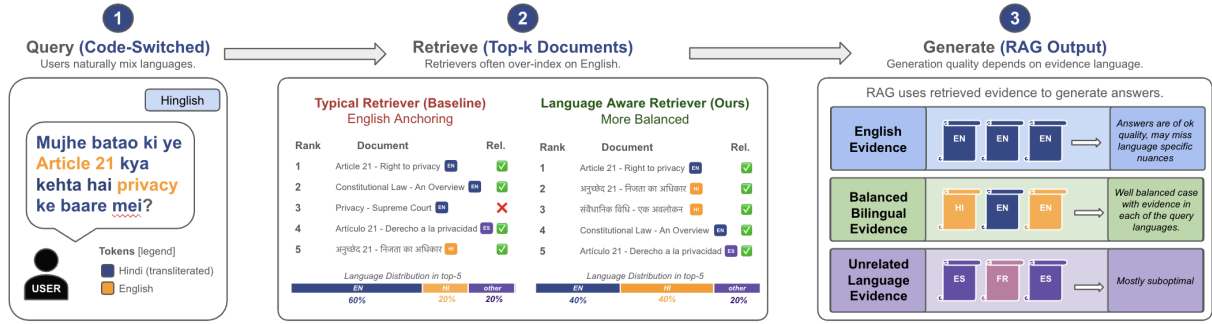


Figure 1: Overview of multilingual retrieval and subsequent generation under code-switched queries.

uate multilingual retrievers under code-switched queries and find that strong performance on standard retrieval metrics, including mean reciprocal rank (MRR) (Voorhees, 1999) and recall, can mask language-skewed retrieval. In particular, the strongest relevance-performing retrievers tend to favor English evidence, while other retrievers favor the query language. These metrics measure if relevant evidence is retrieved but do not consider whether that evidence covers language(s) present in the query. As a result, a retriever can achieve high standard retrieval scores while overlooking relevant evidence in another query language.

RQ2: How do query scripts and romanization affect retrieval, and can induced biases be mitigated? We introduce Language-Targeted Contrastive Training (LTCT) to characterise a trade-off between quality (i.e., document relevance) and language fairness in retrieval. We show that improving query-language alignment comes at a cost to bilingual coverage, highlighting the challenges of jointly optimizing both objectives.

RQ3: How does the language of retrieved evidence affect downstream RAG performance? With retrieval quality held constant, we find that the preferred evidence language varies depending on the query language composition, motivating bilingual retrieval as a robust retrieval policy.

Our results show that how retrieval mass is distributed across languages substantially affects the quality of multilingual retrieval and generation, providing new insight into the limitations of RAG systems for code-switched queries. We release our code.¹

¹medhahira/CodeSwitchRAG

2 Background & Related Work

Code-Switching in NLP Is practice of alternating between two or more languages within a single conversation or utterance, is a well-documented phenomenon in bilingual and multilingual communities (Poplack, 2013; Myers-Scotton, 1993).² Code-switching poses well-documented challenges for multilingual pretrained models such as mBERT and XLM-R (Winata et al., 2021). Recent benchmarks stress-test LLMs directly: Winata et al. (2026) introduce CODEMIXQA and show that models frequently fail on mixed-language inputs, while Oh et al. (2026b) study whether generated outputs align with the language distribution of code-switched prompts. Interestingly, Li et al. (2025) find that enforcing monolingual decoding in bilingual reasoning models reduces MATH500 accuracy by 5.6 points, suggesting that code-switching in reasoning traces can be beneficial. Together, these findings suggest that code-switching can affect different stages of LLM processing, with its effects varying across tasks and domains. We focus on *input-side* code-switching and its implications for multilingual retrieval. Existing work has established that code-switched queries can substantially affect retrieval effectiveness (Zhu et al., 2026; Zeng et al., 2026). However, to our knowledge, it remains unclear how language bias in retrieval affects the coverage of multilingual evidence and, critically, how the language of retrieved evidence shapes downstream generation in RAG pipelines. We address these questions by jointly evaluating retrieval quality, language alignment, and downstream RAG performance.

²Distinct from code-mixing, refers to intra-sentential mixing at the morphological or lexical level, although the terms are often used interchangeably in NLP (Doğruöz et al., 2021).

Cross-Lingual Retrieval, Embedding Bias, and Language Fairness Multilingual dense retrievers like LaBSE (Feng et al., 2022), mContriever (Izacard et al., 2022), multilingual-e5 (Wang et al., 2024) and BGE-M3 (Chen et al., 2024) map languages into shared embedding space, yet exhibit well-documented bias toward English and other high-resource languages (Pires et al., 2019; Conneau et al., 2020; Lawrie et al., 2023).

This bias is further compounded by script variation: romanized text aligns more closely with English embeddings than native-script text, while transliteration significantly degrades dense retrieval performance (Husain et al., 2024; Chari et al., 2025). Prior work shows that both synthetic code-switching augmentation and romanization can partially improve multilingual retrieval robustness (Litschko et al., 2023; Purkayastha et al., 2023). Most closely related to our work, Zeng et al. (2026) introduce CSR-L and CS-MTEB, demonstrating that code-switching substantially degrades retrieval and increases embedding-space divergence across retrievers. We extend this line of work by jointly evaluating retrieval *accuracy* and *language diversity*, studying romanized queries, and connecting retrieval behavior to downstream RAG performance.

This English-skewed embedding geometry poses a fairness concern: certain language communities receive lower-quality retrieval (Ekstrand et al., 2022). Yang et al. (2024) formalize this problem using the Multi-EuP dataset (Yang et al., 2023), while Kirnap et al. (2025) expose intrinsic biases across MLIR technologies. We extend this work to code-switched queries, examining how mixed-language inputs affect retrieval effectiveness and whether retrieved evidence covers the languages present in the query. We introduce metrics for bilingual coverage and retrieval diversity, enabling analysis of the trade-off between retrieval accuracy and language balance.

Multilingual RAG Most RAG research assumes English-centric pipelines (Lewis et al., 2020), but multilingual settings introduce additional retrieval and generation challenges. Prior work shows that cross-lingual evidence can destabilize generation (Chirkova et al., 2024), motivating approaches that either translate retrieved documents into a shared language (Ranaldi et al., 2025) or improve multilingual reranking and corpus selection (Wang et al.,

2026; Lee et al., 2026). We complement this line of work by focusing on *code-switched* queries, examining how language mixing within the query affects which evidence is retrieved, how it is ranked and which languages are represented in the retrieved context.

3 Dataset Construction

To study RAG robustness under code-switched queries, we build a retrieval dataset on top of CODEMIXQA (Winata et al., 2026), a benchmark of $\sim 64k$ naturally code-switched question-answer pairs annotated with $n \geq 1$ source URLs. Although CODEMIXQA provides source URLs for its questions, it was designed for standalone QA rather than retrieval-based evaluation. We therefore construct a multilingual retrieval corpus around it, enabling us to study both retrieval effectiveness and language bias under code-switched queries.

Document corpus We scrape and clean the English text for each unique URL, discarding pages that are inaccessible or empty, yielding $\mathcal{D}_{en}$. Each document is then translated into 11 additional languages via the Google Translate API: Bengali, Spanish, French, Hindi, Indonesian, Italian, Japanese, Korean, Marathi, Urdu, and Chinese (Simplified), resulting in 12 monolingual variants per document. All variants are pooled into a unified multilingual corpus $\mathcal{D} = \bigcup_{\ell \in \mathcal{L}} \mathcal{D}_\ell$, with each document tagged by its language ℓ . This design enables two complementary analyses: (i) retrieval effectiveness on code-switched (English+X) queries, capturing representation quality, and (ii) the language distribution of retrieved evidence across $\mathcal{L}$, capturing induced language bias.

Relevance labels For a query annotated with n source URLs, we treat all $12n$ language variants as relevant, assuming that the translated documents preserve the answer-bearing content of their source documents, consistent with standard practice in cross-lingual retrieval evaluation (Wu et al., 2016; Artetxe and Schwenk, 2019). This relevance definition keeps document content fixed across languages, allowing us to distinguish retrieval of the correct content from the language in which that content is retrieved. Thus, retrieving an English variant for a Hindi-English query is considered relevant while simultaneously providing evidence of

English anchoring. The robustness of our analyses to translation quality is further supported by a back-translation experiment in Appendix B.4, which yields consistent English-anchoring patterns when the English corpus is reconstructed through back-translation.

Retrieval necessity We first verify that retrieved evidence provides meaningful signal for the question-answer instances in our corpus by comparing retrieval-augmented generation (RAG) with zero-shot prompting (ZSP) on Qwen3.5-9B and Llama-3-8B. As shown in Appendix B.1, RAG improves F1 from 0.048 to 0.355 for Qwen3.5-9B and from 0.028 to 0.494 for Llama-3-8B, confirming that retrieved evidence is critical for strong benchmark performance.

4 RQ1: Evaluation Framework and Retrieval Bias

Multilingual retrieval for code-switched queries requires not only ranking relevant documents accurately, but also preserving the query’s language composition, which standard retrieval objectives do not capture. We define the retrieval task (§4.1), introduce our evaluation framework (§4.2), and present the baseline results (§4.3).

4.1 Problem Setup

Consider code-switched query q with languages $\mathcal{L}_q = \{\ell_1, \ell_2\}$. Each document $d \in \mathcal{D}$ (see §3) has a language label ℓ_d . Documents are annotated with query-specific relevance labels and language-alignment labels ($\ell_d \in \mathcal{L}_q$). In our experiments, we focus on the setting where queries mix English with one other language, i.e., $\mathcal{L}_q = \{\ell_{en}, \ell_x\}$.

4.2 Evaluation Protocol

Evaluating retrieval under code-switched queries requires looking beyond standard ranking metrics, which can mask systematic language imbalance in retrieved results. We therefore evaluate retrieval quality along four complementary axes: (i) **relevance**, measuring whether correct documents are retrieved regardless of language; (ii) **language alignment**, capturing the extent to which retrieved documents reflect the languages present in the query; (iii) **language coverage**, quantifying how well systems retrieve relevant documents across

query languages; and (iv) **diversity**, measuring the linguistic spread of retrieved results.

Relevance We evaluate language-agnostic ranking quality using two standard retrieval metrics, mean reciprocal rank (MRR) (Voorhees, 1999) and Recall@ k , computed against query-specific gold relevance labels.

Language Alignment Standard ranking metrics ignore the language of retrieved documents. We introduce **LARS (Language-Aligned Retrieval Score)**, which combines a reciprocal-rank term, rewarding relevant evidence at a high rank, with a language-alignment term that down weights cases where the highest-ranked relevant evidence is not in one of the query languages.

Formally, let $\mathcal{G}_q$ refer to the n gold URLs for query q , each instantiated in $\mathcal{D}$ as 12 monolingual variants, as described in §3. We credit a gold URL as retrieved if any of its monolingual document variants is retrieved. For each $u \in \mathcal{G}_q$, let R_u denote the rank of the highest-ranked retrieved variant of u for query q , and let A_u denote its language-alignment score: $A_u = 1.0$ if that variant is in a query language $\ell \in \mathcal{L}_q$, and $A_u = 0.2$ otherwise. Gold URLs with no retrieved variant within the retrieval depth $K = 2 \times |\mathcal{G}_q|$ receive a score of 0.

$$\text{LARS}(q) = \frac{1}{|\mathcal{G}_q|} \sum_{u \in \mathcal{G}_q} \left(\frac{\alpha}{R_u} + (1 - \alpha) A_u \right) \quad (1)$$

The weighting parameter α balances retrieval rank against language alignment; we set $\alpha = 0.6$. We discuss the choice of α , A_u , the retrieval depth K , and sensitivity to these parameters in Appendix D.

Language Coverage We evaluate how well systems retrieve gold documents across the query languages, where $\mathcal{L}_q = \{en, \ell_x\}$. **Recall_{EN}** measures the fraction of English gold documents retrieved, while **Recall_X** measures coverage of gold documents in ℓ_x . We further define:

- **Joint Coverage**: fraction of queries for which both the English and “X” gold documents are among the top- k results
- **Language Gap**: $\text{Recall}_{EN} - \text{Recall}_X$, measuring directional language bias.

Diversity We measure the linguistic spread of retrieved results using **Language Entropy** $H(L) = -\sum_i p_i \log_2 p_i$, where p_i denotes the proportion

of top- k results in language i across all 12 corpus languages. We use entropy as a measure of the language distribution of retrieved results. **Unique Languages@ K** provides a complementary count-based measure.

4.3 Baseline Retrieval Results

The central finding is that strong ranking performance and balanced language coverage are dissociable: models can score well on MRR while systematically suppressing non-English evidence.

Strong ranking does not imply balanced retrieval. Among baselines (see Table 1), gte-multilingual achieves the highest ranking accuracy, while LaBSE, mContriever, and distiluse exhibit more distributed retrieval, as reflected by $Ent@10$ and $UL@10$. However, greater linguistic diversity is not inherently better; retrieving documents across many unrelated languages can increase diversity without improving the usefulness of the retrieved set, while a less diverse set may provide stronger coverage of the languages present in the query. Thus, standard relevance and diversity metrics capture complementary but incomplete aspects of retrieval quality, motivating the language-aware coverage metrics introduced above.

Strong-performing retrievers exhibit distinct language preferences. Examining the Language Gap in the Gray baseline rows in Table 1 and the top-1 retrieval distributions in Appendix Figure 2, we find that baseline retrievers exhibit distinct forms of language imbalance. Positive Language Gap values indicate greater coverage of English than the query language. Notably, among the stronger-performing multilingual retrievers, multilingual-e5 and distiluse exhibit an English-pivot pattern, with positive gaps and relatively high English-side recall. In contrast, LaBSE and mContriever exhibit negative gaps and favor the query language at the expense of English coverage. bge-large-en also strongly favors English, but this behavior is expected from its English-only training and is therefore not treated as evidence of English anchoring among multilingual retrievers. Overall, the baselines reveal that language imbalance is systematic but not unidirectional; stronger multilingual retrievers can exhibit

English-pivot behavior, while other retrievers favor the query language instead.

No retriever achieves consistent bilingual coverage. Joint Coverage varies widely across baselines, revealing a limitation that is not captured by ranking metrics alone. multilingual-e5 achieves the highest Joint Coverage, but its relatively high English-side recall is accompanied by lower Rec_X . Conversely, LaBSE and mContriever often favor the query language while providing less English coverage. Thus, no baseline consistently retrieves relevant evidence from both languages in a code-switched query, regardless of the direction of its language preference.

5 RQ2: Understanding & Mitigating Language Bias

Having characterized retrieval bias in baseline systems, we next examine how script variation and retriever adaptation interact with these effects. We first analyze the role of script family (§5.1) and romanization (§5.2) in baseline retrieval behavior, and then introduce Language-Targeted Contrastive Training along with its retrieval results (§5.3).

5.1 Script Family Shapes Retrieval Bias

Figure 4 reveals that the direction and magnitude of language imbalance vary by script family. To isolate the role of script, we group queries by writing system: *Indic (non-Latin)* (Hindi and Marathi, Bengali, Urdu), *CJK (Chinese, Japanese, and Korean)*, and *Latin* (Spanish, French, Italian, Indonesian), and measure the average gap $Rec_{EN} - Rec_X$. The non-Latin groups use writing systems that differ substantially from English, while their languages generally have relatively lower representation in multilingual pretraining corpora. We use this grouping to examine whether language imbalance varies systematically across script families and whether language-aware training objectives alter these patterns consistently.

Indic-script queries exhibit the strongest English bias. These queries exhibit the largest English over query-language gaps across all retrievers (Figure 4a). These systems disproportionately retrieve English evidence while under-serving the non-English (X) side of code-switched queries. This pattern occurs across retrievers, indicating that

Retriever	Cond.	Relevance				Diversity		Lang. Align.	Lang. Coverage			
		MRR ↑	Rec@1 ↑	Rec@5 ↑	Rec@10 ↑	Ent@10	UL@10	LARS ↑	Rec _{EN} ↑	Rec _X ↑	Joint ↑	Gap → 0
<i>multilingual-e5</i>												
Baseline		0.898	0.411	0.658	0.749	2.230	5.886	0.599	0.565	0.496	0.590	0.072
LTCT		0.938	0.434	0.794	0.835	<u>0.361</u>	<u>1.615</u>	0.680	0.003	0.817	0.007	−0.814
Abl (no tgt)		0.972	0.460	0.630	0.776	2.820	7.576	0.524	0.156	0.383	0.169	−0.227
Stratified		0.967	0.454	<u>0.854</u>	<u>0.885</u>	0.388	2.163	0.730	0.881	0.067	0.123	0.817
<i>LaBSE</i>												
Baseline		0.591	0.230	0.323	0.382	2.595	6.845	0.256	0.052	0.183	0.087	−0.130
LTCT		0.909	0.411	0.731	0.784	<u>0.543</u>	<u>1.974</u>	0.632	0.008	0.756	0.018	−0.748
Abl (no tgt)		0.964	0.456	0.605	0.742	2.924	8.077	0.518	0.203	0.378	0.250	−0.174
Stratified		0.955	0.444	<u>0.807</u>	<u>0.843</u>	0.194	1.618	0.696	0.838	0.015	0.035	0.826
<i>mContriever</i>												
Baseline		0.484	0.175	0.369	0.453	0.570	2.069	0.323	0.204	0.266	0.101	−0.061
LTCT		0.814	0.357	0.666	0.728	0.192	1.295	0.573	0.004	0.693	0.009	−0.689
Abl (no tgt)		0.958	0.450	0.628	0.798	2.721	7.134	0.531	0.253	0.343	0.238	−0.090
Stratified		<u>0.955</u>	<u>0.444</u>	0.829	0.865	<u>0.207</u>	<u>1.650</u>	0.712	0.861	0.032	0.069	0.832
<i>distiluse-multilingual</i>												
Baseline		0.540	0.202	0.302	0.376	2.673	7.176	0.254	0.199	0.118	0.197	0.082
LTCT		0.806	0.350	0.587	0.652	0.968	2.661	0.503	0.016	0.570	0.035	−0.554
Abl (no tgt)		0.951	0.446	0.591	0.762	2.751	7.252	0.502	0.188	0.317	0.178	−0.129
Stratified		<u>0.937</u>	<u>0.431</u>	0.748	0.791	<u>0.333</u>	<u>2.048</u>	0.656	0.786	0.030	0.066	0.759
<i>bge-large-en-v1.5</i>												
Baseline		0.855	0.385	0.655	0.743	2.073	5.405	0.567	0.520	0.366	0.413	0.156
LTCT		0.704	0.300	0.552	0.603	0.219	1.385	0.474	0.008	0.565	0.016	−0.556
Abl (no tgt)		0.947	0.447	0.670	0.817	2.644	6.809	0.553	0.348	0.295	0.253	0.054
Stratified		<u>0.946</u>	<u>0.445</u>	0.836	0.871	<u>0.348</u>	<u>2.070</u>	0.716	0.866	0.040	0.078	0.829

Table 1: Unified evaluation of retrievers under code-switched queries. We report relevance, language diversity, language alignment (LARS), and language coverage. **Bold** = best per retriever; underline = second-best. ↑ higher is better; ↓ lower is better; Gap evaluated w.r.t. 0 (lower absolute deviation better). Gray = unmodified baseline; Blue = LTCT family (trained); Green = Stratified (trained, alternative objective).

language imbalance is not specific to a particular retriever architecture.

CJK and Latin languages show more variable language imbalance. CJK and Latin languages exhibit smaller and more variable Language Gap values than the Indic group. Across these languages, some retriever-language combinations show little directional skew, while others favor the query language rather than English. Thus, language imbalance is not uniformly English-directed; the strongest and most consistent English skew occurs for Indic-script queries, whereas CJK and Latin queries exhibit a mix of retriever preferences.

Anchoring is not primarily a translation artifact. To test whether English anchoring is driven by the fact that the original English documents are not machine-translated, we replace $\mathcal{D}_{en}$ with English documents obtained by back-translating the corresponding X variants. We then repeat the baseline retrieval experiments while keeping the queries and all other corpus variants unchanged. The observed English-anchoring pattern remains broadly consistent after this intervention, including sim-

ilar query-language recall Rec_X (Appendix B.4, Table 6). This suggests that the observed English preference is not attributable to translation artifacts.

5.2 Romanization Amplifies Retrieval Bias

Romanization introduces language imbalance in multilingual retrieval. Representing non-English queries in Latin script can increase the tendency of retrievers to favor English evidence, even when the underlying query content is unchanged.

Romanized queries increase English concentration in retrieval. Compared with native-script queries, romanized queries retrieve a higher proportion of English documents among the highest-ranked results. This pattern is visible at both top-1 and top-20 retrieval and occurs across multiple retrievers (Appendix Figures 2 and 3). The finding is consistent with prior work showing that romanized text can align more closely with English representations (Husain et al., 2024; Chari et al., 2025).

5.3 Language-Targeted Contrastive Training

To characterize the English-skewed retrieval behavior observed in a subset of baseline retrievers (par-

ticularly the stronger-performing multilingual models §4.3, we introduce **Language-Targeted Contrastive Training** (LTCT), a language-aware training objective that we use to characterize the trade-off between retrieval quality and query-language coverage. We define *anchoring* as the tendency of a retriever to over-weight documents in one language relative to the query-language composition, leading to disproportionate retrieval of evidence from that language in code-switched settings. In this work, we focus primarily on English anchoring because it is the prominent directional pattern among the stronger-performing multilingual base-lines, while the same framework can be applied to other forms of language anchoring. Here, *query-language coverage* refers to whether relevant evidence is retrieved across the languages present in the query, rather than disproportionately favoring one language. LTCT operationalizes this objective through three components: positive sampling, negative sampling, and loss design (pseudocode in Appendix E).

Positive sampling Relevant documents are partitioned into same-language ($\ell_d = \ell_x$) and cross-language groups. We enforce a minimum of one same-language positive (same_lang_positive_min= 1), sampling up to three positives per query, with remaining slots filled from the cross-language pool.

Negative sampling We sample five negatives per query from the non-relevant pool (num_negatives= 5): two hard negatives drawn from non-relevant English documents (hard_negative_ratio= 0.5), one target-language negative ($\ell_d = \ell_x$), and two from the residual pool of documents in any language other than ℓ_x . The single target-language negative prevents collapse to a language-identity shortcut; without it, matching ℓ_x alone would suffice to rank a document highly.

Training objective LTCT combines three complementary inductive biases: same-language positives encourage alignment to the query language; cross-language hard negatives enforce language-aware ranking; and same-language negatives preserve intra-language discrimination. Together, these push the retriever toward surfacing query-language evidence without collapsing to trivial

language-identity shortcuts.

Concretely, the base contrastive loss aggregates:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{lang}} + \lambda_{\text{en}} \mathcal{L}_{\text{en}} + \lambda_{\text{cs}} \mathcal{L}_{\text{cs}} + \lambda_{\text{script}} \mathcal{L}_{\text{script}} \quad (2)$$

where $\mathcal{L}_{\text{lang}}$ is a language-aware contrastive loss (see Appendix E.4 for hyperparameter details), $\mathcal{L}_{\text{en}}$ penalizes retrieval mass allocated to English documents beyond a target proportion $\gamma = 0.25$, $\mathcal{L}_{\text{cs}}$ enforces consistency between original and code-switched query representations, and $\mathcal{L}_{\text{script}}$ aligns native and romanized script representations. We augment $\mathcal{L}_{\text{base}}$ with a target-language mass term that encourages the model to concentrate probability mass on same-language candidates $M = \{i \mid \ell_{d_i} = \ell_q\}$, without requiring explicit relevance:

$$\mathcal{L}_{\text{target-lang}}(q) = -\log\left(\sum_{i \in M} p_i + \epsilon\right) \quad (3)$$

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{target}} \mathcal{L}_{\text{target-lang}} \quad (4)$$

where p_i is the softmax-normalized similarity score for document d_i . Each component of $\mathcal{L}_{\text{base}}$ targets a distinct observed bias: $\mathcal{L}_{\text{en}}$ regularizes against English over-retrieval, $\mathcal{L}_{\text{cs}}$ stabilizes representations under query-level language mixing, and $\mathcal{L}_{\text{script}}$ accounts for the additional retrieval degradation introduced by romanized inputs (§5.2). The target-language mass term then provides a direct training signal toward query-language evidence, complementing these regularizers without overriding the relevance-driven base contrastive objective.

Ablation conditions Three ablation conditions are evaluated: (i) *Targeted* ($\lambda=0.2$, same-language positive); (ii) *No-target-loss* ($\lambda=0$, same-language positive); and (iii) *Stratified* ($\lambda=0$, equal quota from ℓ_q and English, uniform random negatives). Appendix E.4 has more details.

5.4 Retriever Adaptation Results

LTCT improves alignment but concentrates retrieval on the query language. Under the full LTCT objective, LARS improves substantially across all models, confirming that the target-language mass loss effectively steers retrieval toward query-language-aligned documents. However, Joint Coverage collapses to near zero across all models, with Rec_{EN} dropping to ≈ 0.003 across

the board. Thus, while LTCT effectively counteracts English-skewed retrieval, it does so by concentrating retrieval on the query language and suppressing English evidence that may also be useful for downstream generation (see §6).

Ranking accuracy is largely preserved, with one exception. Four of five models improve on standard ranking metrics under LTCT (mean $\uparrow$: +1.6% MRR, +1.8% Recall@10). bge-large-en-v1.5 is the exception (MRR 0.855 $\rightarrow$ 0.704), consistent with the difficulty of adapting an English-only retriever to multilingual settings. Notably, the Stratified condition achieves strong Recall@5 gains across all models, suggesting that balanced bilingual positive sampling alone is sufficient to improve Recall@5 while preserving MRR.

The target-language loss is the primary driver of language concentration. Same-language positive forcing alone (LTCT, no target loss) already meaningfully suppresses English bias and produces the high Joint Coverage among targeted conditions (mContriever: 0.238). Adding $\lambda > 0$ converts these modest gains into near-total language concentration, collapsing Joint to near zero with only marginal additional LARS gains. To reduce English bias while preserving bilingual coverage, positive forcing without the target-language loss offers a better trade-off than the full LTCT.

Stratified training reveals an opposite failure mode. While LTCT collapses English-side recall, Stratified training exhibits the mirror image: although it achieves strong Recall@5 and MRR gains, Rec_{EN} reaches near-ceiling, while Rec_X collapses, indicating that improved ranking performance comes at the expense of coverage of the non-English side of the query. Equal bilingual positive sampling without hard negatives improves retrieval accuracy but favors English evidence rather than achieving balanced coverage of both query languages. Together, LTCT and Stratified illustrate the alignment-coverage trade-off: LTCT over-corrects toward the query language, Stratified favors English, but achieves balanced Joint Coverage.

Training shifts language anchoring but does not eliminate English bias. Appendix Figure 4 shows that the effect of training depends on the training objective. LTCT (panel e) shifts retrieval

Language	D_en	D_x	D_rand
<i>Native Scripts</i>			
Bengali	3.65 $\pm$ 1.42	3.26 $\pm$ 1.32	3.48 $\pm$ 1.37
French	4.06 $\pm$ 1.38	3.80 $\pm$ 1.44	3.36 $\pm$ 1.47
Hindi	3.80 $\pm$ 1.32	3.46 $\pm$ 1.37	3.24 $\pm$ 1.51
Indonesian	3.66 $\pm$ 1.29	3.44 $\pm$ 1.39	3.54 $\pm$ 1.37
Italian	3.71 $\pm$ 1.43	3.80 $\pm$ 1.32	3.20 $\pm$ 1.48
Japanese	3.67 $\pm$ 1.39	3.82 $\pm$ 1.29	3.52 $\pm$ 1.40
Korean	3.58 $\pm$ 1.57	3.80 $\pm$ 1.53	3.42 $\pm$ 1.50
Marathi	3.48 $\pm$ 1.52	3.86 $\pm$ 1.03	3.34 $\pm$ 1.48
Simpl. Chinese	3.23 $\pm$ 1.68	3.80 $\pm$ 1.26	3.43 $\pm$ 1.49
Spanish	3.47 $\pm$ 1.56	3.56 $\pm$ 1.43	3.24 $\pm$ 1.57
Toba Batak	3.58 $\pm$ 1.35	3.50 $\pm$ 1.52	3.16 $\pm$ 1.61
Urdu	3.55 $\pm$ 1.53	3.34 $\pm$ 1.35	3.32 $\pm$ 1.48
<i>Transliterated</i>			
Bengali (translit)	3.86 $\pm$ 1.39	3.14 $\pm$ 1.34	3.57 $\pm$ 1.50
Hindi (translit)	3.98 $\pm$ 1.39	3.38 $\pm$ 1.47	3.52 $\pm$ 1.45
Marathi (translit)	3.66 $\pm$ 1.42	3.49 $\pm$ 1.23	3.72 $\pm$ 1.16
Urdu (translit)	3.64 $\pm$ 1.52	3.10 $\pm$ 1.56	3.86 $\pm$ 1.31

Table 2: LLM-judged RAG quality across languages and corpus settings ([1, 5] scale). **Green** = best per language, **yellow** = second-best, and **red** = worst. Transliterated queries are shown separately.

toward the query language across script families, reducing English anchoring. In contrast, the no-target-language-loss ablation (panel d) retains greater variation across script families, while stratified training (panel f) continues to exhibit strong English anchoring across scripts. These results suggest that improving language alignment requires an explicit language-targeted objective; stratification alone does not reliably resolve English anchoring.

6 RQ3: Impact of Retrieval Language on RAG Performance

Table 2 evaluates how evidence language impacts downstream RAG quality while maintaining retrieval relevance constant. We construct three corpus variants containing the same gold-relevant documents: D_{en} (English evidence), D_x (code-switched-language evidence), and D_{rand} (evidence in an unrelated language). Generation quality is measured using an LLM judge (Appendix G.2).

English evidence dominates most languages. The strongest relevance-performing retrievers in RQ1 tend to exhibit an English preference, and this preference can translate into higher downstream RAG quality for some languages. In Bengali, Hindi, and Urdu, D_{en} outperforms D_x , suggesting that English evidence can be more use-

ful to the generator for these languages. However, this preference is not universal: Japanese, Korean, Marathi, and Chinese perform better with native-language evidence. Thus, the relationship between retrieval preference and generation quality is language-dependent, motivating bilingual retrieval as a robust strategy that preserves access to both English and query-language evidence.

Transliteration amplifies English preference. Transliterated queries show a strong shift toward English evidence (Hindi: 3.98 for D_{en} vs. 3.38 for D_x , compared to a smaller gap in native script). Bengali shows a similar pattern. This suggests transliteration compounds English bias at both retrieval and generation levels, making bilingual retrieval especially important in such settings.

Random-language evidence degrades generation quality. D_{rand} yields the lowest scores on average across languages, indicating that generation quality is sensitive to the language of the retrieved evidence. Together with the zero-shot results, which established the necessity of retrieval (see Appendix B.1), these results show that retrieved evidence provides an important source of language-dependent variation in RAG performance.

7 Discussion & Conclusion

Code-switched queries are an important form of human-AI interaction, yet retrieval systems are evaluated almost exclusively on monolingual inputs. We study multilingual retrieval under code-switched queries in 3 dimensions: we introduce an evaluation framework to rank quality, language alignment, and bilingual coverage; using LTCT as a controlled intervention to expose the alignment-coverage trade-off in RAG, and connecting retrieval behavior to downstream generation quality.

High ranking accuracy masks systematic language bias. Baseline retrieval models achieving strong MRR under code-switched queries do so largely by retrieving English evidence, a behavior invisible to standard metrics but exposed by LARS and our language gap metric. While prior work has documented language bias in multilingual representations (Pires et al., 2019; Conneau et al., 2020) and recent work has identified English preference in code-switched retrieval (Zeng et al., 2026), our findings expand this understanding by showing how

such bias interacts with retrieval evaluation: strong relevance scores can coexist with systematic over-retrieval of English evidence, and the extent of this bias varies with query language and script. This reveals a dimension of retrieval disparity that conventional relevance metrics alone cannot capture, motivating language-aware evaluation alongside standard ranking metrics.

Balancing retrieval quality and query-language fairness remains challenging. Training variations explored reveal a trade-off between retrieval quality, language alignment, and multilingual coverage. While some objectives reduce English anchoring, they can instead concentrate retrieval heavily on the query language, leading to extreme X-anchoring and reduced bilingual coverage. This suggests that language alignment should not be treated as an objective in isolation: effective retrieval for code-switched queries must preserve relevant evidence across both languages while avoiding systematic preference for either one. More broadly, our results highlight joint bilingual coverage as an important and currently underspecified objective for multilingual retrieval training.

Bilingual retrieval is the most robust RAG policy. The optimal evidence language varies across query languages, with some benefiting from English evidence and others from native-language evidence. This heterogeneity aligns with Chirkova et al. (2024) on brittle cross-lingual generation, and partially contradicts Ranaldi et al. (2025), who find translation to English broadly beneficial: our results show this is not universal, particularly for CJK and Indic-script languages. Bilingual retrieval provides a robust hedge that does not require knowing in advance which language the generator handles better for a given query.

Language anchoring is a fairness concern. Code-switched queries receive lower same-language evidence coverage, compounding bias across the retrieval-generation pipeline. While prior work (Yang et al., 2024; Kirnap et al., 2025) highlights such imbalance, our results show that overly aggressive alignment can over-correct by suppressing useful English evidence. Multilingual systems must balance alignment with accessibility so no language in a query is systematically excluded.

8 Limitations

Machine-translated corpus Our corpus is built with the Google Translate API, so translation quality may vary across languages. We test this confound by replacing the English documents with back-translated English and find that the observed retrieval patterns remain broadly consistent (Appendix B.4). This suggests that the observed English anchoring is not primarily due to translation quality, although language-specific translation effects cannot be ruled out entirely.

Scope of Languages Covered We study English+X code-switching across 12 languages, but our framework does not cover three-way code-switching, intra-word mixing, or language pairs that do not include English. Generalizing to these settings remains an open direction.

Generation evaluation RAG performance in §6 is measured using an LLM-based judge. To assess reliability, we performed a small-scale human evaluation on a subset of 260 samples in languages Hindi, Japanese, and Bengali. Found substantial agreement with the automatic judgments (Cohen’s $\kappa = 0.68$). We view the task as relatively straightforward for automatic evaluation, as the judge primarily determines whether generated answers are supported by retrieved evidence rather than requiring nuanced open-ended reasoning.

Reproducibility Our document corpus depends on live URLs scraped from CODEMIXQA and translations obtained via a commercial API. We have released the code and scripts used for document scraping, translation, preprocessing, and relevance-label construction, enabling researchers to reconstruct the corpus.

Scope of LTCT LTCT is evaluated as a post-hoc fine-tuning intervention on existing retrievers. We do not study its interaction with retriever pretraining objectives or its behavior under domain shift.

Alignment trade-offs While LTCT reduces English over-retrieval, it can over-concentrate retrieval mass on the query language and suppress English-side recall. Our ablations suggest that balancing bilingual coverage in a shared embedding space remains challenging: removing the target-language mass loss leaves residual anchor-

ing, while stronger alignment objectives may over-correct toward monolingual retrieval. Future work should explore softer constraints, dynamic training schedules, or explicit bilingual coverage objectives.

Ethical Considerations

This work studies retrieval bias affecting multilingual and code-switching users, a population that includes many speakers of lower-resource languages. We use publicly available benchmark data (CODEMIXQA) and web-scraped documents; no personally identifiable information is collected or retained. Machine translation is used solely to construct a controlled research corpus and is not deployed in any user-facing system. The English anchoring bias documented here represents a fairness concern for non-English-dominant users, and we encourage practitioners to adopt the evaluation framework introduced in §4.2 when auditing multilingual RAG systems for language equity.

Acknowledgments

We acknowledge Carnegie Mellon University and the Sapling Lab for their support and for providing the research environment and resources that made this work possible.

References

- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Adrien Barbaresi. 2021. [Trafilatura: A web scraping library and command-line tool for text discovery and extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.
- Andreas Chari, Iadh Ounis, and Sean MacAvaney. 2025. [Lost in transliteration: Bridging the script gap in neural ir](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’25, page 2900–2905, New York, NY, USA. Association for Computing Machinery.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality,](#)

- multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Nadezhda Chirkova, Benjamin Piwowarski, and Stéphane Clinchant. 2024. [Retrieval-augmented generation in multilingual settings](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. [A survey of code-switching: Linguistic and social perspectives for language technologies](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.
- Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2022. [Fairness in recommender systems](#). In *Recommender Systems Handbook*, pages 679–707, New York, NY. Springer.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Mohammed Safi Ur Rahman Kuriyakodan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2024. [RomanSetu: Efficiently unlocking multilingual capabilities of large language models via romanization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Transactions on Machine Learning Research*, 2022.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Ömer Kirnap and 1 others. 2025. [Language bias in information retrieval: The nature of the beast and mitigation methods](#). *Preprint*, arXiv:2509.06195.
- Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. [LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.
- Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023. [Overview of the TREC 2022 neuclir track](#). In *Proceedings of the 31st Text REtrieval Conference (TREC 2022)*. National Institute of Standards and Technology (NIST).
- Nayeon Lee, Jiwoo Song, and Byeongcheol Kang. 2026. [Coral: Adaptive retrieval loop for culturally-aligned multilingual rag](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 27215–27237.
- 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.
- Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. 2025. [The impact of language mixing on bilingual LLM reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32531–32548, Suzhou, China. Association for Computational Linguistics.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. [XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

- (EMNLP), pages 6008–6018, Online. Association for Computational Linguistics.
- Robert Litschko, Ekaterina Artemova, and Barbara Plank. 2023. [Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11716–11731, Toronto, Canada. Association for Computational Linguistics.
- Carol Myers-Scotton. 1993. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.
- Juhyun Oh, Haneul Yoo, Faiz Ghifari Haznitrana, and Alice Oh. 2026a. [OLA: Output language alignment in code-switched LLM interactions](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46603–46624, San Diego, California, United States. Association for Computational Linguistics.
- Juhyun Oh, Haneul Yoo, Faiz Ghifari Haznitrana, and Alice Oh. 2026b. [Ola: Output language alignment in code-switched llm interactions](#). *Preprint*, arXiv:2601.03589.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Shana Poplack. 2013. “sometimes i’ll start a sentence in spanish y termino en español”: Toward a typology of code-switching. *Linguistics*, 51(s1):11–14.
- Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. [Romanization-based large-scale adaptation of multilingual language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.
- Leonardo Ranaldi, Elena Sofia Ruzzetti, Mirko Lanzi, Giulia Pucci, and Fabio Massimo Zanzotto. 2025. [Multilingual retrieval-augmented generation for knowledge-intensive tasks](#). *Preprint*, arXiv:2504.03616.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. [Overview for the first shared task on language identification in code-switched data](#). In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction fine-tuned open-access multilingual language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- Ellen M. Voorhees. 1999. [The trec-8 question answering track report](#). In *Proceedings of the Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*. National Institute of Standards and Technology (NIST).
- Dan Wang, Guozhao Mo, Yafei Shi, Cheng Zhang, Bo Zheng, Boxi Cao, Xuanang Chen, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. 2026. [All languages matter: Understanding and mitigating language bias in multilingual RAG](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7441–7455, San Diego, California, United States. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual E5 text embeddings: A technical report](#). *Preprint*, arXiv:2402.05672.
- Genta Indra Winata, David Anugraha, Patrick Amadeus Irawan, Anirban Das, Haneul Yoo, Pares Dashore, Shreyas Kulkarni, Ruochen Zhang, Haruki Sakajo, Frederikus Hudi, Anaelia Ovalle, Syrielle Montariol, Felix Gaschi, Michael Anugraha, Rutuj Ravindra Puranik, Zawad Hayat Ahmed, Adril Putra Merin, and Emmanuele Chersoni. 2026. [Can large language models understand, reason about, and generate code-switched text?](#) *Preprint*, arXiv:2601.07153.
- Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yao, Yanxia Xu, Pascale Fung Yu, and Pascale Xu. 2021. [Are multilingual models effective in code-switching?](#) In *Proceedings of the 5th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–109, Online. Association for Computational Linguistics.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016.

Google’s neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024. [Language fairness in multilingual information retrieval](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2487–2491, Washington, DC, USA. ACM.

Jinrui Yang, Timothy Baldwin, and Trevor Cohn. 2023. [Multi-EuP: The multilingual european parliament dataset for analysis of bias in information retrieval](#). In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 282–291, Singapore. Association for Computational Linguistics.

Qingcheng Zeng, Yuheng Lu, Zeqi Zhou, Heli Qi, Puxuan Yu, Fuheng Zhao, Hitomi Yanaka, Weihao Xuan, and Naoto Yokoya. 2026. [Code-switching information retrieval: Benchmarks, analysis, and the limits of current retrievers](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13055–13071, San Diego, California, United States. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. [Multilingual large language models are not \(yet\) code-switchers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Tongyao Zhu, Huang Chao Ming, and Min-Yen Kan. 2026. [When does mixing help? analyzing query embedding interpolation in multilingual dense retrieval](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31544–31562, San Diego, California, United States. Association for Computational Linguistics.

A Analysis of the Document Corpus

Language	Code	Avg. Words/File	Avg. Tokens/File
Japanese	ja	1253.89	6598.38
Chinese (Simplified)	zh-cn	1286.64	5868.55
Korean	ko	2372.63	5731.36
Marathi	mr	2677.59	6143.83
Bengali	bn	2667.22	6658.03
Hindi	hi	2967.17	5875.35
Indonesian	id	2992.01	5479.23
Urdu	ur	3187.55	5083.35
English	en	3218.63	6048.57
Italian	it	3269.30	6054.35
French	fr	3348.06	6206.22
Spanish	es	3433.22	5997.69

Table 3: Average number of words and tokens per file by language (sorted by word count)

Since all documents are parallel translations of the same underlying texts, variation in average

word counts primarily reflects linguistic and structural differences across languages rather than differences in content. Languages such as Chinese and Japanese exhibit substantially lower word counts due to their high information density and logographic writing systems, where meaning is encoded more compactly at the character level. In contrast, Romance languages such as Spanish, French, and Italian show higher word counts, likely driven by richer morphological agreement and more explicit syntactic realization during translation. Languages including Hindi, Urdu, Bengali, and Marathi occupy an intermediate range, reflecting a balance between morphological richness and syntactic compression.

To further analyze these differences in a model-relevant setting, we compute token counts using the **XLM-RoBERTa tokenizer** (Conneau et al., 2020)³. While tokenization reduces some of the extreme disparities observed in word-level statistics, notable variation persists across languages due to differences in subword segmentation efficiency. Languages with non-Latin scripts, such as Bengali, Japanese, and Chinese, tend to exhibit higher token counts relative to their word counts, reflecting less efficient subword decomposition. Conversely, languages such as Urdu and Indonesian show comparatively lower token usage, indicating more compact segmentation under the shared multilingual vocabulary.

Overall, these results suggest that both word- and token-level statistics are inherently language-dependent surface measures and are not directly comparable across languages. Instead, they should be interpreted as artifacts of linguistic structure and tokenizer behavior rather than indicators of semantic content. This highlights that naive length-based statistics introduce systematic cross-lingual bias and should not be used as comparable proxies for semantic content in multilingual corpora.

B Dataset Curation

B.1 Retrieval Necessity Validation

Before studying multilingual retrieval bias, we first verify that retrieval provides meaningful signal for the task. We compare retrieval-augmented generation (RAG) against zero-shot prompting (ZSP),

³xlm-roberta-base

Model	Mode	EM	F1
Qwen3.5-9B	ZSP	0.009	0.048
Qwen3.5-9B	RAG	0.099	0.355
Llama-3-8B	ZSP	0.003	0.028
Llama-3-8B	RAG	0.252	0.494

Table 4: Zero-shot prompting (ZSP) vs. retrieval-augmented generation (RAG) on English-translated queries with $\mathcal{D}_{en}$. Both models use greedy decoding with a five-shot prompt format (Appendix B.2). Retrieval substantially improves performance across all metrics.

where the model answers directly from parametric knowledge without retrieved context.

Table 4 shows that RAG substantially outperforms ZSP across all evaluated models. Qwen3.5-9B improves from F1 0.048 $\rightarrow$ 0.355, while Llama-3-8B improves from 0.028 $\rightarrow$ 0.494. These results confirm that parametric knowledge alone is insufficient for the benchmark and that access to retrieved evidence is critical for strong performance.

B.2 Prompt Templates

We use separate prompts for zero-shot prompting (ZSP) and retrieval-augmented generation (RAG). Both prompts enforce short extractive outputs to ensure compatibility with exact match (EM) and token-level F1 evaluation.

Zero-Shot Prompting (ZSP). The following template is used when the model answers directly from parametric knowledge without retrieved evidence.

Zero-Shot Prompt Template
You are a question answering system. CRITICAL RULES – follow exactly: 1. Output ONLY the answer, nothing else. No explanation, no sentences, no punctuation at the end. 2. Answer using only a single word or short phrase (1-5 words maximum). 3. Do NOT output any commentary, code, or follow-up questions. 4. If the answer is unknown, output exactly: unknown Question: {problem} Answer:

Retrieval-Augmented Generation (RAG). The following template is used when retrieved context documents are provided to the model.

RAG Prompt Template
You are a question answering system. Your task is to extract the exact answer from the provided context. CRITICAL RULES – follow exactly: 1. Output ONLY the answer, nothing else. No explanation, no sentences, no punctuation at the end unless it is part of the answer itself. 2. Copy the shortest span from the context that directly answers the question. 3. Do NOT output "(from Doc X)", "Note:", "#", code blocks, or any commentary. 4. If the answer is not in the context, output exactly: unknown 5. Output must be 1-5 words maximum. Context: {chunk} Question: {problem} Answer:

All experiments use greedy decoding (`do_sample=False`) to ensure deterministic outputs across retrieval conditions.

B.3 Corpus Statistics and Translation Validation

Additional corpus statistics and translation-quality analyses are provided here. We validate that machine-translated documents preserve answer-bearing content sufficiently well for multilingual retrieval evaluation. Translation quality is assessed using COMET-based evaluation and manual inspection of answer extractability across translated document variants.

Corpus construction details. We scraped the English text for each unique URL in CODEMIXQA, retaining pages that returned a successful HTTP response and contained non-empty text after boilerplate removal using *trafilatura* (Barbarese, 2021). The resulting English corpus $\mathcal{D}_{en}$ was translated into the remaining languages using the Google Translate API, producing parallel multilingual document collections.

Table 5 summarizes corpus statistics.

Statistic	Value
Raw unique URLs (EN)	2936
Retained documents (EN)	2378
Languages	12
Total multilingual documents	28536

Table 5: Summary statistics for the multilingual retrieval corpus.

Metric	Equivalent cells
Rec_X (query language)	55 / 55
Recall@10	54 / 55
GDAS	54 / 55
LARS	53 / 55
Joint coverage	52 / 55
Rec_{EN} / Gap	39 / 55

Table 6: Back-translation retrieval ablation across 5 retrievers $\times$ 11 query languages. Values report the number of cells equivalent to the clean-English baseline under TOST (± 0.02). Query-language recall Rec_X is equivalent in all cells, while the observed English anchoring remains broadly consistent after back-translation.

Translation quality validation. A key assumption underlying the corpus is that machine-translated documents preserve answer-bearing content. To validate this assumption, we randomly sampled translated documents in Hindi, Japanese, and Bengali (representing Devanagari, CJK, and Indic scripts respectively) and evaluated translation quality using COMET-22 (Rei et al., 2022).

Mean COMET scores of 0.84, 0.86, and 0.81 respectively indicate that translated documents largely preserve semantic content relevant for retrieval and downstream QA. We additionally verified on a manually inspected sample of 260 QA pairs that gold answers remain extractable from translated documents in the vast majority of cases.

Translation quality is not perfectly uniform across languages. Rather than leaving this as a caveat, we quantify it directly and then test experimentally whether it accounts for the English-side advantages we report; see Appendix B.4. We discuss the residual limitation in §8. Further, we have done some analysis to prove the effectiveness of our translation mechanisms by running experiments on back translated English documents which seconds our translation quality (see Appendix B.4).

B.4 Back-Translation Robustness Check

To test whether the observed English anchoring is an artifact of translation quality, we back-translate each translated document from X to English, producing an English corpus with translation artifacts comparable to those introduced in the forward translations. We then repeat the baseline retrieval experiments from Table 1 on these back-translated documents. The English-anchoring pattern (bias towards retrieval of English documents) remains consistent with the original English corpus, with retrieval results equivalent in 54/55 **language** $\times$ **retriever** cells (where languages $\in \{\text{bn, es, fr, hi, id, it, ja, ko, mr, ur, zh}\}$) under a threshold of ± 0.03 and query-language recall equivalent in 55/55 cells, suggesting that the observed bias is not primarily attributable to translation quality.

Retrieval results. Across the nine relevance metrics, 482 of 495 cells are equivalent (Table 6b). The 13 failures are one-sided: eight are distiluse under the Urdu pivot and all are positive ($+0.040$ to $+0.051$), meaning that back-translated English is retrieved better than the original English. The remaining five are mContriever and occur only at hit@1 and RR. Query-language recall Rec_X is equivalent in 55/55 cells, with every mean delta below 0.005; changes in the language gap therefore arise solely from changes in Rec_{EN} .

These results suggest that the observed English anchoring is not primarily attributable to translation quality. The language-gap analysis is consistent with this conclusion: the Indic group retains the largest gap in all conditions corresponding to the combinations of retriever and query language evaluated for the English-anchoring analysis for the English anchoring retrievers, while the relative ordering across script families remains unchanged under the back-translated corpus.

C English Bias Anchoring in Retrieval (Baseline Models)

Below are heatmaps analyzing the query (X in EN+ X) vs document in $k = 1$ (Figure 2) and $k = 20$ (Figure 3) setting.

D Motivating LARS: Design Choices and Sensitivity Analysis

D.1 Why a Composite Metric?

Standard retrieval metrics such as MRR and Recall@ k measure whether relevant documents are returned, but are blind to *which language* those documents are in. In a code-switched RAG setting where the query mixes English and language ℓ_x , a retriever that consistently returns rank-1 English documents achieves strong MRR while systematically under-serving the non-English side of the query. LARS is designed to expose this failure by jointly penalizing (i) poor ranking and (ii) language misalignment within a single score that preserves comparability across systems.

D.2 LARS Parameter Choices

Choice of $\alpha = 0.6$. We set $\alpha = 0.6$, giving slightly greater weight to ranking than language alignment. The equal-weight setting ($\alpha = 0.5$) produces greater asymmetry between the English-pivot and LTCT-collapsed/query-language-pivot failure modes, while the max-min analysis identifies $\alpha \approx 0.74$ as the point where their LARS gaps are equalized. We therefore use $\alpha = 0.6$ as a compromise between these settings, retaining near-equal weighting while modestly prioritizing ranking.

Choice of $A_u = 0.2$. Non-query-language documents remain relevant under our labeling scheme (§3) and should not receive a score of zero; they retain retrieval value despite being language-misaligned. We therefore set $A_u = 0.2$, assigning one-fifth of the aligned score to non-query-language documents. This reflects a substantial penalty for language misalignment without treating such documents as irrelevant. The English anchoring pattern remains qualitatively consistent across $A_u \in [0.1, 0.4]$, with retriever rankings preserved.

D.3 Limitations of LARS

LARS collapses ranking and alignment into a single number, which can obscure the direction of failure. A system with moderate ranking and poor alignment can score similarly to one with strong ranking but no alignment. We therefore recommend reporting LARS alongside Rec_{EN}, Rec_X, and

the language gap as a diagnostic suite rather than as a standalone summary statistic.

E Training Details

Algorithm 1 gives the pseudocode for the full LTCT training procedure, and Algorithm 2 for the Stratified ablation condition.

E.1 Language-Aware Loss Decomposition and Hyperparameters

This section details the components of the base contrastive objective and their associated hyperparameters.

Language-aware contrastive loss. The term $\mathcal{L}_{\text{lang}}$ is a language-conditioned contrastive objective that differentiates between same-language and cross-language positives. It assigns a higher weight α to same-language positive pairs and a lower weight β to cross-language positive pairs:

$$\alpha = 1.8, \quad \beta = 0.8.$$

Let q be a query and $\mathcal{P}_q$ its positive documents. Each positive $d \in \mathcal{P}_q$ is assigned:

$$w(d, q) = \begin{cases} \alpha & \text{if } \ell_d = \ell_q \\ \beta & \text{otherwise} \end{cases}$$

The language-aware contrastive loss is:

$$\mathcal{L}_{\text{lang}}(q) = -\log \frac{\sum_{d \in \mathcal{P}_q} w(d, q) \exp(s(q, d))}{\sum_{d' \in \mathcal{D}} \exp(s(q, d'))}$$

This induces a controlled language bias in the embedding space, encouraging stronger clustering of documents sharing the query language while preserving cross-lingual alignment.

English mass regularization. Let $\mathcal{D}_{\text{en}}$ denote English documents. The loss penalizes excessive probability mass assigned to English documents beyond a target threshold γ :

$$\gamma = 0.25.$$

Define:

$$p(d|q) = \frac{\exp(s(q, d))}{\sum_{d' \in \mathcal{D}} \exp(s(q, d'))}$$

$$\mathcal{L}_{\text{en}}(q) = \max \left(0, \sum_{d \in \mathcal{D}_{\text{en}}} p(d|q) - \gamma \right)$$

This prevents English dominance in multilingual retrieval and stabilizes ranking distributions.

Code-switch consistency. $\mathcal{L}_{cs}$ enforces consistency between representations of original and code-switched queries, improving robustness to mixed-language inputs.

Script alignment. $\mathcal{L}_{script}$ aligns native-script and romanized-script representations of the same content.

Target-language mass objective. An additional term encourages concentration of probability mass over documents in the same language as the query:

$$\lambda_{target} = 0.2.$$

This term promotes language-consistent retrieval without requiring explicit relevance annotations.

Hyperparameters summary.

$$\alpha = 1.8, \quad \beta = 0.8, \quad \gamma = 0.25$$

E.2 Language-Targeted Contrastive Training

Algorithm 1 Language-Targeted Contrastive Training

```

1: for each query  $q$  do
2:   Collect relevant documents  $\mathcal{P}(q)$ 
3:   Partition  $\mathcal{P}(q)$  into:
      $\mathcal{P}_{same} \leftarrow \{d \in \mathcal{P}(q) \mid \ell_d = \ell_x\}$ 
      $\mathcal{P}_{cross} \leftarrow \mathcal{P}(q) \setminus \mathcal{P}_{same}$ 
4:   Sample positives  $\mathcal{P}^+$  with minimum same-language
     constraint
5:   Sample negatives  $\mathcal{N}$  using language-aware hard sam-
     pling
6:   Compute similarity scores  $\{s_i\}$ 
7:   Compute losses:
      $\mathcal{L}_{base}$ 
      $\mathcal{L}_{target-lang}$ 
8:   Update model parameters using combined loss
9: end for

```

E.3 Stratified Training Algorithm (Ablation)

Algorithm 2 Stratified Contrastive Training

```

1: for each query  $q$  with language  $\ell_q$  do
2:   Collect relevant documents  $\mathcal{P}(q)$ 
3:   Partition into language buckets:
      $\mathcal{P}_{\ell_q} \leftarrow \{d \in \mathcal{P}(q) \mid \ell_d = \ell_x\}$ 
      $\mathcal{P}_{en} \leftarrow \{d \in \mathcal{P}(q) \mid \ell_d = en\}$ 
4:   Sample positives  $\mathcal{P}^+$ : draw  $\lfloor k/2 \rfloor$  from each bucket;
     top-up from either if one is small
5:   Sample negatives  $\mathcal{N}$ : uniform random from  $\mathcal{D} \setminus \mathcal{P}(q)$ 
     across all languages,  $|\mathcal{N}| = |\mathcal{P}^+|$ 
6:   Compute loss  $\mathcal{L} = \mathcal{L}_{InfoNCE}(\mathcal{P}^+, \mathcal{N})$  ( $\lambda_{target} = 0$ )
7:   Update model parameters
8: end for

```

E.4 Hyperparameters for training

Fixed hyperparameters. All reported runs use $\alpha = 1.8$, $\beta = 0.8$, $\gamma = 0.25$, $\lambda_{target} = 0.2$, temperature $\tau = 0.05$, `hard_negative_ratio`= 0.5, and `same_lang_positive_min`= 1. These are fixed rather than swept to isolate the structural contribution of each component across ablation conditions. We train with AdamW and gradient clipping.

F Analysis of Top-k Documents Under Different Training Paradigms

In this section, we analyze the retrievers under different training paradigms. We plot heatmaps to show the average top- k language gaps in the percentage of English vs X documents retrieved for a code-switched query which combines English and another language.

G Prompt Templates and Experimental Implementation for RQ3

We implement a lightweight multilingual retrieval-augmented generation (RAG) evaluation framework using Meta-Llama-3-8B-Instruct as both the answer generator and the LLM judge is GPT-4o. For each query, documents are retrieved under different language conditions (native-language, English, and random-language retrieval), concatenated into a context window, and passed to the model for generation. Generated answers are subsequently evaluated by the same model using an LLM-as-a-judge setup.

G.1 RAG Answer Generation Prompt

The following prompt is used for retrieval-augmented answer generation when supporting documents are available.

RAG Generation Prompt

```

Use only the documents below to answer the
question.
Be concise – one sentence or a few words.
Respond in the same language as the question
(language code: {lang_code}).

```

```
{context}
```

```

Question: {question}
Answer:

```

When no retrieved documents are available, we instead use the following fallback prompt:

Fallback Prompt (No Retrieved Context)

Answer the following question briefly
(a few words or one sentence).
Respond in the same language as the question.

Question: {question}
Answer:

G.2 LLM-as-a-Judge Prompt

To evaluate answer quality, we prompt the same model to act as an expert multilingual QA evaluator. The judge assigns a score from 1–5 and provides a short rationale.

LLM Judge Prompt

You are an expert evaluator for multilingual QA systems.

Score the model's answer on a scale of 1 to 5:
5 = fully correct and complete
4 = mostly correct, minor gap
3 = partially correct
2 = barely relevant
1 = wrong or hallucinated

Question: {question}
Reference answer: {reference}

Model answer: {answer}

Return exactly one JSON object and nothing else

Output schema:
{
 "score": <int 1-5>,
 "rationale": "<one sentence>"
}

G.3 Experimental Conditions

For each query, we evaluate retrieval under up to three language conditions:

- **Native-language retrieval:** documents retrieved in the query language.
- **English retrieval:** documents retrieved only from English corpora.
- **Random-language retrieval:** documents retrieved from a randomly selected non-English language distinct from the query language.

Each retrieved document is truncated to at most 1200 characters before concatenation into the prompt context. Generation uses greedy decoding (`do_sample=False`) for reproducibility.

G.4 Generator Robustness for RQ3

To assess whether the RQ3 findings depend on the choice of generator, we repeat the experiment using Qwen-2.5-7B as the generator on a representative subset of 50 examples per retrieval condition. The retrieval setup and evidence-language conditions are identical to those in §6.

The qualitative conclusions remain consistent with those reported in the main paper. In aggregate, English evidence achieves the highest average judge score (3.15), followed by native-language evidence (3.05), with unrelated-language evidence performing worst (2.84). At the individual-language level, the preferred evidence language remains language-dependent rather than universal: for example, Bengali and Hindi favor English evidence, whereas Indonesian, Italian, and Simplified Chinese favor native-language evidence.

These results suggest that the observed language-dependent generation effects are not specific to a single generator, although broader evaluation across multilingual generators remains an important direction for future work.

G.5 Implementation Details

The implementation is based on the transformers library. Models are loaded in either FP16 or 4-bit quantized mode depending on available hardware. We use a maximum generation budget of 64 new tokens for answer generation and 128 new tokens for judge outputs.

H Disclosure of AI Assistance

The artificial intelligence assistants (in specific: ChatGPT and Gemini) were used to support grammar correction, writing clarity, and the general editing of the paper. AI tools were also used for limited coding assistance, including basic script writing and debugging. All of the decisions about experimental design, analysis, and final content were made and verified by us (the authors).

I Artifact Usage and Intended Use

I.1 Use of Existing Artifacts

CODEMIXQA. We use CODEMIXQA (Winata et al., 2026) as the source of code-switched question-answer pairs and associated URLs. Our use is consistent with its intended purpose as a

research benchmark for multilingual and code-switched NLP evaluation. Our derivative dataset is intended solely for non-commercial research use.

Web-scraped documents. We scrape English documents from URLs provided in CODEMIXQA using *trafilatura* (Barbaresi, 2021). Scraped content is used exclusively for constructing a controlled research corpus and is not deployed in any user-facing system. To respect the terms of source websites, we will release **only document URLs and relevance labels** rather than raw scraped text, following standard practice in information retrieval dataset releases (Lewis et al., 2020).

Google Translate API. Translations are generated via the Google Translate API solely for research corpus construction. Consistent with the API terms of service, translated documents will **not be directly redistributed**. Instead, we will release the translation pipeline and language labels to allow reproduction of the corpus.

Meta-Llama-3-8B-Instruct. This model is used as both the answer generator and LLM judge under Meta’s community license, which permits research use. Our use is consistent with its intended research purpose and does not involve fine-tuning or redistribution of model weights.

I.2 Artifacts Created in This Work

The artifacts released by this work: including the evaluation code, the LARS metric implementation, document relevance labels, retrieval model checkpoints, and the LTCT training pipeline are all intended for **research use only**. They should not be used in production systems or commercial applications without further evaluation, particularly given the fairness and coverage limitations documented in this paper (see §8).

We have released all artifacts under a license, with the exception of model checkpoints, which inherit the license terms of their respective base models.

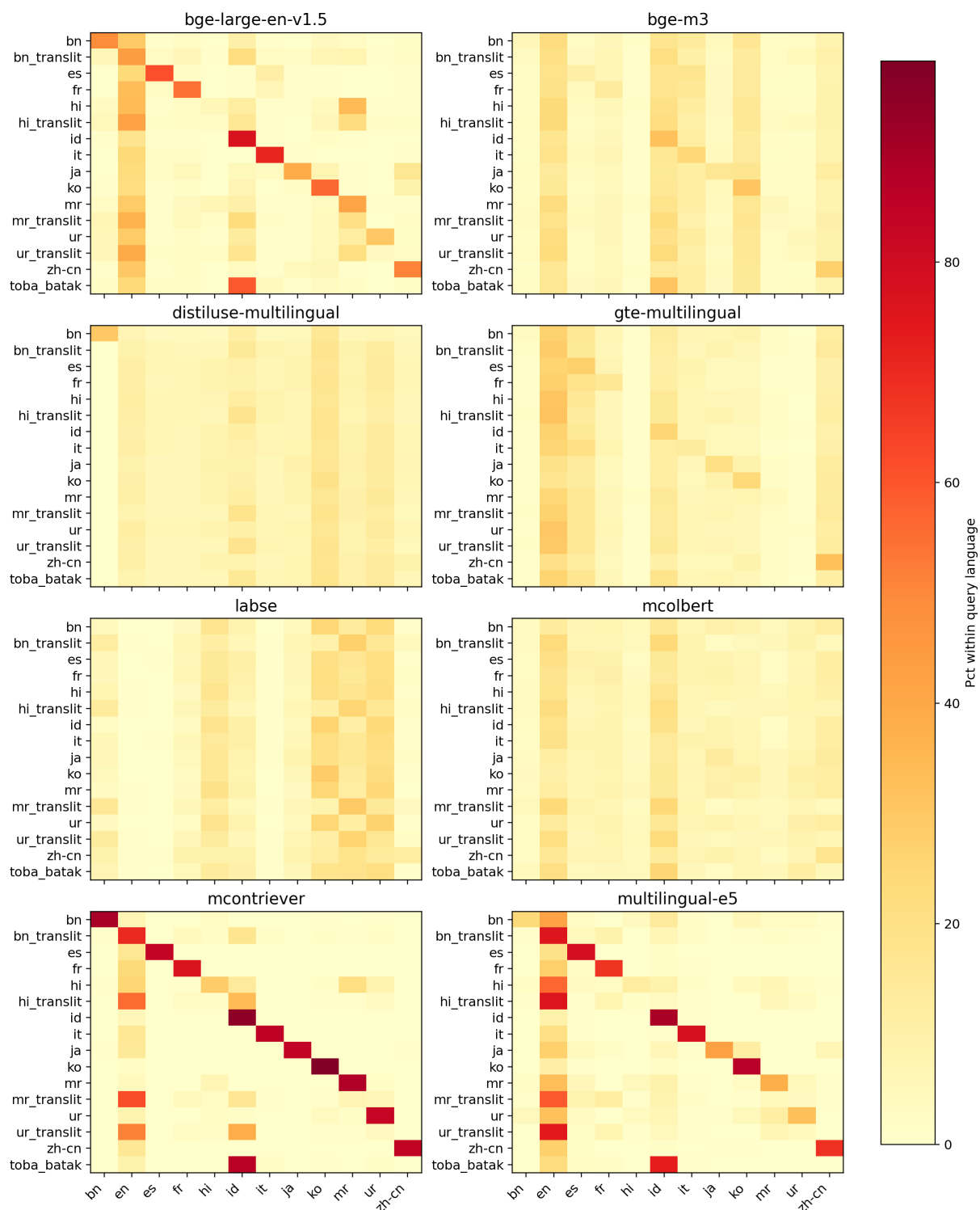


Figure 2: Query-document language distribution at $k=1$ across retrievers. Rows correspond to query languages (including code-switched variants) and columns to document languages. Values indicate the percentage of retrieved documents per language. English-pivot models show strong concentration in the English column, while alignment-based models exhibit more distributed cross-lingual retrieval.

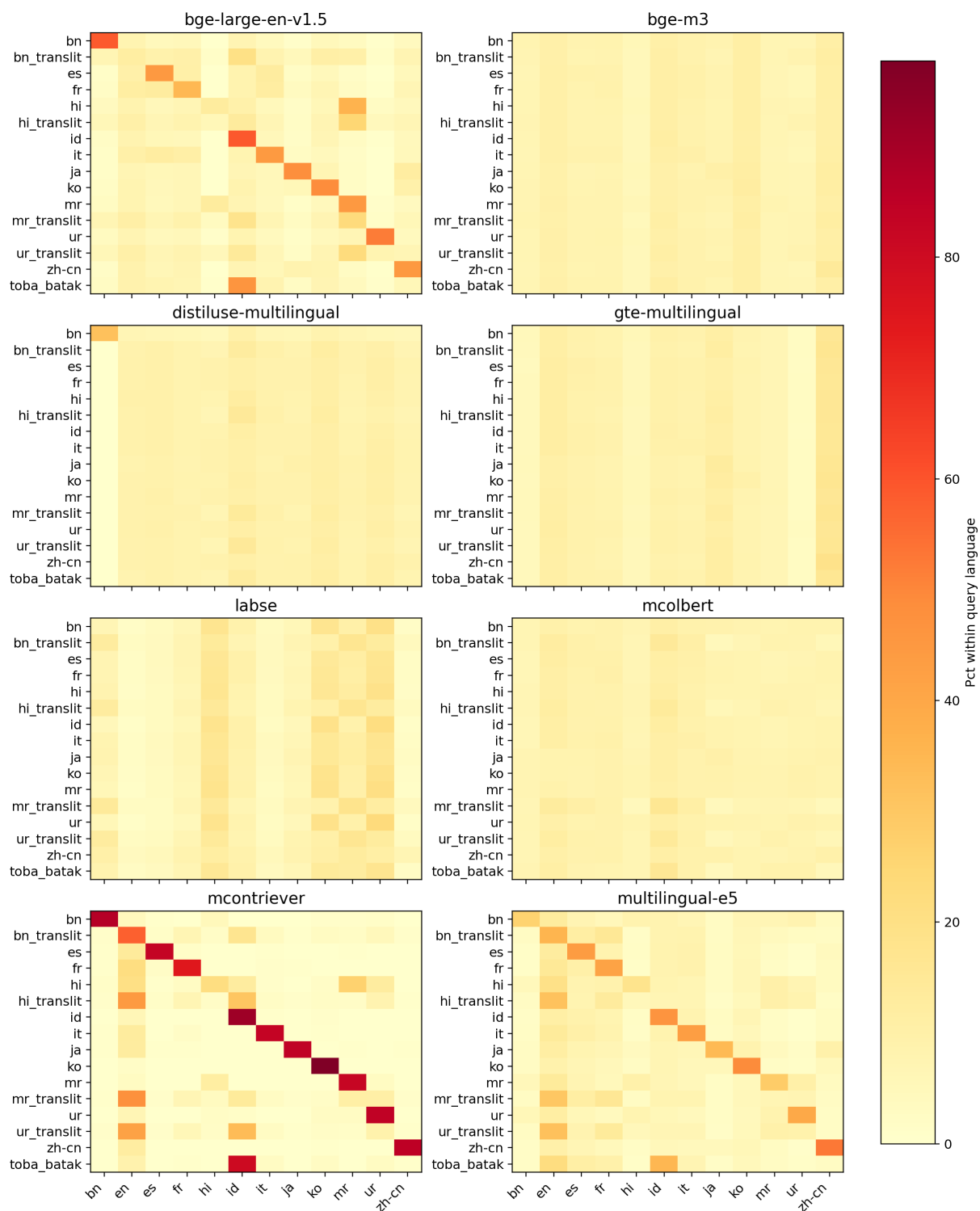
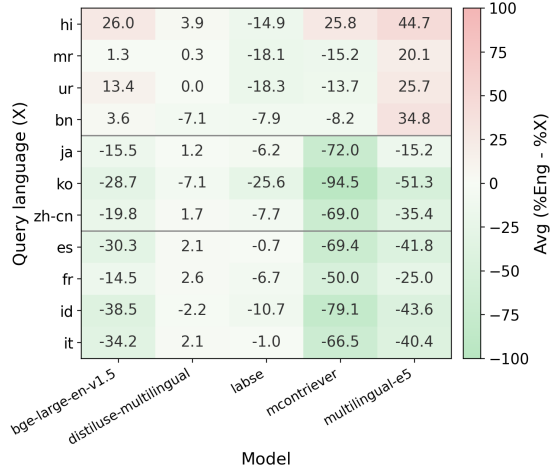
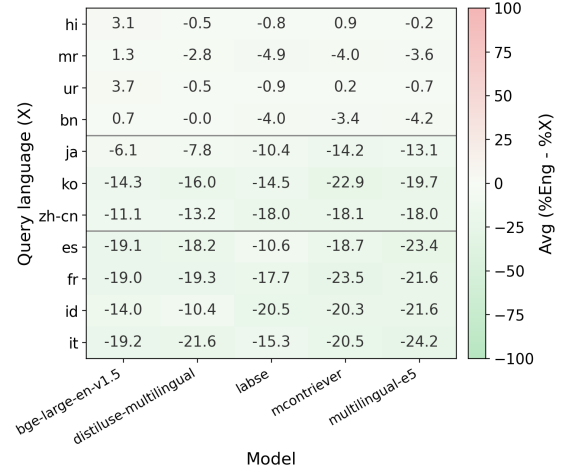


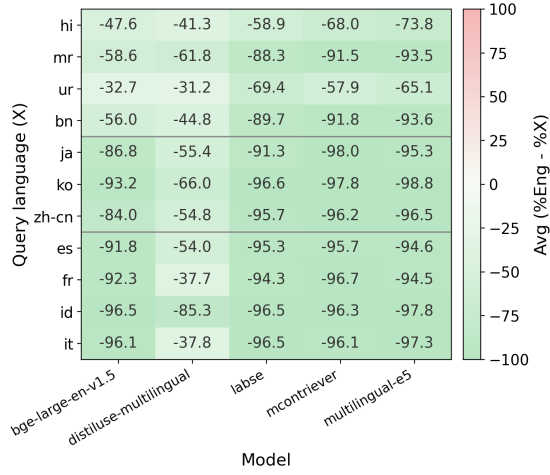
Figure 3: Query-document language distribution at $k = 20$. Compared to Figure 2, retrieval becomes more diverse, with reduced English dominance and increased coverage of non-target languages, indicating that cross-lingual matches emerge at deeper ranks.



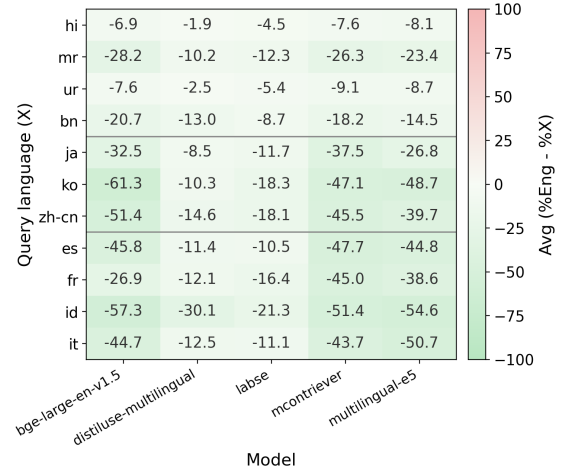
(a) Baseline retriever: average language gap in top- k retrieval, where gap is $\%Eng - \%X$.



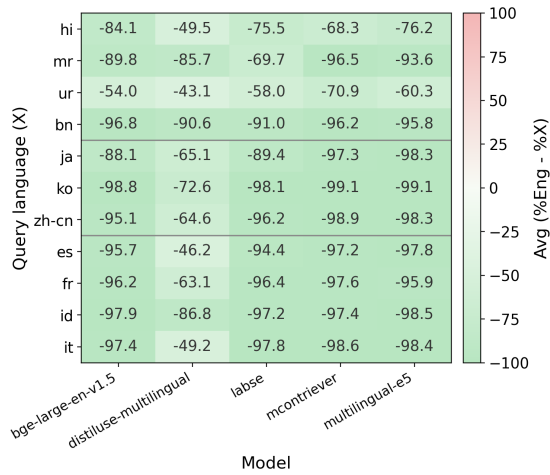
(b) Ablation (no target-language loss): average top- k language gap $\%Eng - \%X$.



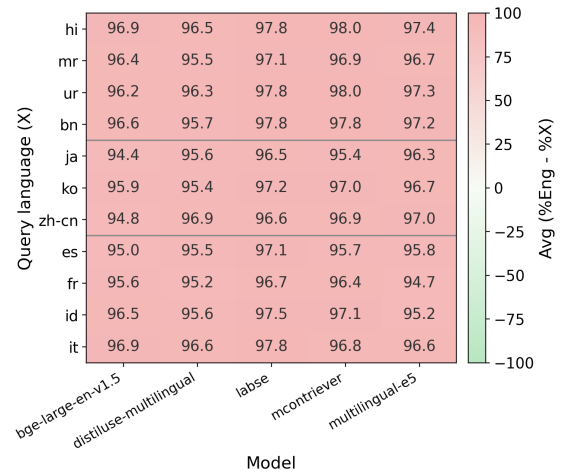
(c) Ablation (random position strategy): average top- k language gap $\%Eng - \%X$.



(d) Targeted training without target-language loss: average top- k language gap $\%Eng - \%X$.



(e) Trained retriever setting: average top- k language gap $\%Eng - \%X$.



(f) Stratified training setting: average top- k language gap $\%Eng - \%X$.

Figure 4: Cross-setting language preference in retrieval: average $\%Eng - \%X$ among the top- k documents (k = number of gold documents per query; X = annotated query language). Red (+) = English-skewed, green (-) = X -skewed. Rows are grouped by script family.