Code-Switching Information Retrieval: Benchmarks, Analysis, and the Limits of Current Retrievers

Qingcheng Zeng[1*], **Yuheng Lu**[2*], **Zeqi Zhou**[3], **Heli Qi**[2,4], **Puxuan Yu**[5], **Fuheng Zhao**[6], **Hitomi Yanaka**[7,4], **Weihao Xuan**[7,4†], **Naoto Yokoya**[7,4]
[1]Northwestern University, [2]Waseda University, [3]Brown University
[4] RIKEN AIP, [5]Snowflake Inc., [6]University of Utah, [7]The University of Tokyo

Abstract

Code-switching is a pervasive linguistic phenomenon in global communication, yet modern information retrieval systems remain predominantly designed for, and evaluated within, monolingual contexts. To bridge this gap, we present a holistic study of code-switching IR. We introduce the **C**ode-**S**witching **R**etrieval benchmark-**L**ite (**CSR-L**), a human-annotated benchmark designed to capture natural mixed-language queries, and evaluate statistical, dense, cross-encoder, and late-interaction retrieval methods on it. The results show that code-switching is a persistent performance bottleneck, degrading even strong multilingual models. We further show that this failure is associated with substantial divergence between monolingual and code-switched query embeddings. To test whether the pattern generalizes beyond retrieval, we construct **CS-MTEB**, a benchmark covering 11 diverse tasks, where performance drops reach up to 27%. Finally, we examine lexicon-based vocabulary expansion and find that, while it yields partial gains, it does not close the gap to monolingual performance. These findings underscore the fragility of current systems and establish code-switching as a crucial frontier for future IR optimization.

1 Introduction

Information retrieval (IR) stands as a cornerstone infrastructure for a wide array of intelligent applications, serving as the backbone for modern search engines, retrieval-augmented generation (RAG) systems (Lewis et al., 2021), and autonomous search agents (Jin et al., 2025; Zhao et al., 2025). Its ability to efficiently locate relevant data is critical for grounding generative models and enabling users to access vast repositories of knowledge. The underlying algorithms powering IR have undergone a significant evolution, shifting from traditional statistical methods like BM25 (Robertson and Zaragoza, 2009) to semantic-aware dense retrieval (Gao et al., 2021) and sophisticated late interaction architectures (Khattab and Zaharia, 2020). Crucially, as digital information becomes increasingly globalized, the field has expanded far beyond English-centric approaches. We have witnessed a vital shift toward robust multilingual IR (Yu et al., 2024) and complex cross-lingual IR (Zuo et al., 2025), which are essential for processing the diverse linguistic landscapes of the real world.

Despite the extensive evaluation of IR across multiple languages, one pervasive linguistic phenomenon remains critically understudied in current literature: code-switching (Chanda and Pal, 2025). This omission is striking given that code-switching is a fundamental aspect of global communication, particularly as approximately 70% of the world population consists of bilingual speakers (Li, 2007). Sociolinguistic studies highlight this frequency; for instance, Ahmed (2024) investigated speech communities in three countries and observed that code-switching occurs more than 15 times every 10 minutes. Within the context of search, Gupta et al. (2014) conducted a large-scale analysis of Microsoft Bing logs and identified a substantial volume of code-switching queries. This trend was notably pronounced in the entertainment domain, where mixed-language inputs constituted around 27% of overall traffic. Collectively, these findings underscore the urgent need to address code-switching in retrieval systems. Yet, we still lack a systematic evaluation of code-switching IR capabilities.

In this paper, we present the first holistic study of code-switching IR. Our framework, summarized in Figure 1, proceeds in three stages. First, we build the **C**ode-**S**witching **R**etrieval benchmark-**L**ite (**CSR-L**), a human-annotated benchmark that captures natural code-switched queries, and evaluate statistical, dense, cross-encoder, and late-

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 13055–13071

interaction retrieval methods on it. This analysis shows that even simple query-side code-switching substantially degrades retrieval quality, including for strong multilingual retrievers, and that the degradation is accompanied by a clear shift in embedding space. Second, we scale the study beyond standard retrieval by introducing **CS-MTEB**, an MTEB-style benchmark covering 11 tasks across 7 task types; across these tasks, advanced embedding models still exhibit performance drops of up to 27%. Third, we test whether lexicon-based vocabulary expansion can mitigate the problem. Although this intervention improves English-centric retrievers, it still falls short of restoring monolingual performance. Taken together, these results identify code-switching as a major robustness gap in current IR systems. Code and datasets are publicly available in our GitHub repository and the CS-MTEB and CSR-L Hugging Face collections.

2 Related Work

IR and Embedding Models Evaluation The field of IR has undergone a fundamental transformation in its backbone methodology, evolving from lexical matching to semantic representation. In modern applications, the mainstream IR pipeline typically adopts a "retrieve-then-rerank" architecture to balance efficiency and precision. For the initial retrieval step, the paradigm has shifted toward dense retrieval, where most embedding models are trained using contrastive learning in a bi-encoder fashion. This approach encodes queries and documents into independent vector spaces, allowing for efficient similarity calculation via dot product or cosine similarity during inference. Training these models often involves sophisticated negative sampling strategies and loss functions, such as InfoNCE (Oord et al., 2018), to optimize the separation between relevant and irrelevant passages. Following retrieval, a reranking stage is often employed—frequently utilizing cross-encoders—to re-score the top candidates with finer granularity by capturing the full interaction between query and document tokens. To rigorously assess these advancements, the community has developed a wide range of benchmarks. Early efforts like BEIR (Thakur et al., 2021) focused on measuring zero-shot generalization across diverse domains, while MTEB (Muennighoff et al., 2023) expanded the scope to massive text embedding tasks beyond just retrieval. More recently, benchmarks such as BRIGHT (Su et al., 2025) have been proposed to test models on highly challenging, realistic queries that require deep reasoning, pushing the boundaries of current embedding capabilities.

Multilingual and Cross-lingual Retrieval Multilingual and cross-lingual retrieval performance of embedding models has received increasing attention. For example, MMTEB (Enevoldsen et al., 2025) evaluated embedding models in over 250 languages and across more than 500 tasks. Litschko et al. (2025) evaluated IR models on cross-dialect retrieval. For training, Wang et al. (2024b); Yu et al. (2024); Zhang et al. (2025) represent some recent attempts to build multilingual retrievers using open-source and synthetic data. However, one crucial linguistic phenomenon, code-switching, remains relatively underexplored. Litschko et al. (2023); Do et al. (2024) represent two preliminary attempts to use code-switching data to enhance multilingual and cross-lingual IR. Although Winata et al. (2024); Kim et al. (2025) touches on code-switching evaluation, it remains task- and setting-specific (e.g., sentiment analysis and bitext retrieval, focusing on late-interaction models) and does not provide a holistic picture of code-switching in embedding-based IR, which we address in this paper.

3 Code-Switching Retrieval Benchmark-Lite (CSR-L)

The naturalness of code-switched text has been examined from both theoretical (Poplack, 2020; Myers-Scotton, 1997) and empirical (Pratapa et al., 2018; Hsu et al., 2023) perspectives. However, the field remains without a single standard or automatic metric to reliably judge the naturalness of code-switching, which severely limits the scalability of benchmarks. Consequently, in this section, we employ human annotators to rewrite queries within IR benchmarks. This approach allows us to overcome the limitations of automated metrics, ensuring high data quality and facilitating a more reliable evaluation.

3.1 Building CSR-L

We selected four representative datasets containing a limited number of queries to facilitate rewriting: (1) Touché 2020 (Bondarenko et al., 2020) for argument retrieval; (2) HumanEval (Chen et al., 2021) for code retrieval; (3) TRECCOVID (Roberts et al., 2021) for biomedical IR; and (4) FollowIR (Weller et al., 2025) for evaluating instruction-following ca-

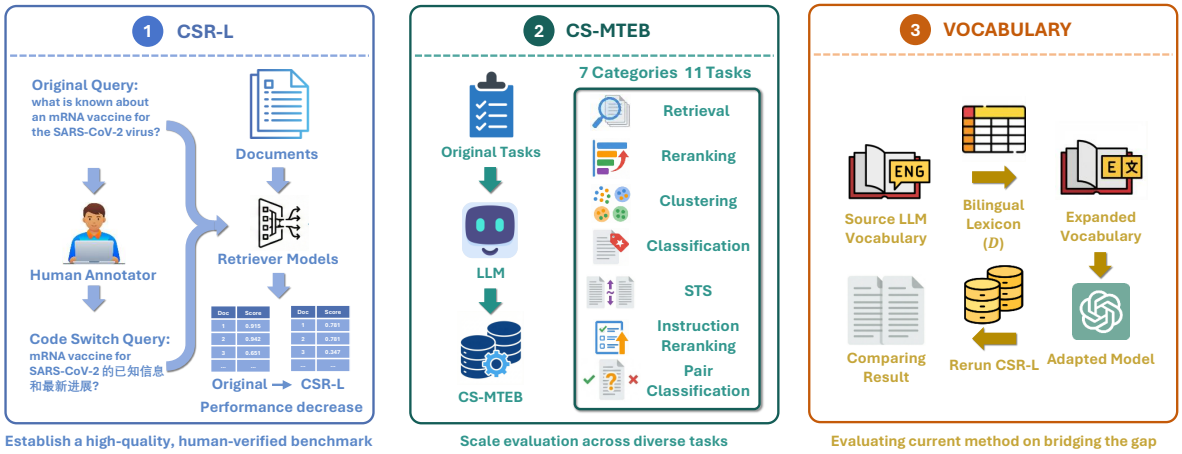


Figure 1: Overview of our comprehensive study on Code-Switching IR. Our framework proceeds in three stages: (1) CSR-L: We establish a high-quality, human-verified retrieval benchmark to assess natural mixed-language queries. (2) CS-MTEB: We scale the evaluation to 11 diverse tasks across 7 categories using LLM-assisted generation. (3) Vocabulary Expansion: We investigate lexicon-based vocabulary adaptation as a strategy to bridge the embedding space divergence between pure and code-switched text.

	Total Number			Avg. Length		Examples
Dataset	Q	$\mathcal{D}$	$\mathcal{D}^+$	Q	$\mathcal{D}$	
Touché 2020	49	303,732	34.94	16.82	451.51	Table 6
HumanEval	158	158	1.00	64.76	98.20	Table 7
TRECCOVID	50	171,332	493.46	24.36	223.51	Table 8
FollowIR	198	98,312	30.07	111.15	465.39	Table 9

Table 1: Statistics of datasets in CSR-L-Chinese. Q: number of queries; D: corpus size; D^+: average positive documents per query. Avg. Length is measured in the GPT-2 (Radford et al., 2019) tokenizer. Our query examples can be seen in the tables in Appendix.

pabilities. As these datasets are originally English-only, we rewrote the queries to introduce code-switching in two languages: Mandarin Chinese and Japanese.

Three authors of this paper participated in the query rewriting task. All are native Chinese speakers with professional proficiency in both English and Japanese, developed through their undergraduate and postgraduate education. The rewriting process followed two steps: (1) one annotator first rewrote the query into a code-switched form; and (2) a second annotator validated the result, with the authority to edit the text or discard the rewrite when necessary. The detailed instructions are provided in Appendix A. Statistics for the final Chinese dataset are presented in Table 1, while the Japanese statistics are reported in Table 10 in the appendix.

3.2 Evaluation Setup

We evaluate CSR-L with four families of IR methods: (1) the lexical baseline BM25 (Robertson and Zaragoza, 2009); (2) bi-encoder retrievers, including *all-MiniLM-L12-v2* (Reimers and Gurevych, 2019), *e5-large-v2* (Wang et al., 2024a) and *multilingual-e5-large (mE5-large)* (Wang et al., 2024b), *bge-m3* (Chen et al., 2024), *Arctic-Embed-m/l-v2.0* (Yu et al., 2024), and *Qwen3-Embedding-0.6/4/8B* (Zhang et al., 2025); (3) cross-encoder rerankers, including *jina-reranker-v3* (Wang et al., 2025), *bge-reranker-v2-m3* (Chen et al., 2024), and *Qwen3-Reranker-0.6/4/8B* (Zhang et al., 2025); and (4) the late-interaction retriever *ColBERT v2* (Santhanam et al., 2022).

We use nDCG@10 as the primary metric throughout the evaluation, with the exception of FollowIR, where we report pairwise-MRR (*p-MRR*). For each method, we compare performance on the original queries and their code-switched counterparts. For the cross-encoder results in CSR-L, we score each query–document pair directly over the full document set, rather than reranking a top-k candidate pool produced by a separate first-stage retriever. Accordingly, the absolute cross-encoder numbers in Table 2 and Table 5 should be interpreted as direct full-corpus scoring results rather than as conventional two-stage reranking performance. Additional metric details are provided in Appendix B.

Method Family	Model	Touché 2020		HumanEval		TRECCOVID		FollowIR		Avg	Avg	Drop
		Orig	CSR-L	Orig	CSR-L	Orig	CSR-L	Orig	CSR-L	Orig	CSR-L	Δ
Statistical	BM25	60.32	37.68	35.02	41.79	55.62	46.43	-0.62	-1.8	37.59	31.03	-6.56
Bi-encoder	*e5-large-v2*	42.52	22.88	80.70	72.93	66.64	50.42	-0.99	-4.97	47.22	35.32	-11.90
	all-MiniLM-L12-v2	49.22	23.85	70.08	60.37	51.17	39.51	-0.66	-3.36	42.45	30.09	-12.36
	mE5-large	49.32	42.75	81.15	74.04	71.56	56.54	-3.38	-2.28	49.66	42.76	-6.90
	bge-m3	55.02	50.00	61.33	59.26	54.70	52.32	-2.94	-3.00	42.03	39.65	-2.38
	Arctic-Embed-m-v2.0	65.29	48.46	78.7	75.05	80.45	74.15	-3.20	-4.32	55.31	48.34	-6.97
	Arctic-Embed-l-v2.0	64.05	54.91	71.27	68.94	83.63	76.99	-2.45	-2.47	54.13	49.59	-4.54
	Qwen3-Embedding-0.6B	71.65	61.30	94.24	94.43	89.43	81.66	5.10	4.07	65.11	60.37	-4.74
	Qwen3-Embedding-4B	75.07	66.67	98.12	96.17	92.95	88.67	11.87	8.91	69.50	65.11	-4.39
	Qwen3-Embedding-8B	75.77	68.55	99.22	98.90	94.68	89.72	9.86	7.63	69.88	66.20	-3.68
Cross-encoder	*jina-reranker-v3*	22.68	24.96	85.53	84.43	81.32	68.07	-0.27	-0.17	47.32	44.32	-3.00
	bge-reranker-v2-m3	35.48	27.86	43.74	49.77	79.00	67.17	-1.38	0.32	39.21	36.28	-2.93
	Qwen3-Reranker-0.6B	29.15	23.91	83.74	84.11	84.30	71.19	1.40	-0.01	49.65	44.80	-4.85
	Qwen3-Reranker-4B	37.76	28.34	85.29	84.11	85.44	70.86	2.33	-1.01	52.71	45.58	-7.13
	Qwen3-Reranker-8B	40.91	32.01	85.53	84.62	84.58	69.88	2.74	0.56	53.44	46.77	-6.67
Late-interaction	ColBERT v2	61.62	29.30	40.30	42.46	69.30	53.74	-0.95	-0.46	42.57	31.26	-11.31

Table 2: nDCG@10 and *p-MRR* on the original (Orig) and code-switched (CSR-L) queries across four IR benchmarks on English-Chinese code-switching. Avg is the macro-average over the four datasets. Drop Δ is computed as Avg(CSR-L) - Avg(Orig); negative values indicate performance degradation under code-switching.

4 CSR-L Results

4.1 General Results

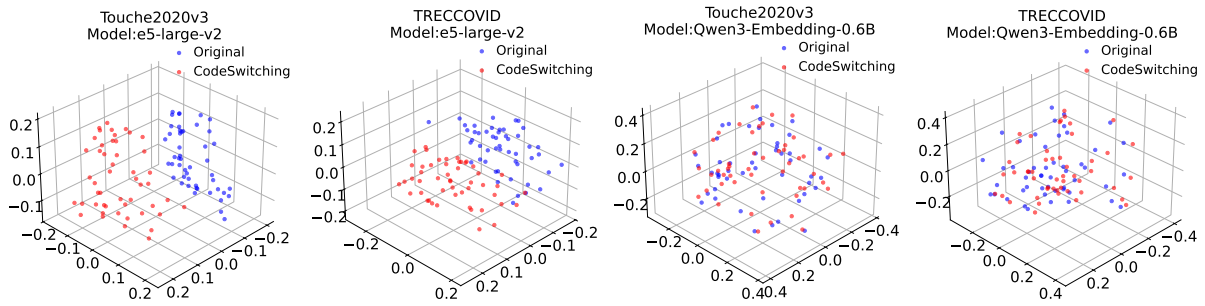
The Chinese results are shown in Table 2, while the Japanese results are reported in Table 5 in Appendix C. The overall pattern is highly consistent across the two languages. First, query-side code-switching alone substantially degrades performance on the main retrieval datasets, even though the underlying documents remain unchanged. The newly added multilingual bi-encoder baselines, *mE5-large* and *bge-m3*, follow the same trend, showing that multilingual encoders are more robust but not immune. The degradation is especially large on Touché 2020 and TRECCOVID, whereas it is milder on HumanEval, likely because that benchmark is structurally simpler. Among English-centric bi-encoders such as *e5-large-v2*, the drop reaches roughly 15 points on the two general retrieval datasets. Even for the *Qwen3-Embedding* series, which is comparatively more robust, the decrease on Touché 2020 and TRECCOVID still exceeds 8 points in some settings. Model scaling helps, but even the 8B variant does not eliminate the gap.

While all evaluated models struggle, an important distinction emerges between English-centric systems and multilingual retrievers. Multilingual models generally exhibit a smaller relative decline than their English-only counterparts. For instance, when controlling for model size, *Arctic-Embed-m-v2.0* experiences a substantially smaller drop than *e5-large-v2*. This relative stability suggests that exposure to diverse languages during training provides a meaningful benefit, helping the model interpret code-switching patterns and partially absorb the disruption caused by linguistic mixing.

Finally, we observe no significant variation in robustness across different retrieval paradigms. For example, despite the higher computational cost associated with cross-encoders, these models do not exhibit superior resistance to the performance drops caused by code-switched queries. We note, however, that the cross-encoder scores in Table 2 and Table 5 come from direct full-corpus scoring rather than a standard retrieve-then-rerank pipeline, so their absolute values should not be read as directly comparable to conventional reranking benchmarks. This vulnerability is equally prevalent in statistical methods and late-interaction frameworks. Taken together, our results on CSR-L suggest that while multilingual pre-training offers partial mitigation, code-switching poses a fundamental challenge that neither architectural complexity nor current scaling strategies can fully overcome.

4.2 Embedding Space Analysis

Visualizing embedding spaces provides valuable insights into the underlying causes of retrieval failure. In this subsection, we focus on Touché 2020 and TRECCOVID, two datasets where models exhibited significant performance degradation. We selected *e5-large-v2* and *Qwen3-Embedding-0.6B* as representative models and visualized their query representations in a three-dimensional space using Principal Component Analysis (PCA), as shown in



(a) *e5* on Touché (b) *e5* on TREC (c) *Qwen 0.6B* on Touché (d) *Qwen 0.6B* on TREC

Figure 2: The visualization of *e5* and *Qwen 0.6B* embeddings on two IR datasets.

Figure 2.

Our analysis uncovers distinct geometric behaviors. For the English-centric retriever *e5-large-v2*, code-switching induces a drastic shift in the embedding space: the original and code-switched queries separate into two disjoint dense clusters rather than forming a shared semantic distribution. In contrast, multilingual models exhibit greater stability. For example, on Touché 2020, the centroid distance is smaller for *Qwen3-Embedding-0.6B* than for *e5-large-v2* (0.20 vs. 0.25), and the two query sets overlap much more strongly. This geometric resilience likely contributes to the more moderate performance declines observed in multilingual models. Nevertheless, the gap does not disappear, suggesting that code-switching introduces semantic difficulties that go beyond what standard multilingual pre-training currently resolves.

5 CS-MTEB

In the preceding section, our results demonstrated that regardless of model scale, IR paradigm, or multilingual pre-training, current methods consistently fail to maintain parity with monolingual performance when processing code-switched queries. Recognizing the critical need to assess the universality of this deficit, we now expand our evaluation beyond standard retrieval tasks. In this section, we leverage LLMs to scale our investigation, covering a broader spectrum of task types, datasets, and language pairs. By aligning with the rigorous standards of the general-purpose MTEB benchmark, we introduce **CS-MTEB**. This comprehensive framework is designed to provide a holistic diagnosis of text embedding models, systematically uncovering the boundaries of their success and failure in mixed-language scenarios.

5.1 Covered Tasks

To ensure a comprehensive evaluation, we curated a diverse set of tasks for this benchmark, spanning the following categories:

- Instruction Reranking: We incorporate *FollowIR* (Weller et al., 2025), aligning our setup with standard MTEB protocols.
- Retrieval: Beyond the three datasets established in CSR-L, we expand the scope by including *Arguana* (Wachsmuth et al., 2018) and *ClimateFEVERHardNegatives* (Diggelmann et al., 2021).
- Clustering: We utilize *ArXivHierarchicalClusteringP2P* (Enevoldsen et al., 2025). For this task, we randomly introduce code-switching into 10% of the document text to quantify the resulting impact on clustering stability.
- Classification: We adopt the test set of *TweetSentimentExtractionClassification* (Enevoldsen et al., 2025) to serve as the representative classification benchmark.
- Semantic Textual Similarity (STS): We employ the *STS Benchmark* (Enevoldsen et al., 2025) to assess the models' semantic understanding. Specifically, we apply code-switching to one sentence within each pair to test cross-lingual alignment.
- Reranking: We leverage *AskUbuntuDupQuestions* (Lei et al., 2016) to evaluate reranking capabilities. Consistent with our retrieval setup, we apply code-switching exclusively to the query side.
- Pair Classification: We utilize *TwitterSemEval2015* (Xu et al., 2015) for pair classification. Similarly, we introduce code-switching into one sentence of the pair to challenge the model's judgment.

By encompassing 11 distinct tasks across these 7 categories, we aim to construct a holistic picture

Model	Setting	Instr. Rerank (1)	Retrieval (5)	Clust. (1)	Cls. (1)	STS (1)	Rerank (1)	Pair Cls. (1)	Total (11)
e5-large-v2	Original	-0.99	51.78	62.00	73.97	84.55	60.17	59.88	55.91
	Chinese	-4.97	40.49	55.71	57.47	59.42	22.39	54.42	40.70
	Japanese	-1.93	40.14	54.87	62.76	65.71	25.75	55.17	43.21
	German	-2.54	40.80	55.65	61.74	67.54	26.20	56.11	43.64
	Spanish	-2.31	41.49	59.03	63.06	69.55	26.38	57.79	45.00
Arctic-Embed-m-v2.0	Original	-3.20	64.21	60.09	64.78	75.97	62.37	58.09	54.62
	Chinese	-4.32	57.54	54.09	56.94	62.48	34.30	56.29	45.33
	Japanese	-3.13	57.79	54.45	58.80	64.73	37.15	56.43	46.60
	German	-3.83	60.39	53.05	62.66	68.72	37.09	57.62	47.98
	Spanish	-3.43	61.60	58.52	62.09	68.12	38.08	58.17	49.02
Qwen3-Embedding-0.6B	Original	5.10	73.67	68.21	72.07	91.14	63.09	75.55	64.12
	Chinese	4.07	69.69	61.70	68.50	86.69	37.13	72.90	57.24
	Japanese	4.11	68.68	63.10	68.58	85.23	37.33	72.29	57.05
	German	3.54	66.26	64.83	68.00	85.86	36.04	73.08	56.80
	Spanish	3.75	69.80	63.77	68.71	86.16	36.20	74.25	57.52

Table 3: CS-MTEB results by model and evaluation setting. Columns correspond to CS-MTEB task categories, with the number of tasks per category in parentheses. The result is the macro average over 7 task categories / Mean (TaskType).

of how code-switching influences the performance of text embedding models.

5.2 Experimental Setup

Because fully human rewriting is infeasible at MTEB scale, we use an LLM to generate the code-switched variants. We refined the prompt templates through several iterations, grounding the design in the human-authored CSR-L queries to preserve both naturalness and information need. The final prompt template is provided in Appendix D, and a manual quality check on 50 sampled rewritten queries is reported in Table 19.

We selected `MiMo-V2-Flash` (Core Team et al., 2026) as the backbone model for this generation task. Our evaluation incorporates 9 languages mixed with English, including Chinese, Japanese, German, Spanish, Korean, French, Italian, Portuguese, and Dutch. For the experimental analysis, we assess the following models: *e5-large-v2* (Wang et al., 2024a), *Arctic-Embed-m-v2.0* (Yu et al., 2024), and *Qwen3-Embedding-0.6B* (Zhang et al., 2025).

5.3 Results

The main results of CS-MTEB on four languages are presented in Table 3, with an additional five languages reported in Table 15. The same qualitative pattern from CSR-L reappears: across models, tasks, and language pairs, code-switching remains a broad and persistent bottleneck.

First, for English-centric models such as *e5-large-v2* (Wang et al., 2024a), the performance degradation is substantial and consistent. We observe a significant drop in the average score across all four language mixtures, ranging from approximately 10 to 15 points. Critically, this decline occurs regardless of linguistic proximity; the model suffers similar losses whether English is mixed with typologically distinct languages like Chinese and Japanese, or with closer European relatives such as German and Spanish. This universality suggests that without explicit multilingual alignment, the embedding space is highly fragile to the semantic noise introduced by code-switching.

In contrast, the multilingual retriever *Arctic-Embed-m-v2.0* (Yu et al., 2024) exhibits greater resilience, although it is not immune. Across the same set of languages, the performance decline is noticeably mitigated compared to the monolingual baseline. For instance, in the Spanish-English setting, the model experiences a drop of approximately 5 points, compared to the $\sim$10 point drop observed in *e5-large-v2*. While this indicates that exposure to diverse linguistic data provides a foundational robustness, the persistence of these gaps underscores that standard multilingual training alone is insufficient to fully bridge the code-switching deficit.

Analyzing performance across different task categories reveals distinct sensitivities. While retrieval tasks exhibit a consistent decline, finer-grained objectives can be substantially more fragile, with reranking showing the sharpest failures. For example, in the Japanese code-switching setting, *e5-large-v2* suffers a catastrophic degradation in reranking, plummeting from a baseline of 60.17 to just 25.75. In contrast, more decision-oriented tasks such as pair classification tend to be comparatively less sensitive, likely because they can

rely on coarse semantic cues rather than the precise ordering and alignment required by ranking-based objectives. Taken together, these results establish that code-switching challenges vary significantly by task demands, motivating targeted optimization beyond simple scale.

6 Vocabulary Expansion for Retrieval

The CSR-L and CS-MTEB results establish two consistent observations: code-switching hurts end-task performance, and it also perturbs the representation space. This suggests that at least part of the failure arises at the input and encoding level. One plausible contributor is vocabulary and tokenization coverage: when tokens from a secondary language are split into many low-frequency subwords, the resulting representations can become noisy and drift away from the monolingual manifold. This motivates a controlled, low-cost intervention: lexicon-based vocabulary expansion, which extends the tokenizer with high-frequency missing words from the target language while leaving the main body of the retriever unchanged. If this intervention recovers a meaningful portion of the performance drop, it would indicate that vocabulary coverage is an important bottleneck; otherwise, it would imply that code-switching failures stem from deeper representation and training mismatches beyond the tokenizer. Within multilingual NLP, a significant body of work focuses on extending monolingual models to multilingual settings. For example, Wang et al. (2022) adapted monolingual models to cover thousands of languages using lexicon-based techniques, and Zeng et al. (2023) introduced a vocabulary expansion algorithm designed to elicit robust multilingual performance from monolingual backbones. In this section, we ask whether the same idea can improve robustness to code-switched queries.

6.1 Lexicon-Based Vocabulary Expansion

To investigate whether bridging the semantic gap between languages can mitigate the performance degradation observed in code-switching retrieval, we implement a lexicon-based vocabulary expansion strategy. This method, adapted from the initialization techniques proposed by Zeng et al. (2023), allows us to project the semantic capabilities of a well-aligned source language (e.g., English) onto the target language without requiring extensive multilingual pre-training.

Formally, we utilize an independent bilingual lexicon $\mathcal{D} = \{(w_t, w_s)\}$, which consists of word-level translation pairs mapping the target language to the source language. Distinct from this linguistic resource, the pre-trained source model operates on a subword vocabulary $\mathcal{V}_{pre}$ (e.g., WordPiece or BPE tokens) with a corresponding embedding matrix $\mathbf{E}_{pre} \in \mathbb{R}^{|\mathcal{V}_{pre}| \times d}$. Our objective is to initialize the embeddings for the target tokens $\mathcal{V}_{target}$ based on their semantic equivalents in the lexicon.

A significant challenge lies in the granularity mismatch: entries in the lexicon $\mathcal{D}$ are typically whole words, whereas the model vocabulary $\mathcal{V}_{pre}$ consists of subword units. To address this, we define the tokenizer as $T(\cdot)$ that decomposes a linguistic word w into a sequence of subword tokens $[k_1, k_2, \ldots, k_m]$, where each $k_i \in \mathcal{V}_{pre}$.

For a given target word w_t, we first identify its set of source translations $\mathcal{N}(w_t) = \{w_s \mid (w_t, w_s) \in \mathcal{D}\}$. To obtain a vector representation for a specific source word w_s, we tokenize it into its constituent subwords and average their pre-trained embeddings:

$$\mathbf{v}_{w_s} = \frac{1}{|T(w_s)|} \sum_{k \in T(w_s)} \mathbf{e}_k \tag{1}$$

where $\mathbf{e}_k \in \mathbf{E}_{pre}$ is the embedding of the subword token k. Finally, to initialize the embedding for the target token w_t, we aggregate the representations of all its valid source translations:

$$\mathbf{e}_{w_t} = \frac{1}{|\mathcal{N}(w_t)|} \sum_{w_s \in \mathcal{N}(w_t)} \mathbf{v}_{w_s} \tag{2}$$

In cases where a target token has no translation in the lexicon (i.e., $\mathcal{N}(w_t) = \emptyset$), we initialize $\mathbf{e}_{w_t}$ using a standard normal distribution $\mathcal{N}(0, \sigma^2)$. This hierarchical aggregation, from subword to word, then from source word to target word, ensures that the messy, fragmented nature of the pre-trained vocabulary does not hinder the effective transfer of semantic information to the code-switching context.

6.2 Experiments and Results

Experimental Setup We apply our vocabulary expansion strategy to two representative English-only retrievers: *all-MiniLM-L12-v2* and *e5-large-v2*. To construct the semantic mapping, we utilize the high-quality bilingual lexicons provided by Conneau et al. (2018). We independently expand these models to support both Chinese and

Model	Settings	Touché 2020 (Argument)	HumanEval (Code)	TRECCOVID (Science)	FollowIR (IF)	Avg
all-MiniLM-L12-v2	orig model + CSR-L-Chinese	23.85	60.37	39.51	-3.36	30.09
	adapted model + CSR-L-Chinese	40.01	64.44	48.32	-1.87	**37.73**
	orig model + CSR-L-Japanese	22.12	62.18	36.12	-0.65	29.94
	adapted model + CSR-L-Japanese	30.86	65.72	41.64	-0.88	**34.34**
e5-large-v2	orig model + CSR-L-Chinese	22.88	72.93	50.42	-4.97	35.32
	adapted model + CSR-L-Chinese	38.55	74.18	64.26	-2.99	**43.50**
	orig model + CSR-L-Japanese	22.88	72.49	45.34	-1.93	34.70
	adapted model + CSR-L-Japanese	26.98	76.77	56.96	-1.52	**39.80**

Table 4: Performance comparison of original and vocabulary-adapted models on the CSR-L benchmarks.

Japanese, subsequently evaluating the performance of the adapted versions on our CSR-L benchmark.

Results As shown in Table 4, lexicon-based vocabulary expansion consistently improves robustness to code-switching for both evaluated English-only retrievers. For `all-MiniLM-L12-v2`, adaptation increases the macro-average from 30.09 to 37.73 on CSR-L-Chinese and from 29.94 to 34.34 on CSR-L-Japanese, indicating a clear but partial recovery. Similarly, `e5-large-v2` benefits from adaptation, with the average improving from 35.32 to 43.50 (Chinese) and from 34.70 to 39.80 (Japanese). The gains are driven primarily by the two general retrieval benchmarks (e.g., Touché 2020 and TRECCOVID), while improvements on HumanEval are comparatively smaller. Overall, these results mirror the earlier finding that code-switching imposes a substantial bottleneck, and demonstrate that vocabulary expansion provides a low-cost mitigation, but does not fully eliminate the deficit.

7 Discussion

In this paper, we identified code-switching as a persistent and universal bottleneck for modern IR. Across our constructed CSR-L and CS-MTEB benchmarks, we observed significant performance degradation regardless of whether the system employs statistical, dense, or late-interaction architectures. While multilingual training offers a degree of geometric stability, mitigating the severity of these drops compared to English-centric baselines, it fails to fully immunize models against the semantic disruption caused by mixed-language queries. Furthermore, our experiments with lexicon-based vocabulary expansion provide a nuanced insight: although this low-cost intervention yields measurable performance improvements, the resulting models still trail significantly behind English-only settings. This persistent gap underscores that code-switching is not merely a vocabulary coverage issue resolvable by surface-level patches, but a complex semantic challenge that necessitates dedicated architectural or training innovations to achieve true parity with monolingual systems.

Semantic Alignment vs. Retrieval Relevance Our findings reveal a critical, often overlooked distinction between *semantic alignment* and *retrieval relevance* in code-switching contexts. While recent benchmarks such as **MINERS** (Winata et al., 2024) demonstrate that multilingual models can achieve competitive performance in semantic tasks like bitext mining without fine-tuning, our results on CSR-L and CS-MTEB paint a significantly more complex picture. We observe that while current models effectively align synonyms across languages, which explains their resilience in simpler retrieval tasks, they struggle profoundly when tasked with the nuanced relevance modeling required for high-precision IR. This fragility is even preserved in large-scale foundation models; for instance, on CS-MTEB reranking tasks, the performance of *Qwen3-Embedding-0.6B* plummets from a monolingual baseline of 63.09 to 37.33 in the Japanese setting. To sum up, these results underscore the immense heterogeneity across text embedding applications: proficiency in cross-lingual alignment does not guarantee robustness in code-switching IR, further validating the need for the specialized development of code-switching retrieval systems.

The Limits of Direct Multilingual Interaction Zuo et al. (2025) recently established that while LLMs excel as rerankers when inputs are translated (noisy monolingual IR), they fall severely short when interacting directly with multilingual bi-encoder outputs without intermediate translation. Our work extends this conclusion to the code-switching domain: just as models struggle with direct cross-lingual retrieval, they are similarly fragile when processing fluidly mixed-language queries. The failure of our vocabulary expansion experiments further corroborates this, indicating that surface-level fixes cannot compensate for the

model's fundamental inability to process "native" mixed-language sequences. Collectively, these findings imply that future progress depends not on better translation or alignment but on developing training data that treat code-switching as a distinct linguistic modality.

Limitations

We identify two main limitations in our work.
Language and phenomenon coverage. Our benchmarks operationalize code-switching as natural, query-level mixing between English and a small set of partner languages, which keeps the evaluation controlled and aligns with common search behavior where English technical terms appear inside otherwise non-English queries. At the same time, code-switching in the wild spans a broader space (e.g., romanization, transliteration and spelling variation, community-specific conventions, and mixed-language documents), which is not the primary focus of this study and remains a straightforward direction for future benchmark extensions.
Annotation and generation noise. Working with code-switched text inevitably involves human judgment about what constitutes a natural switch while preserving the original information needs. We mitigate this through bilingual annotators, validation checks, conservative rewrite guidelines, and a manual spot check of generated CS-MTEB queries, but modest stylistic variation and occasional generation artifacts are difficult to eliminate entirely at scale. Accordingly, we emphasize consistent trends across models and settings, and release our resources to facilitate replication and expansion under alternative annotation or generation protocols.

Acknowledgments

This work was supported by JST CRONOS (Grant Number JPMJCS25K5) and JST NEXUS (Grant Number JPMJNX25CA). Weihao Xuan is supported by RIKEN Junior Research Associate (JRA) Program.

References

Yusuf M. Ahmed. 2024. Code-switching in multilingual communities: Case studies from kenya, malaysia, and the UAE. *Journal of International English Research Studies (JIERS)*, 2(4):13–21.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of touché 2020: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 384–395, Cham. Springer International Publishing.

Supriya Chanda and Sukomal Pal. 2025. Overview of the shared task on code-mixed information retrieval from social media data. In *Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '24, page 29–31, New York, NY, USA. Association for Computing Machinery.

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. *Preprint*, arXiv:2402.03216.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Preprint*, arXiv:1710.04087.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, and 107 others. 2026. Mimo-v2-flash technical report. *Preprint*, arXiv:2601.02780.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Junggeun Do, Jaeseong Lee, and Seung-won Hwang. 2024. ContrastiveMix: Overcoming code-mixing dilemma in cross-lingual transfer for information retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 197–204, Mexico City, Mexico. Association for Computational Linguistics.

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, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *Preprint*, arXiv:2502.13595.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Parth Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 677–686, New York, NY, USA. Association for Computing Machinery.

I-Hung Hsu, Avik Ray, Shubham Garg, Nanyun Peng, and Jing Huang. 2023. Code-switched text synthesis in unseen language pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5137–5151, Toronto, Canada. Association for Computational Linguistics.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Jonghwi Kim, Deokhyung Kang, Seonjeong Hwang, Yunsu Kim, Jungseul Ok, and Gary Lee. 2025. MiLQ: Benchmarking IR models for bilingual web search with mixed language queries. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22643–22659, Suzhou, China. Association for Computational Linguistics.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1279–1289, San Diego, California. Association for Computational Linguistics.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Wei Li. 2007. The bilingualism reader / edited by li wei.

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 3096–3108, Toronto, Canada. Association for Computational Linguistics.

Robert Litschko, Oliver Kraus, Verena Blaschke, and Barbara Plank. 2025. Cross-dialect information retrieval: Information access in low-resource and high-variance languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10158–10171, Abu Dhabi, UAE. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Shana Poplack. 2020. Sometimes i'll start a sentence in spanish y termino en español: toward a typology of code-switching. In *The bilingualism reader*, pages 213–243. Routledge.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of trec-covid. *Preprint*, arXiv:2104.09632.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *Preprint*, arXiv:2112.01488.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2025. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *Preprint*, arXiv:2407.12883.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for listwise document reranking. *Preprint*, arXiv:2509.25085.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. 2024. MINERS: Multilingual language models as semantic retrievers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2742–2766, Miami, Florida, USA. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *Preprint*, arXiv:2412.04506.

Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, Rob Voigt, and Jie Yang. 2023. Greenplm: cross-lingual transfer of monolingual pre-trained language models at almost no cost. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Fuheng Zhao, Jiayue Chen, Yiming Pan, Tahseen Rabbani, Divyakant Agrawal, and Amr El Abbadi. 2025. Access paths for efficient ordering with large language models. *arXiv preprint arXiv:2509.00303*.

Longfei Zuo, Pingjun Hong, Oliver Kraus, Barbara Plank, and Robert Litschko. 2025. Evaluating large language models for cross-lingual retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11415–11429, Suzhou, China. Association for Computational Linguistics.

A Instructions Given to Annotators for Rewriting the Queries

For Chinese CSR-L query rewriting, the annotators are native Chinese speakers with years of experience in English-speaking environments. They are asked to rewrite each original query into a natural code-switched form that reflects realistic conversational and search behavior. For Japanese CSR-L, the annotators are familiar with both English and Japanese usage, and the same instructions are applied. We reproduce the Chinese instructions below:

Instructions to Annotators

Task: Rewrite the given IR benchmark query into a **natural code-switched query** that mixes **English and the target language**, while keeping the **same information need**.

Instructions:
- Read the original query and identify the exact information need (what the user wants to find).
- Rewrite it as a single query that a bilingual person might realistically type, mixing English and the target language.
- **Do not change meaning:** do not add new details or constraints; do not remove important details (time, location, stance, entities, domain terms).
- **Keep key terms stable:** keep named entities, product names, dataset identifiers, and technical terms unchanged unless there is a widely used and unambiguous target-language form.
- **Both languages must contribute:** avoid a fully monolingual query and avoid adding only one borrowed word. Each language should contain meaningful content.
- **Naturalness:** prefer phrase-level mixing (often keep technical keywords in English, express framing in the target language). Avoid word-by-word literal translation.
- **Length:** keep the rewrite roughly similar in length (about $\pm 20\%$). Do not add explanations or extra sentences.
- **If the query is about code:** do not translate code tokens, function names, variable names, operators, or error messages. Only rewrite the surrounding natural language.
- If you feel the query is impossible to rewrite to be natural code-switching queries, write **None**.

Example
Rewritten query: What are the causes and treatments of 神经性厌食症and 神经性贪食症?

Now, here is the query awaiting rewritten:
{query}

B Additional Details on Evaluation

Unless otherwise specified, all benchmark settings follow the standard MTEB evaluation process. Due to VRAM limitations, however, we set the batch size to 4 rather than 32 for all retrieval tasks except HumanEvalRetrieval. To speed up inference, we also enable FlashAttention 2 whenever the model supports it, which may lead to slight differences from the public MTEB leaderboard. The metrics used for each task type are as follows:

- Retrieval: nDCG@10
- Instruction Reranking: p-MRR (Weller et al., 2025)
- Clustering: V-measure (Rosenberg and Hirschberg, 2007)
- Classification: Accuracy
- STS: Cosine Spearman correlation
- Reranking: MAP@1000
- Pair Classification: mean average precision

C CSR-L Results on Japanese

The results for CSR-L-Japanese are presented in Table 5.

D Prompt Template for Doing Code-switching

We take the prompt for CSR-L-Chinese tasks as an example, shown below:

Prompt Template: English-Chinese Code-Switching Rewrite

Role: You rewrite English paragraphs/sentences into natural English–Chinese code-switched texts, like something a bilingual Mainland Chinese/English user would type. If the input is a paragraph, you should do **{num_cs}** code-switching changes in the original text.
Task:

- **Input:** one dense English paragraph/sentence (for search).
- **Output:** one single English-Chinese mixed paragraph/sentence.

Style & rules:

1. **Base language**
   - The query should be mainly in English.
   - Use Simplified Chinese only where Chinese is the more natural or widely used expression for Chinese users.
2. **Preserve meaning**
   - Keep all important keywords and intent from the original sentence.

Method Family	Model	Touché 2020		HumanEval		TRECCOVID		FollowIR		Avg	Avg	Drop
		Orig	CSR-L	Orig	CSR-L	Orig	CSR-L	Orig	CSR-L	Orig	CSR-L	Δ
Statistical	BM25	60.32	39.17	35.02	34.75	55.62	43.48	-0.62	0.48	37.59	29.47	-8.12
Bi-encoder	*e5-large-v2*	42.52	22.88	80.70	72.49	66.64	45.34	-0.99	-1.93	47.22	34.70	-12.52
	all-MiniLM-L12-v2	49.22	22.12	70.08	62.18	51.17	36.12	-0.66	-0.65	42.45	29.94	-12.51
	mE5-large	49.32	40.54	81.15	75.28	71.56	53.85	-3.38	-2.26	49.66	41.85	-7.81
	bge-m3	55.02	45.53	61.33	58.85	54.70	45.41	-2.94	-2.84	42.03	36.74	-5.29
	Arctic-Embed-m-v2.0	65.29	47.89	78.70	74.29	80.45	73.35	-3.20	-3.13	55.31	48.10	-7.21
	Arctic-Embed-l-v2.0	64.05	53.19	71.27	70.60	83.63	79.07	-2.45	-2.04	54.13	50.21	-3.92
	Qwen3-Embedding-0.6B	71.65	58.77	94.24	94.80	89.43	78.63	5.10	4.11	65.11	59.08	-6.03
	Qwen3-Embedding-4B	75.07	60.33	98.12	96.15	92.95	88.11	11.87	11.26	69.50	63.96	-5.54
	Qwen3-Embedding-8B	75.77	68.69	99.22	98.52	94.68	88.14	9.86	8.73	69.88	66.02	-3.86
Cross-encoder	*jina-reranker-v3*	22.68	24.49	85.53	83.41	81.32	71.28	-0.27	-0.03	47.32	44.79	-2.53
	bge-reranker-v2-m3	35.48	28.94	43.74	39.91	79.00	66.14	-1.38	-0.76	39.21	33.56	-5.65
	Qwen3-Reranker-0.6B	29.15	23.93	83.74	81.90	84.30	72.80	1.40	0.55	49.65	44.80	-4.85
	Qwen3-Reranker-4B	37.76	27.96	85.29	83.75	85.44	73.81	2.33	0.55	52.71	46.52	-6.19
	Qwen3-Reranker-8B	40.91	32.21	85.53	84.26	84.58	71.71	2.74	1.22	53.44	47.35	-6.09
Late-interaction	ColBERT v2	61.62	31.18	40.30	34.23	69.30	48.86	-0.95	0.87	42.57	28.79	-13.78

Table 5: nDCG@10 and *p-MRR* on the original (Orig) and code-switched (CSR-L) queries across four IR benchmarks on English-Japanese code-switching. Avg is the macro-average over the four datasets. Drop Δ is computed as Avg(CSR-L) - Avg(Orig); negative values indicate performance degradation under code-switching.

- Keep product names, library names, frameworks, and proper nouns in English (e.g., Python, React, Kubernetes, Notion, Apple, RTX 4060), unless there is an extremely standard Chinese name everyone uses.

3. **When to use Chinese** (use Chinese especially for idioms, for example)
   - Learning/guide phrases: 入门教程, 入门课程, 学习路线, 使用教程, 实战教程, 完整指南, 保姆级教程, 选购指南, 选购建议, 配置推荐, 排行榜, 使用心得, 避坑, 踩坑经验, 速查表.
   - Evaluation/preference words: 性价比, 便宜一点, 高性价比, 对比, 推荐, 怎么选, 适合新手, 适合大学生.
   - Very common Chinese net-terms that replace simple English nouns, for example:
     - "gaming laptop" → 游戏本
     - Keep most other general nouns in English unless the Chinese term is clearly more common among Chinese users.
4. **How to mix**
   - Keep the overall structure and most content words in English.
   - Insert Chinese at the phrase level (e.g., "... 入门教程", "... 性价比对比", "best 游戏本 for college students 性价比"), not every other word.
   - Keep it short: do not turn it into a full explanation.
5. **Do NOT**
   - Do NOT translate the whole sentence into Chinese.
   - Do NOT restate the sentence twice in different languages.
   - Do NOT add romanization or language labels (no "[in Chinese]", etc.).
   - Do NOT add commentary or explanation.
   - Do NOT use emoji to do code switching.
   - Do NOT output any chain-of-thought, reasoning, or thinking process.
   - Do NOT include any preamble like "Here is the rewritten version:".
6. **Be faithful**
   - Keep the same length as the original paragraph/sentence.
   - Keep the same sentence order as the original input.
7. **Output format**
   - Your response MUST start directly with `<code_switched_output>` and end with `</code_switched_output>`.
   - Output ONLY the XML tags with your answer inside. Nothing else.
   - No text before or after the XML tags.

Output XML format:

```
<code_switched_output>Your
answer</code_switched_output>
```

Examples of desired style (do NOT output these literally):

- `Python data analysis 入门教程`
- `best 游戏本 for college students 性价比`
- `Notion personal knowledge management 入门教程`
- `markdown 速查表`

Query
Should teachers get 终身教职?

Table 6: CSR-L-Chinese Touché 2020 Code-Switching Example.

Query
Given an array of non-negative integers, return a sorted copy: if sum(first index value, last index value) is odd sort ascending, if even sort descending. 注意: 不要修改原数组。示例: sort_array([]) => [], sort_array([5]) => [5], sort_array([2,4,3,0,1,5]) => [0,1,2,3,4,5]?

Table 7: CSR-L-Chinese HumanEval Code-Switching Query Example.

Query
Will SARS-CoV2 infected people develop immunity，交叉保护是否可能?

Table 8: CSR-L-Chinese TRECCOVID Code-Switching Query Example.

Keep the similar length (measured by number of sentences) as the original input. Do not over-simplify the query.
Now rewrite this paragraph/sentence into a natural English–Chinese code-switched version:
{your_query_here}

E CSR-L-Chinese Query Examples

Examples of rewritten code-switching queries in CSR-L-Chinese are listed in tables below from Table 6 to Table 9.

F Japanese-CSR-L Statistics And Query Examples

Statistics and the examples of rewritten code-switching queries in Japanese-CSR-L are listed in tables below from Table 10 to Table 14.

G Additional Results on CS-MTEB

In Table 15, we report CS-MTEB results on additional five languages.

Query-og
What efforts have been made to stabilize the 比萨斜塔, and how successful have the efforts been?
Instruction-og
Relevant documents provide discussions of the current condition of the tower, describe the 加固措施 taken, and/or provide measurements reflecting change in the tower.
Query-changed
What efforts have been made to stabilize the 比萨斜塔, and how successful have the efforts been?
Instruction-changed
Relevant documents provide discussions of the current condition of the tower, describe the 加固措施 taken, and/or provide measurements reflecting change in the tower. Exclude documents mentioning the year 1990.

Table 9: CSR-L-Chinese FollowIR Code-Switching Query And Instruction Example. It conforms to the original format of MTEB, which uses the same query as query-og and query-changed but has difference in instruction-og and instruction-changed.

	Total Number			Avg. Length		Examples
Dataset	**Q**	$\mathcal{D}$	$\mathcal{D}^+$	**Q**	$\mathcal{D}$	
Touché 2020	49	303,732	34.94	16.39	451.51	Table 11
HumanEval	158	158	1.00	88.48	98.20	Table 12
TRECCOVID	50	171,332	493.46	22.98	223.51	Table 13
FollowIR	208	98,312	30.00	120.86	465.39	Table 14

Table 10: Statistics of datasets in Japanese-CSR-L. Q: number of queries; D: corpus size; D^+: average positive documents per query. Avg. Length is measured in tokens. Examples can be seen in the tables in Appendix.

Query
Do violent video games contribute to 若者の暴力?

Table 11: Japanese-CSR-L Touché 2020 Code-Switching Example.

H Additional Discussion on a Newly Curated Retrieval Benchmark

To further probe whether the CSR-L findings are overly tied to well-known benchmark suites, we add an extra retrieval-only check on *AILACaseDocs*, a dataset that was newly introduced in the recent RTEB leaderboard and is not part of the

Query
Given a string, find out how many distinct characters (大文字・小文字を問わず) does it consist of

Table 12: Japanese-CSR-L HumanEval Code-Switching Query Example.

Query
best masks for Covid-19 感染予防おすすめ

Table 13: Japanese-CSR-L TRECCOVID Code-Switching Query Example.

Query-og
What standards do cruise ships use for 衛生と安全の維持?
Instruction-og
Relevant documents refer to 衛生と安全 practices and standards for レジャークルーズ船. Not relevant are standards for small pleasure craft or commercial freight ships, tankers, etc. Documents referring to a specific ship's problems are not relevant.
Query-changed
What standards do cruise ships use for 衛生と安全の維持?
Instruction-changed
Relevant documents refer to 衛生と安全 practices and standards for レジャークルーズ船, but don't include information about Royal Caribbean or Royal Viking. Not relevant are standards for small pleasure craft or commercial freight ships, tankers, etc. Documents referring to a specific ship's problems are not relevant.

Table 14: Japanese-CSR-L FollowIR Code-Switching Query And Instruction Example. It conforms to the original format of MTEB, which uses the same query as query-og and query-changed but has difference in instruction-og and instruction-changed.

commonly used MMTEB leaderboard. Following the same query-side prompting procedure described for CS-MTEB, we construct a Chinese code-switched version and evaluate *mE5-large* together with *Qwen3-Embedding-0.6B* under the same protocol.

The results in Table 16 show that the same phenomenon persists on this newer benchmark: both models degrade when the query is code-switched, with the drop being substantial for *mE5-large* and smaller but still non-trivial for *Qwen3-Embedding-0.6B*. While we do not claim that *AILACaseDocs* is completely isolated from broader benchmark-ecosystem effects, this additional check reduces the concern that our conclusions are driven solely by in-domain adaptation to a small set of long-standing public evaluation datasets.

I Additional Discussion on Two-Stage Reranking

Because the CSR-L cross-encoder results in the main tables are obtained by direct full-corpus scoring, we additionally test a standard two-stage setup on CSR-L-Chinese. Specifically, we use *Qwen3-Embedding-0.6B* as the first-stage retriever, keep the top-100 candidates for each query, and then rerank them with *jina-reranker-v3* or *Qwen3-Reranker-0.6B*.

The results in Table 17 show that the same code-switching degradation persists under a strong and standard retrieval pipeline: even after retrieving with *Qwen3-Embedding-0.6B* and reranking the top-100 candidates, both rerankers still perform worse on the code-switched queries than on the original English ones. This confirms that the performance drop is not an artifact of the direct full-corpus cross-encoder setup alone.

J Additional Discussion on Non-English Monolingual Baselines

To separate code-switching effects from simply moving away from English, we add a Chinese-centric evaluation in which both the monolingual baseline queries and the document collection are Chinese. Concretely, we use the Chinese subset of *MIRACLRetrievalHardNegatives*, then convert the original Chinese queries into Chinese–English code-switched queries with the same prompting procedure while keeping the documents unchanged.

The results in Table 18 show that the degradation persists even when the monolingual baseline is non-English: all three retrievers perform worse on the code-switched queries than on the original Chinese ones. This indicates that the effect we observe is not merely a consequence of moving away from English as the highest-resource language, but also

Model	Setting	Instr. Rerank (1)	Retrieval (5)	Clust. (1)	Cls. (1)	STS (1)	Rerank (1)	Pair Cls. (1)	Total (11)
e5-large-v2	Original	-0.99	51.78	62.00	73.97	84.55	60.17	59.88	55.91
	Korean	-0.50	37.92	24.78	64.43	55.52	61.25	55.01	42.63
	French	-0.8	44.18	26.29	70.28	57.6	64.97	52.57	45.01
	Italian	-0.68	40.93	24.72	65.14	56.1	60.32	51.63	42.59
	Portuguese	-1.61	43.39	25.68	67.83	56.37	62.72	56.85	44.46
	Dutch	-2.11	39.93	26.28	64.26	55.1	62.02	58.2	43.38
Arctic-Embed-m-v2.0	Original	-3.20	64.21	60.09	64.78	75.97	62.37	58.09	54.62
	Korean	-2.52	55.26	36.7	61.29	55.89	59.21	53.57	45.63
	French	-3.18	60.85	36.53	66.61	57.63	62.71	52.47	47.66
	Italian	-3.53	61.50	37.64	66.85	57.34	62.41	50.21	47.49
	Portuguese	-2.94	61.29	37.19	66.47	56.12	61.89	58.12	48.31
	Dutch	-3.97	58.13	36.63	65.14	54.99	60.02	55.89	46.69
Qwen3-Embedding-0.6B	Original	5.10	73.67	68.21	72.07	91.14	63.09	75.55	64.12
	Korean	2.72	67.34	36.82	84.08	73.17	67.73	59.33	55.89
	French	2.28	67.97	35.33	85.57	73.18	68.46	59.31	56.02
	Italian	3.34	68.85	35.49	83.74	72.29	68.6	58.64	55.85
	Portuguese	4.84	69.55	35.51	84.69	72.2	67.9	64.08	56.97
	Dutch	2.42	65.87	36.25	82.91	70.83	66.32	63.54	55.45

Table 15: CS-MTEB results by model and evaluation setting. Columns correspond to CS-MTEB task categories, with the number of tasks per category in parentheses. The result is the macro average over 7 task categories / Mean (TaskType).

Model	Orig	Chinese	Drop
mE5-large	41.89	23.83	-18.06
Qwen3-Embedding-0.6B	34.80	31.85	-2.95

Table 16: Additional results on AILACaseDocs. Drop is computed as Chinese - Orig.

Model	Touché O	Touché C	HE O	HE C	TREC O	TREC C	FIR O	FIR C	Avg O	Avg C	Drop
jina-reranker-v3	62.04	58.21	98.22	98.07	89.37	84.20	4.65	3.13	63.57	60.90	-2.67
Qwen3-Reranker-0.6B	73.05	67.65	97.33	97.29	91.69	88.36	0.08	0.93	65.54	63.56	-1.98

Table 17: Additional two-stage reranking results on CSR-L-Chinese. O/C denote original and Chinese code-switched queries, respectively. Drop is computed as Avg C - Avg O.

Model	Orig	CS	Drop
jina-embeddings-v3	57.89	50.50	-7.39
Qwen3-Embedding-0.6B	60.19	55.56	-4.63
Arctic-Embed-l-v2.0	61.18	53.30	-7.88

Table 18: Additional results on the Chinese subset of MIRACLRetrievalHardNegatives. CS denotes Chinese–English code-switched queries, and Drop is computed as CS - Orig.

appears in a Chinese-centric retrieval setting where the comparison axis is monolingual Chinese versus Chinese–English code-switching.

K Additional Discussion on Query Quality Verification

To provide a direct quality check for the automatically generated CS-MTEB queries, we manually inspected 50 sampled rewritten queries. Two raters independently scored each query on a 1–10 scale along two axes: *naturalness*, which measures whether the code-switching pattern resembles a plausible bilingual user query, and *information preservation*, which measures whether the rewritten query retains the original information need.

Criterion	Rater 1	Rater 2	Mean
Naturalness	9.02	9.30	9.16
Information Preservation	9.80	9.76	9.78

Table 19: Manual quality check on 50 sampled CS-MTEB rewritten queries. Each query is rated independently by two raters on a 1–10 scale.

As shown in Table 19, the sampled queries receive high scores from both raters on both dimensions. In particular, information preservation is consistently close to the ceiling, indicating that the rewritten queries largely maintain the original search intent, while the naturalness scores also remain high, suggesting that the inserted language switches are generally fluent and plausible. Although this spot check does not replace full-scale human verification, it provides additional evidence that the automatic rewriting procedure yields sufficiently reliable queries for benchmark construction.

L GenAI Statement

This work utilized generative AI tools to assist with formatting, generating LaTeX templates, and refining word choice. The authors reviewed and verified all AI-assisted content to ensure factual accuracy and academic integrity.

M License Statement

In this project, we use the MTEB evaluation framework (Muennighoff et al., 2023), which is released under the Apache License 2.0. Our evaluation datasets are largely accessed through the MTEB suite and their original sources (for example, the Hugging Face Hub); each dataset is used in accordance with its respective license terms.

We also use the following publicly released model checkpoints under their stated licenses: *all-MiniLM-L12-v2* (Apache License 2.0), *e5-large-v2* (MIT License), *Arctic-Embed-m/l-v2.0* (Apache License 2.0), *Qwen3-Embedding-0.6/4/8B* (Apache License 2.0), *jina-reranker-v3* (CC BY-NC 4.0), *bge-reranker-v2-m3* (Apache License 2.0), *Qwen3-Reranker-0.6/4/8B* (Apache License 2.0), and *ColBERT v2* (MIT License). We use these models for research and evaluation purposes and comply with the corresponding license requirements (including non-commercial restrictions where applicable).