MiLQ: Benchmarking IR Models for Bilingual Web Search with Mixed Language Queries

Jonghwi Kim[1], Deokhyung Kang[1], Seonjeong Hwang[1], Yunsu Kim[3]*, Jungseul Ok[1,2], Gary Geunbae Lee[1,2],
[1] Graduate School of Artificial Intelligence, POSTECH, Republic of Korea ,
[2] Department of Computer Science and Engineering, POSTECH, Republic of Korea,
[3] Lilt, Inc. {jonghwi.kim, deokhk, seonjeongh, jungseul.ok, gblee}@postech.ac.kr, yunsu.kim@lilt.com

Abstract

Despite bilingual speakers frequently using mixed-language queries in web searches, Information Retrieval (IR) research on them remains scarce. To address this, we introduce **MiLQ**, **Mi**xed-**L**anguage **Q**uery test set, the first public benchmark of mixed-language queries, qualified as realistic and relatively preferred. Experiments show that multilingual IR models perform moderately on MiLQ and inconsistently across native, English, and mixed-language queries, also suggesting code-switched training data's potential for robust IR models handling such queries. Meanwhile, intentional English mixing in queries proves an effective strategy for bilinguals searching English documents, which our analysis attributes to enhanced token matching compared to native queries.[1]

1 Introduction

Code-switching[2], where bilingual speakers alternate languages within a context, is a prevalent linguistic behavior in multilingual communities (Auer, 1999; Gardner-Chloros, 2009; Auer, 2013). This phenomenon extends to Human-Computer Interaction (HCI), especially via AI agents like ChatGPT (OpenAI, 2023), where understanding mixed-language input critically affects their perceived reliability by bilingual users (Bawa et al., 2020; Choi et al., 2023). Information Retrieval (IR) systems also face the challenge of effectively handling such mixed-language queries (Sitaram et al., 2019).

Meanwhile, recent IR research has expanded beyond Monolingual IR (*MonoIR*) settings to diverse multilingual settings. The benchmarks (Asai et al., 2021; Lawrie et al., 2023b,a; Soboroff, 2023; Adeyemi et al., 2024; Litschko et al., 2025) are widely utilized, representing diverse language scenarios. However, research on mixed-language queries remains sparse and outdated (Fung et al., 1999; Gupta et al., 2014; Sequiera et al., 2015), with no publicly available benchmark.

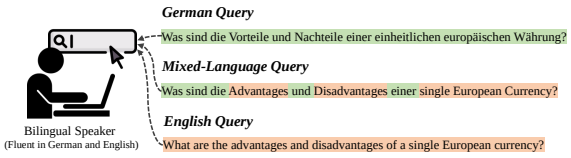


Figure 1: Illustration of a bilingual user freely using German, English, and mixed-language queries. German elements are in green, and English in orange.

To address these gaps, we introduce **MiLQ**, the first **Mi**xed-**L**anguage **Q**uery benchmark created by actual bilingual users (Figure1). Using MiLQ, we explore three main research questions: **(RQ1)** How realistic are our mixed-language queries, and which query language do bilingual users prefer? **(RQ2)** How well do existing multilingual IR models perform in Mixed-language Query Information Retrieval (*MQIR*)? **(RQ3)** Is the behavior of intentionally mixing English terms into query, noted in HCI studies (Fu, 2017, 2019), an effective strategy?

The main contributions of our work are:

- We introduce **MiLQ**, the first public benchmark of mixed-language queries, qualified as realistic and relatively preferred by bilinguals.
- We provide a comprehensive performance analysis of multilingual IR models on **MiLQ**, establishing initial baselines for *MQIR*.
- We show intentionally mixed-language queries are effective for English document retrieval across diverse methods, providing token-level analysis of their rationale.

*This work was done when the author was at aiXplain

[1]The code and data for this work are available at : https://github.com/jonghwi-kim/milq.

[2]In this study, code-switching, mixed-language, and code-mixing are used synonymously.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22655–22671
November 4-9, 2025

Retrieval Scenario (Q→D)	Native Lang (XX)	Num. of Query	Title Query CMI (XX→MiLQ)	Title Query GPT-Eval Acc.	Title Query GPT-Eval Flu.	Title Query Human-Eval Acc.	Title Query Human-Eval Flu.	Title Query Human-Eval Real.	Description Query CMI (XX→MiLQ)	Description Query GPT-Eval Acc.	Description Query GPT-Eval Flu.	Description Query Human-Eval Acc.	Description Query Human-Eval Flu.	Description Query Human-Eval Real.
Mixed→EN	SW	151	8.4→38.6	2.35	2.39	2.83	2.65	2.66	5.6→30.7	2.83	2.44	2.83	2.62	2.61
	SO	151	16.2→59.6	2.38	2.34	2.73	2.58	2.95	5.4→36.1	2.76	2.34	2.63	2.51	2.77
	FI	151	7.3→40.2	2.48	2.52	2.79	2.70	2.59	2.2→45.3	2.63	2.15	2.63	2.44	2.28
	DE	151	9.1→61.8	2.67	2.68	2.61	2.50	2.21	2.1→41.1	2.55	2.11	2.43	2.15	1.80
	FR	151	5.7→35.0	2.52	2.55	2.84	2.51	2.31	2.3→32.9	2.80	2.30	2.84	2.51	2.31
Mixed→XX	ZH	47	0.3→13.7	2.85	2.85	2.79	2.79	2.64	2.4→9.0	2.89	2.91	2.65	2.70	2.50
	FA	45	2.2→15.0	2.98	2.98	2.87	2.82	2.64	0.1→5.6	3.00	2.93	2.91	2.81	2.68
	RU	44	0.0→51.7	2.89	2.50	2.72	2.30	2.16	0.6→51.7	2.93	2.45	2.73	2.16	2.14
Average		**111.4**	**6.2→39.5**	**2.64**	**2.60**	**2.78**	**2.59**	**2.57**	**5.0→31.6**	**2.80**	**2.45**	**2.76**	**2.45**	**2.46**

Table 1: Quality measurements for **MiLQ** (Title & Description queries). Code-Mixing Index (CMI) is on a 0-100 scale (Original Query CMI → Mixed-language Query (**MiLQ**) CMI). For GPT-Eval (Accuracy [Acc.] & Fluency [Flu.]) and Human-Evaluation (Acc. & Flu. & Realism [Real.]), both on a 1-3 scale, cell backgrounds are colored in a fine-grained red gradient from lightest red (scores ≈1.0) to darkest red (scores ≈3.0). The 'Average' row is **bolded**. "XX" denotes the native language.

2 MiLQ: Mixed-Language Query test set

Data Construction We started with queries from two Cross-Language IR (*CLIR*) benchmarks: CLEF (Braschler, 2003) and NeuCLIR22 (Lawrie et al., 2023a), addressing native-to-English and English-to-native retrieval, respectively. These were selected to ensure diverse language scenarios while maintaining quality, based on three criteria: (1) availability of parallel English and native-language queries, (2) widespread use for performance comparison, and (3) budgetary feasibility. Both follow the TREC format (Voorhees, 2005), including short Title and longer Description queries, for which we created mixed-language versions.

Bilingual speakers, experienced in both languages and mixed-language search, crafted natural mixed-language queries from original English and native query pairs, while preserving the original search intent. To reflect realistic code-switching patterns, we adopt Matrix Language Frame theory (Myers-Scotton, 1997) and follow prior studies (Fu, 2017, 2019; Yong et al., 2023; Winata et al., 2023) that describe common code-switching as featuring native language as the grammar-governing matrix and English language as embedded. Accordingly, annotators integrated English terms into the native language structure only when conceptually necessary and linguistically sound. Annotation guidelines are in Appendix A.1, and MiLQ samples are in Appendix A.2 (Figures 5, 6).

Quality Measurement and Analysis We measured MiLQ's quality considering its language mixing, meaning preservation, naturalness, and realism (Table 1). First, for language mixing, we used the **Code-Mixing Index (CMI)** (Das and Gambäck, 2014) (0-100 scale, higher=more mixing; Appendix A.3). Average CMI increased from 6.2 to 39.5 (Title) and 5.0 to 31.6 (Description), showing substantially more mixing than originals. Next, **GPT-Eval** (GPT-4o) using Kuwanto et al.'s framework (high human alignment, Kendall's Tau > 0.5) assessed MiLQ (1-3 scale; rubrics in Appendix A.4) for **Accuracy (Acc.)** (meaning preservation, correct term use) and **Fluency (Flu.)** (naturalness, readability, seamlessness). MiLQ achieved strong average GPT-Eval scores: Acc. 2.64 / Flu. 2.60 (Title) and Acc. 2.80 / Flu. 2.45 (Description). Lastly, for **Human-Eval**, three bilingual annotators per query assessed MiLQ on a 1-3 scale (detailed guidelines in Appendix A.5). This evaluation covered **Accuracy (Acc.)** and **Fluency (Flu.)**, using criteria consistent with GPT-Eval, and an additional **Realism (Real.)**. Realism specifically assessed how naturally bilingual speakers might use the given mixed-language query in real search scenarios. Human evaluators rated MiLQ highly, with average scores: Acc. 2.78 / Flu. 2.59 / Real. 2.57 (Title) and Acc. 2.76 / Flu. 2.45 / Real. 2.46 (Description). These consistently high scores in all metrics affirm the quality and reliability of MiLQ.

Native Lang.	Title Agr.(%)	Title Nat.	Title Mix.	Title Eng.	Description Agr.(%)	Description Nat.	Description Mix.	Description Eng.
SW	78.8	0.44	0.82	2.01	88.7	0.30	1.05	1.68
SO	98.0	0.73	2.26	0.01	100.0	0.19	2.76	0.05
FI	70.9	1.60	1.16	1.26	81.5	1.84	0.34	1.16
DE	53.0	1.11	0.46	1.38	77.5	0.79	0.38	1.85
FR	69.5	1.02	1.86	0.47	78.8	0.82	1.89	0.38
ZH	53.2	1.40	0.92	1.08	70.2	0.88	0.94	1.48
FA	60.0	0.78	2.00	0.37	66.7	0.73	2.00	0.47
RU	72.7	0.59	1.69	0.91	75.0	0.52	1.76	0.79
AVG.	69.3	0.87	1.43	0.89	79.6	0.60	1.54	0.96

Table 2: User preference for Native (Nat.), Mixed-language (Mix.), and English (Eng.) queries. Agr.(%): Percentage of queries where a majority (2+ of 3) of annotators agreed on preferred query type(s). Nat./Mix./Eng. values represent average annotator votes (0-3) for each type. Background color intensity indicates preference strength.

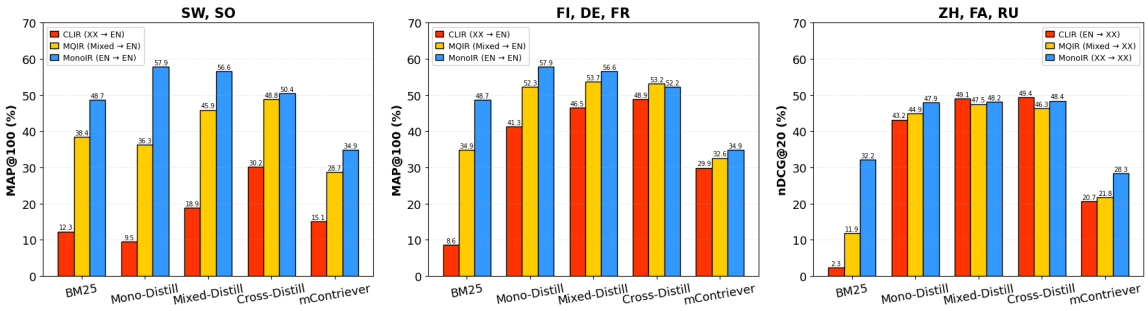


Figure 2: Performance of retrieval models across *CLIR*, *MQIR* (**MiLQ**), and *MonoIR* scenarios. Results are averaged by language group: low-resource (SW, SO; MAP@100) [left], high-resource (FI, DE, FR; MAP@100) [middle], and diverse document language (ZH, FA, RU; nDCG@20) [right]. Models include **BM25**, specialized multi-vector dense retrievers (**Mono-, Mixed-, Cross-Distill**), and **mContriever**. See Appendix B.4 for per-language details.

To investigate user preferences for Native (Nat.), Mixed-language (Mix.), and English (Eng.) query formulations, we asked annotators to select their preferred formulation(s), allowing for multiple selections. For robust assessment, Table 2 presents results for queries in which a majority of annotators (2+ of 3) agreed on their preferred formulation. The scores for each formulation type (0-3) represent the average number of annotators who selected that type as preferred. Overall, Mix. received the highest average scores, with 1.43 for Title and 1.54 for Description queries, outperforming Nat. and Eng. formulations. However, the degree of preference varied across languages. Notably, Somali (SO) exhibited the strongest preference for mixed-language (e.g., Title: 2.26, Description: 2.76). To uncover the reasons for such variations, we conducted interviews with annotators in all languages. These discussions revealed that Somali speakers frequently code-switch, primarily using English to express modern concepts due to Somali's limited contemporary vocabulary—findings aligned with prior literature (Andrzejewski, 1979, 1978; Kapchits, 2019). Further interview insights, including common themes on mixed-language query usage, are provided in Appendix A.6.

In summary, this section addressed **(RQ1)**, confirming that MiLQ is perceived as highly realistic and that bilingual users prefer mixed-language query formulations. Additional details of MiLQ are in Appendix A.7.

3 Experimental Setup

This section details our experimental setup, designed to evaluate various multilingual IR models on mixed-language queries using **MiLQ**.

Test Scenarios & Dataset We evaluate three retrieval scenarios: *MQIR* (**MiLQ**) (Mixed→XX), *MonoIR* (XX→XX), and *CLIR* (XX→YY). Document collections include NeuCLIR22[3] (Lawrie et al., 2023a) and CLEF00-03 (Braschler, 2000, 2002a,b, 2003) (statistics in Appendix B.1). Following prior works (Huang et al., 2023; Yang et al., 2024), queries are concatenations of Title and Description, with MAP@100 and nDCG@20 serving as the primary metrics (detailed in Appendix B.2).

Retrieval Models To create retrieval models specialized for distinct language scenarios, we developed three ColBERT-based (Khattab and Zaharia, 2020) dense retrievers: **Mono-Distill**, **Cross-Distill**, and **Mixed-Distill**. Based on a multilingual pretrained language model, these models are trained via Knowledge Distillation (KD) adapting Translate-Distill strategy (Yang et al., 2024) where English IR training data is translated into target languages. Thus, their specialization for each scenario arises solely from the training data used. **Mono-Distill** is trained for *MonoIR* (e.g., XX→XX or EN→EN) with monolingual query-document pairs (original MSMARCO (Nguyen et al., 2016) or translated version). **Cross-Distill** is trained for *CLIR* (e.g., XX→EN or EN→XX) with cross-lingual query-document pairs derived from MSMARCO. **Mixed-Distill** is trained for *MQIR* (e.g., Mixed→EN or Mixed→XX) with artificially code-switched query-document pairs, generated via bilingual lexicon (Kamholz et al., 2014; Conneau et al., 2017) without translation.

We also include the following baselines: **mContriever** (Izacard et al.) serves as a multilingual single vector dense retriever pre-trained for broad language coverage. **BM25** (Robertson et al., 2009)

[3]https://catalog.data.gov/dataset/2022-neuclir-dataset

is a standard sparse lexical matching retriever. **Translate-Test** first translates queries into the document's language via Neural Machine Translation (NMT), then applies BM25 or Mono-Distill for retrieval. Detailed implementation specifics for all models are in Appendix B.3.

4 Results and Analysis

Main Results In response to **(RQ2)**, MiLQ (*MQIR* in Figure 2) shows that multilingual IR models like Mono-Distill and Cross-Distill achieve moderate performance in *MQIR*, performing between their *MonoIR* and *CLIR* performance. This pattern, also observed with the lexical-based BM25, is attributable to *MQIR*'s intermediate level of lexical cues compared to *MonoIR* and *CLIR* settings.

Further observations underscore specialization's limitations. For instance, Mono-Distill (*MonoIR*-optimized) outperformed Cross-Distill (*CLIR*-optimized) in *MonoIR* settings, and vice-versa. Additionally, mContriever consistently trails specialized models. Notably, Mixed-Distill trained with artificial code-switched text shows well-balanced performance, often outperforming Cross-Distill in *MonoIR* and Mono-Distill in *CLIR/MQIR*. This highlights potential benefits of using mixed-language queries in training for a robust bilingual IR system—a core challenge MiLQ addresses: developing a single robust IR model for bilingual users freely querying in native, English or mixed language. To better harness this potential of code-switched training data explored in prior studies (Litschko et al., 2023; Liu et al., 2025), future work could explore advanced methods, like multilingual LLMs, beyond simple lexicon augmentation.

Regarding **(RQ3)**, intentionally using mixed-language queries offers context-dependent benefits. While native queries are optimal for retrieving native content (*MonoIR*, XX→XX), mixed-language queries (*MQIR*, Mixed→EN) prove superior to native ones (*CLIR*, XX→EN) when bilinguals searching English content, thus offering a clear strategic advantage. Notably, in low-resource *MQIR* for English document retrieval (Figure 2, left), BM25 outperforms neural models like mContriever and Mono-Distill. Consequently, for low-resource languages where neural models struggle with native queries, mixed-language queries with BM25 present a more effective IR system.

Effectiveness of Translate-Test Translate-Test, applying NMT at test time, is widely used in *CLIR*

Method	Low Resource		High Resource		*MonoIR*
	CLIR	*MQIR*	*CLIR*	*MQIR*	
BM25	12.35	38.35	8.56	34.92	48.71
NMT→BM25	41.07	48.10	46.01	47.08	
Mono-Distill	9.53	36.34	41.32	52.26	57.93
NMT→Mono-Distill	50.14	56.92	56.25	56.78	

Table 3: Performance of BM25 and Mono-Distill before and after applying NMT. The metric used is MAP@100 (%).

(Nair et al., 2022). We evaluated its effectiveness for English document retrieval (XX or Mixed → EN), projecting native and mixed-language into English. Table 3 shows introducing NMT for both query types consistently improved performance, bringing them closer to the *MonoIR* scenario. Notably, NMT on mixed-language queries (Mixed→EN) surpassed NMT on native queries (XX→EN). This suggests English terms in mixed queries aid translation, making NMT on these intentionally mixed queries (relevant to **RQ3**) more effective. However, current research on Code-Switching Translation (Huzaifah et al., 2024) has been limited to specific language pairs, underscoring the need for tailored NMT models to better support *MQIR*.

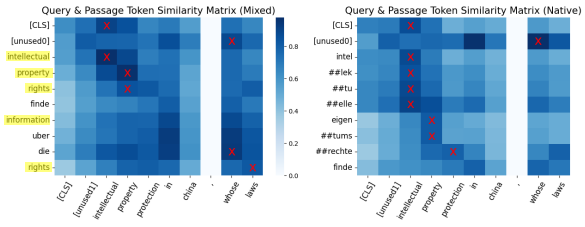


Figure 3: Token-level similarity matrices from Cross-R for German and mixed-language queries on ground truth passage. The y-axis shows tokenized queries (mixed-language left, native right), and the x-axis represents the tokenized English passage. MaxSim tokens are marked by ×, and the code-switched parts are highlighted.

Token-Level Analysis for MQIR The mechanism of multi-vector retriever (e.g., ColBERT) involves identifying the most similar document tokens for each query token. While prior research (Wang et al., 2023; Liu et al., 2024) has explored this in *MonoIR*, its behavior in other language contexts remains unexplored. This token-level analysis offers a rationale for a key aspect of **(RQ3)**: understanding why mixed-language queries can outperform native queries for English document retrieval.

Our analysis compared MaxSim token pair similarity (a query token and its maximal similarity

document token) in mixed-language versus native queries. Figure 3 (left) shows mixed-language queries, by including English terms (e.g., "Intellectual Property Rights" from German "Intellektuelle Eigentumsrechte"), allow these English tokens to form MaxSim pairings (×) with accurate, higher similarity scores. Conversely, native queries (right) rely on cross-lingual interpretation of native tokens (e.g., German "Intellekt," "Eigen") to map English concepts. While MaxSim pairings are also identified (×), this mapping yields weaker similarity for such crucial English concepts. Thus, intentionally mixing English terms improves MaxSim matching through higher similarity scores for English terms—a key rationale (**RQ3**) for *MQIR*'s enhanced English retrieval.

5 Conclusion

This study addressed the prevalent yet understudied phenomenon of mixed-language querying among bilingual speakers by introducing **MiLQ**—the first public user-crafted *MQIR* benchmark, validated for both realism and high user preference. Our comprehensive experiments on MiLQ revealed that current IR models exhibit inconsistent performance across diverse query types, highlighting the need for more robust retrieval systems and demonstrating the promising potential of code-switched training data. Finally, we discovered that intentionally mixing English terms into queries serves as an effective strategy for enhancing English document retrieval among bilingual users.

6 Limitations

While MiLQ is a valuable first public *MQIR* benchmark, it shares limitations common to the broader multilingual IR field. A key challenge is the test set scale; unlike large monolingual English benchmarks (e.g., MS-MARCO (Bajaj et al., 2018), NQ (Kwiatkowski et al., 2019) with thousands of queries), *CLIR* benchmarks typically comprise only tens to hundreds of queries (Asai et al., 2021; Lawrie et al., 2023b,a; Soboroff, 2023; Adeyemi et al., 2024). This is because creating numerous high-quality multilingual test sets is highly resource-intensive. Larger *MQIR* benchmarks would be beneficial, allowing for more robust methodological comparisons and fostering advancements in the field.

MiLQ currently focuses on English-native language pairs, excluding non-English/non-English combinations; future inclusion of these diverse pairings is desirable. Furthermore, while realistic, MiLQ's user-crafted queries may not capture all code-switching patterns, as these are shaped by individual cultural and linguistic experiences. Broader participant involvement could enrich future datasets with more diverse, authentic patterns. Budgetary constraints also limited MiLQ's initial language and domain scope, suggesting future expansions for wider utility.

These limitations and the need for larger test collections highlight promising future directions. Beyond creating larger *MQIR* benchmarks, key research avenues include expanding linguistic diversity (with non-English/non-English pairs), investigating broader code-switching patterns via more diverse annotators, and leveraging advanced techniques like multilingual LLMs to enhance *MQIR*.

Ethical Considerations

Dataset Licensing and Usage Our work uses three primary datasets: NeuCLIR22 (Lawrie et al., 2023a), CLEF00-03 (Braschler, 2003), and our newly introduced MiLQ dataset. We have verified the licensing terms for all existing datasets and ensured our usage is consistent with their intended research purposes. The NeuCLIR22 dataset is published by NIST with public access level and is subject to the NIST Open License. The CLEF00-03 data is distributed by ELDA under an End-User Agreement for Evaluation Packages for Research Use, which permits evaluation purposes. Our MiLQ dataset will be distributed by ELDA under a free evaluation license for academic organizations. The dataset is accessible through our code repository. The MiLQ dataset and our findings are intended solely for academic research purposes in multilingual information retrieval. We discourage commercial deployment without further evaluation of potential societal impacts and biases.

Human Annotation For our MiLQ dataset, queries were created by bilingual speakers and subsequently validated by a different group of bilingual speakers to ensure quality and reduce bias. All participants in both stages were compensated fairly for their work based on regional wage standards and time estimates for each task. The detailed annotation guidelines and compensation structure for each stage are described in Appendix A.1 and A.5. We obtained informed consent from all participants, clearly explaining how their contributions would

be used in research and dataset creation.

Potential Risks The scope of our study is limited to the news domain and nine languages: English, Swahili, Somali, Finnish, German, French, Chinese, Persian, and Russian. Therefore, our findings may not generalize to other domains, genres, or languages not represented in our evaluation. Our findings may inadvertently favor certain language pairs or retrieval approaches that work better for high-resource languages, potentially contributing to digital language divides. Regarding personal information, we followed the existing privacy protection measures established by NIST and ELDA for the original datasets.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2025-RS-2020-II201789, Contribution Rate: 47.5%) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Development of an AI-Based Korean Diagnostic System for Efficient Korean Speaking Learning by Foreigners, Project Number: RS-2025-02413038, Contribution Rate: 47.5%); and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH), Contribution Rate: 5%).

The CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package, ELRA catalogue (http://catalog.elra.info), ISLRN: 317-005-302-361-6, ELRA ID: ELRA-E0008

References

Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, Abdul-Hakeem Omotayo, Idris Abdulmumin, Naome A Etori, Toyib Babatunde Musa, et al. 2024. Ciral: A test collection for clir evaluations in african languages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 293–302.

Bogumił W Andrzejewski. 1978. The development of a national orthography in somalia and the modernization of the somali language. *Horn of Africa*.

BW Andrzejewski. 1979. Language reform in somalia and the modernization of the somali vocabulary. *Northeast African Studies*, pages 59–71.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Peter Auer. 1999. From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. *International journal of bilingualism*, 3(4):309–332.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

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

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Michael Bendersky and Oren Kurland. 2008. Utilizing passage-based language models for document retrieval. In *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30*, pages 162–174. Springer.

Hamed Bonab, James Allan, and Ramesh Sitaraman. 2019. Simulating clir translation resource scarcity using high-resource languages. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 129–136.

Martin Braschler. 2000. Clef 2000—overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 89–101. Springer.

Martin Braschler. 2002a. Clef 2001 — overview of results. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 9–26, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Braschler. 2002b. Clef 2002—overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 9–27. Springer.

Martin Braschler. 2003. Clef 2003–overview of results. In *Workshop of the cross-language evaluation forum for european languages*, pages 44–63. Springer.

Yunjae J Choi, Minha Lee, and Sangsu Lee. 2023. Toward a multilingual conversational agent: Challenges and expectations of code-mixing multilingual users. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–17.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 985–988.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387.

Hengyi Fu. 2017. Query reformulation patterns of mixed language queries in different search intents. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 249–252.

Hengyi Fu. 2019. Mixed language queries in online searches: A study of intra-sentential code-switching from a qualitative perspective. *Aslib Journal of Information Management*, 71(1):72–89.

Pascale Fung, Xiaohu Liu, and Chi-Shun Cheung. 1999. Mixed language query disambiguation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 333–340.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge university press.

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*, pages 677–686.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1048–1056.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, volume 14, pages 3145–3150.

Georgi Kapchits. 2019. on the somali temporal lexicon. *Bildhaan: An International Journal of Somali Studies*, 19(1):7.

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*, pages 39–48.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

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

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2023a. Overview of the trec 2022 neuclir track. *arXiv preprint arXiv:2304.12367*.

Dawn Lawrie, James Mayfield, Douglas W Oard, Eugene Yang, Suraj Nair, and Petra Galuščáková. 2023b. Hc3: A suite of test collections for clir evaluation over informal text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2880–2889.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *arXiv preprint arXiv:2102.10073*.

Robert Litschko, Ekaterina Artemova, and Barbara Plank. 2023. Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data. *arXiv preprint arXiv:2305.05295*.

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.

Andrew Liu, Edward Xu, Crystina Zhang, and Jimmy Lin. 2025. The impact of incidental multilingual text on cross-lingual transfer in monolingual retrieval. In *European Conference on Information Retrieval*, pages 165–173. Springer.

Qi Liu, Gang Guo, Jiaxin Mao, Zhicheng Dou, Ji-Rong Wen, Hao Jiang, Xinyu Zhang, and Zhao Cao. 2024. An analysis on matching mechanisms and token pruning for late-interaction models. *ACM Transactions on Information Systems*, 42(5):1–28.

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

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer.

Shuyo Nakatani. 2010. Language detection library for java.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

OpenAI. 2023. Chatgpt. `https://chat.openai.com`. `https://chat.openai.com`.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, et al. 2015. Overview of fire-2015 shared task on mixed script information retrieval. In *FIRE workshops*, volume 1587, pages 19–25.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Ian Soboroff. 2023. The better cross-language datasets. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3047–3053.

EM Voorhees. 2005. Trec: Experiment and evaluation in information retrieval.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023. Reproducibility, replicability, and insights into dense multi-representation retrieval models: from colbert to col. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2552–2561.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

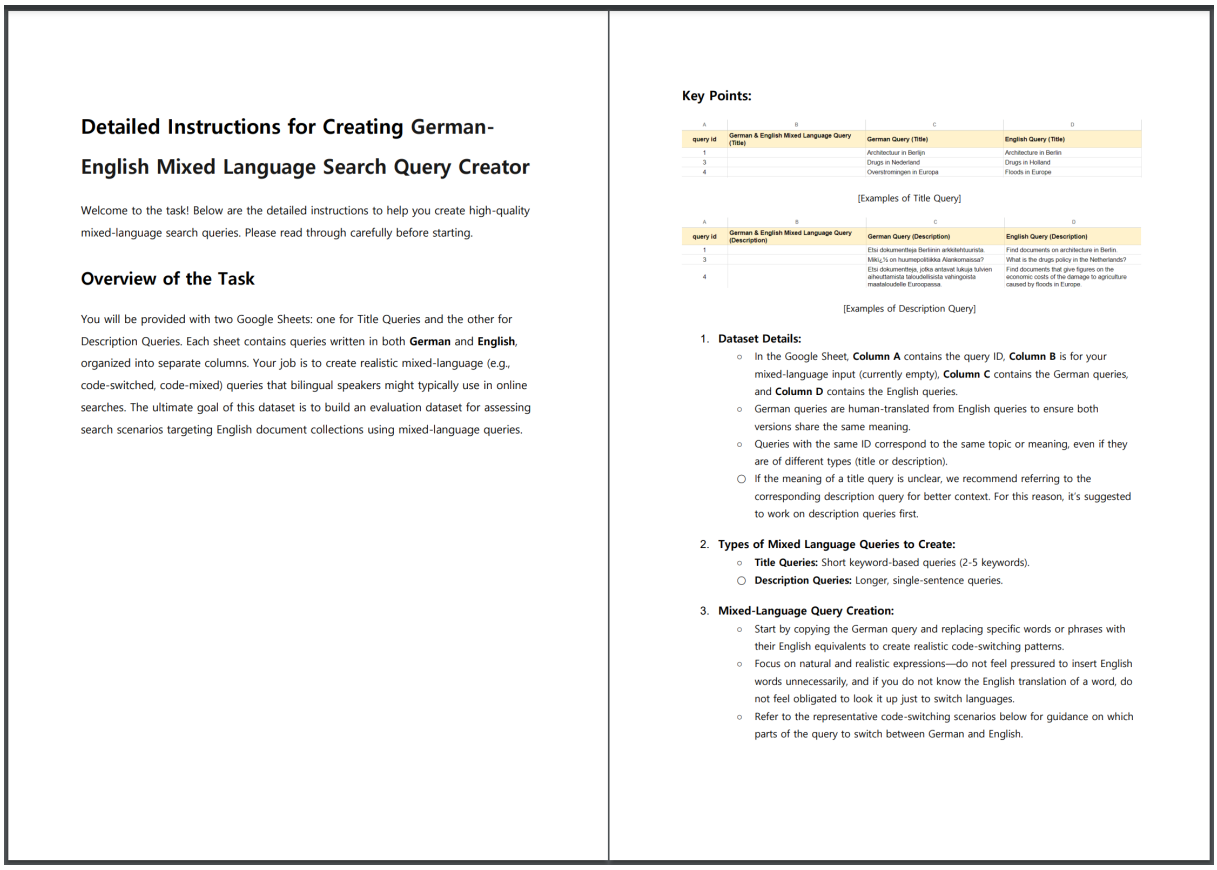
Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *European Conference on Information Retrieval*, pages 50–65. Springer.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, et al. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63.

A Data Annotation

A.1 Details of the Employment and Annotation

We recruited bilingual speakers through Upwork[4], who were fluent in both English and one of the following languages: Swahili (SW), Somali (SO), Finnish (FI), German (DE), French (FR), Chinese (ZH), Persian (FA), or Russian (RU). These annotators were selected based on their proficiency in both languages and their extensive experience in translation activities between English and their respective languages. We provided the annotators with clear guidelines, as shown in Figure 4. The payment was based on the number of queries, with SW, SO, FI, DE, and FR totaling 302 queries (Title + Description) for $40. For ZH, FA, and RU, we created 94, 90, and 88 queries, respectively, with a total cost of $20 per language.



Detailed Instructions for Creating German-English Mixed Language Search Query Creator

Welcome to the task! Below are the detailed instructions to help you create high-quality mixed-language search queries. Please read through carefully before starting.

Overview of the Task

You will be provided with two Google Sheets: one for Title Queries and the other for Description Queries. Each sheet contains queries written in both **German** and **English**, organized into separate columns. Your job is to create realistic mixed-language (e.g., code-switched, code-mixed) queries that bilingual speakers might typically use in online searches. The ultimate goal of this dataset is to build an evaluation dataset for assessing search scenarios targeting English document collections using mixed-language queries.

Key Points:

query id	German & English Mixed Language Query (Title)	German Query (Title)	English Query (Title)
1		Architectuur in Berlijn	Architecture in Berlin
3		Drugs in Nederland	Drugs in Holland
4		Overstromingen in Europa	Floods in Europe

[Examples of Title Query]

query id	German & English Mixed Language Query (Description)	German Query (Description)	English Query (Description)
1		Etsi dokumentteja Berliinin arkkitehtuurista.	Find documents on architecture in Berlin.
3		Mikä on huumepolitiikka Alankomaissa?	What is the drugs policy in the Netherlands?
4		Etsi dokumentteja, jotka antavat lukuja tulvien aiheuttamista taloudellisista vahingoista maatalouteen Euroopassa.	Find documents that give figures on the economic costs of the damage to agriculture caused by floods in Europe.

[Examples of Description Query]

1. **Dataset Details:**
   - In the Google Sheet, **Column A** contains the query ID, **Column B** is for your mixed-language input (currently empty), **Column C** contains the German queries, and **Column D** contains the English queries.
   - German queries are human-translated from English queries to ensure both versions share the same meaning.
   - Queries with the same ID correspond to the same topic or meaning, even if they are of different types (title or description).
   - If the meaning of a title query is unclear, we recommend referring to the corresponding description query for better context. For this reason, it's suggested to work on description queries first.
2. **Types of Mixed Language Queries to Create:**
   - **Title Queries:** Short keyword-based queries (2-5 keywords).
   - **Description Queries:** Longer, single-sentence queries.
3. **Mixed-Language Query Creation:**
   - Start by copying the German query and replacing specific words or phrases with their English equivalents to create realistic code-switching patterns.
   - Focus on natural and realistic expressions—do not feel pressured to insert English words unnecessarily, and if you do not know the English translation of a word, do not feel obligated to look it up just to switch languages.
   - Refer to the representative code-switching scenarios below for guidance on which parts of the query to switch between German and English.

Figure 4: Guideline for German-English mixed-language search query annotators.

[4]https://www.upwork.com

A.2 Examples of Title and Description queries in MiLQ

This appendix illustrates Title and Description mixed-language queries (MiLQ) from our dataset, derived from native and English sources. The figures highlight code-switched segments and indicate their Code-Mixing Index (CMI), calculated by 1.

Source Benchmark	Lang (XX)	Native Title Query		Mixed-Language Title Query (MiLQ)		English Title Query
CLEF00-03	SW	Tatizo ya Mafuta nchini Serbia	(0.0)	Tatizo ya Mafuta nchini Siberia	(20.0)	Siberian Oil Catastrophe
	SO	Masiibada Saliida Saybeeriyaan	(0.0)	Masiibada saliidda Siberian	(33.3)	
	FI	Siperian öljykatastrofi	(0.0)	Siberian öljykatastrofi	(50.0)	
	DE	Ölkatastrophe in Sibirien	(0.0)	Ölkatastrophe in Siberia	(33.3)	
	FR	Catastrophe* petrolière en Sibérie	(25.0)	Catastrophe pétrolière in Siberia	(75.0)	
NeuCLIR22	ZH	巴米揚的大佛像	(0.0)	Bamiyan 大佛像	(50.0)	Buddhas of Bamiyan
	FA	بوداهای بامیان	(0.0)	بوداهای Bamiyan	(50.0)	
	RU	Будды в Бамиане	(0.0)	Бамианские статуи of Buddha	(50.0)	

Figure 5: Examples of Title queries from the MiLQ dataset. Code-switched segments are highlighted, and CMI values are shown in parentheses. (*Note: Although 'Catastrophe' is also a French word, it was identified as English by the language model in this instance.)

Source Benchmark	Lang (XX)	Native Description Query		Mixed-Language Description Query (MiLQ)		English Description Query
CLEF00-03	SW	Pata maelezo kuhusu kupasuka kwa bomba la mafuta nchini Serbia.	(0.0)	Pata maelezo kuhusu kupasuka kwa oil pipeline nchini Siberia	(33.3)	Find information on the rupture of an oil pipeline in Siberia.
	SO	Hel warbixinta dilaaca tuubada saliida ee Saybeeriya.	(0.0)	Hel reports-ka ku saabsan rupturé-ka oil pipeline ee Siberia.	(33.3)	
	FI	Etsi tietoja öljyputken murtumisesta Siperiassa.	(0.0)	Etsi tietoja oil pipeline rupture Siperiassa.	(50.0)	
	DE	Bruch einer Ölpipeline in Sibirien	(0.0)	Information über den Rupture einer Oil Pipeline in Siberia.	(66.7)	
	FR	Rupture d'un pipeline en Sibérie	(0.0)	Rupture d'un oil pipeline en Siberia.	(50.0)	
NeuCLIR22	ZH	我在找有關阿富汗巴米揚大佛的文章。	(0.0)	我在找有關阿富汗 Bamiyan 大佛的文章。	(11.1)	I'm looking for articles on Buddhas of Bamiyan
	FA	دنبال مقالاتی در مورد بوداهای بامیان هستم	(0.0)	دنبال مقالاتی در مورد بوداهای Bamiyan هستم	(14.3)	
	RU	Ищу статьи о Буддах в Бамиане	(0.0)	Ищу статьи on Buddhas of Bamiyan	(66.7)	

Figure 6: Examples of Description queries from the MiLQ dataset, corresponding to the same query IDs as the Title examples shown in Figure 5. Code-switched segments are highlighted, and CMI values are indicated in parentheses.

A.3 Code-Mixing Index (CMI)

The formula for the Code-Mixing Index (CMI) (Das and Gambäck, 2014) is as follows:

$$CMI = \begin{cases} 100 \times \left(1 - \frac{\max(w_i)}{n-u}\right) & \text{if } n > u \\ 0 & \text{if } n = u, \end{cases} \quad (1)$$

where w_i is the word count in language i, $\max(w_i)$ is the word count in the primary language, n is the total word count, and u is the number of language-independent tokens (e.g., numbers, hashtags). In our analysis, we treat the primary language as the native language. We used GPT-4o (Hurst et al., 2024) instead of existing tools for more precise language identification. While existing tools such as language-detection (Nakatani, 2010) and fastText (Joulin et al., 2016) have been widely used for language identification, we observed certain inconsistencies in accuracy. Therefore, we leveraged LLMs for more accurate data analysis. First, we tokenize the text at the word level using NLTK[5]. For Chinese text, we apply Jieba[6], a specialized tokenizer optimized for Chinese word segmentation. After tokenization, we utilize GPT-4o to classify each token's language using the prompt template shown in Figure 7.

[5]https://www.nltk.org/

[6]https://github.com/fxsjy/jieba

Identify the language of each word in the given list using language codes (e.g., ISO 639-1, ISO 639-2, ISO 639-3).
If a word is language-independent (e.g., punctuation, numbers, or symbols), assign it the code 'unknown'.
Return a one-to-one mapping of each word with its corresponding language code, followed by a final list of language codes in order.

Example:

Words: ['I', 'möchte', 'ein', 'new', 'laptop', 'kaufen', ',', 'but', "it's", 'too', 'teuer', '.']

One-to-One Matching:
'I' → 'en'
'möchte' → 'de'
'ein' → 'de'
'new' → 'en'
'laptop' → 'en'
'kaufen' → 'de'
',' → 'unknown'
'but' → 'en'
"it's" → 'en'
'too' → 'en'
'teuer' → 'de'
'.' → 'unknown'

Final List of Language Codes:
['en', 'de', 'de', 'en', 'en', 'de', 'unknown', 'en', 'en', 'en', 'de', 'unknown']

Input:

Words: `{token_list}`

One-to-One Matching:

Figure 7: Prompt template for language identification.

A.4 GPT-Evaluation Rubric

For GPT-based evaluation, we adopted the Accuracy and Fluency rubrics from Kuwanto et al., using their publicly available prompts and evaluation code framework.[7] While their assessments utilized GPT-4O-mini, our study employed the more powerful GPT-4o. The model was instructed to evaluate generated code-switched sentences against the original monolingual sentences on a 1 (lowest) to 3 (highest) scale for each criterion.

Accuracy This criterion measures how well the generated sentence preserves the meaning and information of the original sentence, and whether the code-switched terms are used correctly and appropriately.

Score 1 (Low): Significant deviation from original meaning; key information missing, altered, or redundantly repeated. Code-switched terms incorrect/inappropriate. Introduces new information.

Score 2 (Moderate): Minor deviation from original meaning; most key information present but may have slight errors. Most code-switched terms appropriate with minor mistakes.

Score 3 (High): Fully preserves original meaning; all key information present and correct. Code-switched terms accurate and appropriately used.

Fluency This criterion measures how natural and easy to understand the generated sentence is, considering grammar, syntax, and the smooth integration of code-switching.

Score 1 (Low): Sentence is difficult to understand or awkward; poor grammar/syntax in either language. Code-switching disrupts sentence flow.

Score 2 (Moderate): Sentence is understandable but may have awkward/unnatural phrasing; acceptable grammar/syntax. Code-switching somewhat smooth but not perfectly integrated.

Score 3 (High): Sentence is natural and easy to understand; good grammar/syntax in both languages. Code-switching is smooth and seamless, enhancing flow.

[7] https://github.com/gkuwanto/ezswitch

A.5 Human-Evaluation Guidelines and Rubrics

For the human evaluation of mixed-language queries (MiLQ), we again recruited bilingual speakers via Upwork. Eligibility required proficiency in English and one target language at least at the B2 CEFR level, plus prior translation or linguistic experience, ensuring high-quality judgments. Annotators received detailed instructions (see Figure 8) and evaluated MiLQ quality using three criteria: Accuracy, Fluency, and Realism, rated on a 1-3 scale.

The payment scheme for this evaluation reflected task complexity and language availability: SO and SW annotators were compensated at $20 per annotator; FI, FR, and DE annotators at $30; and FA, ZH, and RU annotators at $15 each.

Accuracy and Fluency Rubrics

Accuracy and Fluency rubrics mirrored those used in GPT-Evaluation (see Appendix A.4). Accuracy measures how well a MiLQ preserves the original query's meaning and appropriately integrates code-switched terms. Fluency assesses the naturalness and clarity of language mixing, ensuring smooth integration of both languages.

Realism This criterion, specific to human evaluation, assesses the likelihood that a bilingual speaker would naturally produce or use the given MiLQ in a real online search context.

Score 1 (Low): Query feels unnatural or forced; unlikely to be used in real search scenarios.

Score 2 (Moderate): Query could be used in real searches, but has noticeable awkwardness or unnatural elements.

Score 3 (High): Query feels natural and comfortable; would likely be used in real search situations.

Somali/English Mixed-Language Query Quality Evaluator

Summary

We are seeking meticulous native Somali speakers who are fluent in English to evaluate the quality and realism of mixed-language (code-switched) search queries. If you regularly mix Somali and English in your online searches, your expertise will be crucial to our research on improving multilingual search experiences.

Task Description

You will be provided with a shared Google Spreadsheet to perform the following evaluations. This work involves assessing a total of 302 queries: 151 'Title' queries (short phrases, typically 2-5 keywords) and 151 'Description' queries (structured as single sentences). Please refer to the 'Examples of Title Queries' and 'Examples of Description Queries' .

query id	Somali Query (Title)	English Query (Title)	Somali & English Mixed Language Query (Title)	Accuracy (1-3)	Fluency (1-3)	Realism (1-3)	Preference (N or M or E) (N=Native, M=Mixed, E=English)
1	Qaabdhismeedka Berlin	Architecture in Berlin	Qaab-dhismeedka Berlin Architecture				
3	Daroogada Holland	Drugs in Holland	Daroogada Holland				
4	Fatahaadaha Yurub	Floods in Europe	Fatahaadaha Europe				
5	Xubinnimada Ururka Yurubiyaanka	European Union Membership	Xubinnimada European Union				

Examples of Title Queries

query id	Somali Query (Description)	English Query (Description)	Somali & English Mixed Language Query (Description)	Accuracy (1-3)	Fluency (1-3)	Realism (1-3)	Preference (N or M or E) (N=Native, M=Mixed, E=English)
1	Hel dokumintiyada qaabdhismeedka Berlin.	Find documents on architecture in Berlin.	Hel dokumintiyada architecture ee Berlin.				
3	Waa maxay seerka daroogada ee Netherlands?	What is the drugs policy in the Netherlands?	Waa maxay drugs policy ee Netherlands?				
4	Hel dokumintiyada bixiyaa qiyaasaha qarashaadka dhaqaalaha ee waxyeelada beeraha ee ay sababaan fatahaada Yurub.	Find documents that give figures on the economic costs of the damage to agriculture caused by floods in Europe.	Hel dokumintiyada bixiyaa figures-ka qarashaadka dhaqaalaha ee waxyeelada beeraha ee ay sababaan floods-ka Yurub.				
5	Aqoonso dabeecadaha wadamada aan xubinta ahayn ee dhinaca ku biirista Bulshada Yurubta ama Ururka Yurubka.	Identify attitudes of non-member countries toward joining the European Community or European Union.	Aqoonso dabeecadaha wadamada aan xubinta ka ahayn ee ku biirista European Community ama European Union.				

Examples of Description Queries

The evaluation process will require you to:

- **Quality Assessment for Mixed-Language Queries**:
  - You will assign a score from 1 to 3 to each mixed-language query based on three criteria: Accuracy, Fluency, and Realism.
  - *Detailed evaluation rubrics for these criteria are provided further below.*
  - **Accuracy**: How well the query preserves the original meaning and uses appropriate code-switched terms.
  - **Fluency**: How natural and understandable the query is, with smooth integration of both languages.
  - **Realism**: How likely a bilingual speaker would naturally use this specific mixed-language query in a real search.
- **Preference Evaluation for Search Use:**
  - Separately, you will be asked to choose the query version(s) you would actually prefer to use in a real online search.
  - Please make quick, instinctive selections based on your typical search habits.
  - If multiple options feel equally suitable, you may select all that apply (e.g., "M, N"). However, please try to choose the one that works best for you.
    1. **N**: Native Language Query
    2. **E**: English Query
    3. **M**: Mixed-Language Query

Evaluation Criteria

1. **Accuracy**
   - **1 (Low)**: Significant meaning deviations; key information missing, altered, or repeated; code-switched terms incorrect/inappropriate; new information introduced.
   - **2 (Moderate)**: Minor meaning deviations; most key information present with slight errors; most code-switched terms appropriate with minor mistakes.
   - **3 (High)**: Fully preserves original meaning; all key information present and correct; code-switched terms accurate and appropriately used.
2. **Fluency**
   - **1 (Low)**: Difficult to understand or awkward; poor grammar/syntax in either language; code-switching disrupts sentence flow.
   - **2 (Moderate)**: Understandable but may have awkward/unnatural phrasing; acceptable grammar/syntax; code-switching somewhat smooth.
   - **3 (High)**: Natural and easy to understand; good grammar/syntax in both languages; code-switching is smooth and seamless.
3. **Realism**
   - **1 (Low Realism)**: Feels unnatural or forced; unlikely to be used in real search scenarios.
   - **2 (Moderate Realism)**: Could be used in real searches, but has noticeable awkwardness or unnatural elements.
   - **3 (High Realism)**:Feels natural and comfortable; would likely be used in real search situations.

General Responsibilities

In performing these core tasks, you will be expected to:

- Strictly adhere to detailed annotation rubrics and guidelines (provided below).
- Provide consistent and objective ratings across all entries.

Requirements

- Native speaker of Somali and fluent in English.
- Familiarity with typical online search behavior and the use of web search engines (e.g., Google Search).
- Experience with, or a clear understanding of, code-switching in everyday communication.
- Excellent attention to detail, with the ability to strictly follow provided guidelines.
- Strong sense of responsibility and adherence to task deadlines.
- (Optional) Previous experience with annotation, data labeling, or evaluation tasks.

Demographic Information Collection

For statistical analysis in our research, we kindly request some basic, anonymized demographic information from selected annotators. This information will be kept strictly confidential and used solely for statistical analysis in the context of our study.

- **Age Range (please select one)**:
  - Under 20 years
  - 20-29 years
  - 30-39 years
  - 40-49 years
  - 50-59 years
  - 60+ years
- **Region (Country of Residence)**
- **Language Proficiency**

Please indicate your proficiency level for both **English** and **Somali** using the scale below. These levels follow the Common European Framework of Reference for Languages (CEFR):

- **A1 (Beginner):** Able to understand and use very basic expressions; limited communication ability.
- **A2 (Elementary):** Able to communicate in simple and routine tasks involving direct information exchange.
- **B1 (Intermediate):** Can handle everyday situations and produce simple connected text on familiar topics.
- **B2 (Upper Intermediate):** Can interact with native speakers with a good degree of fluency and spontaneity.
- **C1 (Advanced):** Able to express ideas fluently and spontaneously without much obvious searching for expressions.
- **C2 (Proficient/Native-like):** Can understand with ease virtually everything heard or read; expresses self effortlessly and precisely.

Figure 8: An example of the detailed annotation guidelines provided to bilingual evaluators, in this case for Somali-English mixed-language search queries. Similar guidelines were adapted for other language pairs.

A.6 Insights from Annotator Interviews on Mixed-Language Query Usage

To gain deeper insights into why and when bilingual users employ mixed-language queries (MiLQ) in real-world online searches, we conducted semi-structured interviews with all annotators. A common question posed was: "*In what situations are mixed-language queries commonly used in real-world online search contexts, and for what reasons?*" Table 4 summarizes the key themes derived from their responses.

Key Theme	Description: Why Users Mix Languages in Queries	Relevant Languages Mentioned
Lexical Gaps	Native language lacks suitable or clear terms for specific concepts (esp. modern/technical), or English equivalents offer greater precision, familiarity, or avoid awkward/ambiguous translations.	Swahili, Somali, French, Finnish, German, Chinese, Persian, Russian
Broader Information Access	English terms are used to retrieve a broader range or larger volume of online results, especially when native-language queries are perceived as too restrictive or yield insufficient/biased information.	Swahili, Somali, Finnish, German, Chinese, Persian
Querying Efficiency	English terms are preferred for faster query formulation, due to shorter terms, keyboard convenience, or greater cognitive accessibility (i.e., English terms come to mind more readily or are more familiar).	Swahili, Somali, French, Finnish, German, Chinese, Russian
Grammatical / Orthographic Simplification	English terms are selected to bypass complex native grammatical constructions (e.g., inflections, case agreement for foreign words) or challenges posed by keyboard layouts and non-Latin scripts.	Russian
Language Modernization	Reliance on established English terms arises from insufficient institutional efforts to standardize native terminology for contemporary concepts, especially in digital and tech domains	Somali

Table 4: Summary of Key Motivations for Mixed-Language Query Usage (Condensed to 5 Points) from Annotator Interviews

In essence, these interviews highlight that bilinguals employ MiLQ for diverse, practical reasons. Key drivers include bridging lexical gaps or seeking terminological precision when native terms are inadequate, especially for modern or technical concepts. Users also mix languages to expand information access, retrieving broader or more diverse results than native-only queries might yield, or to overcome perceived biases. Querying efficiency and fluency are other significant factors, with English often offering faster input or more readily accessible terms. Furthermore, mixed-language can serve to simplify grammatical or orthographic complexities inherent in some native languages, or address deficiencies in language modernization where native terminology for contemporary concepts is lacking.

It is important to note that the specific motivations and patterns of Mixed-language query usage are often highly speaker- and context-dependent, influenced by individual linguistic backgrounds, cognitive habits, the nature of the information need, and even momentary contextual factors. Understanding these varied drivers is crucial for developing IR systems that can effectively cater to the nuanced and dynamic search behaviors of bilingual users worldwide.

A.7 Part-of-Speech Distribution of Code-Switched Words in Queries

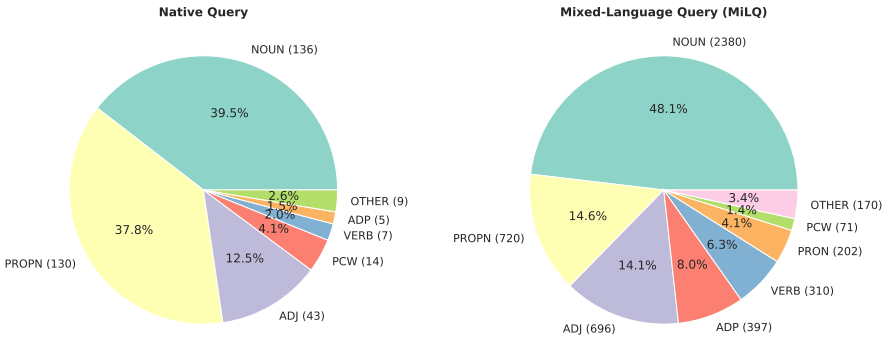


Figure 9: POS distribution of English code-switched words in queries from NeuCLIR22 and CLEF00-03 (left) and MiLQ dataset (right). PCW refers to punctuation-combined words.

The distribution of English words in both native and mixed-language queries predominantly shows that nouns and proper nouns are the most common parts of speech. However, in our MiLQ dataset, nouns outnumber proper nouns, which contrasts with the distribution observed in native queries. Moreover, our dataset exhibits code-switching not only in nouns and proper nouns but also in a broader range of parts of speech, including adjectives, prepositions, verbs, and pronouns, showing a more diverse pattern of code-switching compared to existing datasets.

B Experiment Details

B.1 Benchmark Statistics

	NeuCLIR 22			CLEF00-03					
	ZH	FA	RU	SW	SO	FI	DE	FR	EN
# Queries	47	45	44	151	151	151	151	151	151
# Documents	3.2M	2.2M	4.6M	–	–	–	–	–	113k
# Passages	19.8M	14.0M	25.1M	–	–	–	–	–	1.01M

Table 5: NeuCLIR22 and CLEF00-03 benchmark statistics.

Following previous research (Huang et al., 2023), we use 151 queries from the CLEF C001 – C200 topics, excluding those with no relevant judgments. English documents are sourced from the Los Angeles Times corpus, which includes 113k news articles. For high-resource languages such as Finnish, German, and French, queries are directly provided by the CLEF campaign. In contrast, for low-resource languages, Bonab et al. provided Somali and Swahili translations of English queries.

B.2 Evaluation Metrics

We evaluate retrieval performance using two standard Information Retrieval metrics: **MAP@100 (Mean Average Precision at 100)**: Evaluates ranked lists by averaging precision scores after each relevant binary-judged document is retrieved, up to 100 results. Higher scores indicate better overall retrieval. **nDCG@20 (normalized Discounted Cumulative Gain at 20)**: Assesses ranked lists by measuring cumulative gain from graded-relevance documents within the top 20, discounted by rank and normalized by the ideal gain. Higher scores mean better top-ranking of highly relevant items.

B.3 Implementation Details

Model Configuration For BM25, we utilized the Pyserini toolkit (Lin et al., 2021), which provides reproducible sparse retrieval through Lucene [8]. Dense retrieval experiments with mContriever (Izacard et al.) employed `facebook/mcontriever-msmarco` model[9].

Our primary retrieval experiments use the ColBERT architecture (Khattab and Zaharia, 2020), a multi-vector approach for dense retrieval. We utilized the publicly available PLAID-X implementation[10] for all model training and inference. Consistent with standard ColBERT practices, most training artifacts and hyperparameters were adopted directly. Our primary modification involved setting the maximum document passage length to 180 tokens. Following established methods (Bendersky and Kurland, 2008; Dai and Callan, 2019), documents longer than this threshold were segmented into 180-token passages. During evaluation, the score for each document was determined using the maximum passage score (MaxP) strategy.

All experiments were conducted with a single run. Due to the approximate nearest neighbor (ANN) search employed in ColBERT, experimental results may exhibit minor variations depending on the indexing process. However, we observed that such variations do not lead to substantial differences.

Model Backbones and Computational Resources We fine-tuned distinct ColBERT models for each source benchmark dataset, selecting multilingual Pre-trained Language Model (mPLM) backbones based on practices in prior relevant research (Yang et al., 2024; Huang et al., 2023).

- **For NeuCLIR22 (ZH, FA, RU):** ColBERT was initialized using the XLM-RoBERTa Large model[11]. This model contains approximately 561 million parameters. Fine-tuning for these languages was conducted about 48 hours.
- **For CLEF00-03 (SW, SO, FI, DE, FR):** mBERT-base-uncased[12] was employed as the ColBERT encoder. This model has approximately 179 million parameters. Fine-tuning for these languages took about 24 hours.

All models were trained on a system equipped with four NVIDIA A100-80GB GPUs. During the training process, model is trained with 6 passages for each query.

Hyperparameters A common set of optimization hyperparameters was used for fine-tuning all models. We employed the AdamW optimizer with a learning rate of 5e-6. All models underwent training for 200,000 steps. The total effective batch size was 64, achieved by using a batch size of 16 per GPU across the four GPUs.

[8] https://github.com/castorini/pyserini
[9] https://huggingface.co/facebook/mcontriever-msmarco
[10] https://github.com/hltcoe/ColBERT-X
[11] https://huggingface.co/FacebookAI/xlm-roberta-large
[12] https://huggingface.co/google-bert/bert-base-multilingual-uncased

B.4 Performance in Individual Languages

Query Lang	BM25	Mono-Distil	Mixed-Distil	Cross-Distil	mContriever	NMT→BM25	NMT→Mono-Distil
SW	11.56	15.93	26.64	35.56	24.52	43.70	**51.00**
SW&EN	35.82	38.21	48.04	47.77	32.27	47.76	**56.15**
EN	48.71	**57.93**	56.53	50.77	34.88	48.71	**57.93**
SO	13.14	3.12	11.07	24.91	5.58	38.43	**49.28**
SO&EN	40.88	34.46	43.69	49.77	25.18	48.44	**57.69**
EN	48.71	**57.93**	56.72	49.94	34.88	48.71	**57.93**
FI	7.02	29.65	40.50	41.99	27.23	45.06	**55.14**
FI&EN	40.59	45.87	49.94	49.07	32.23	45.50	**56.20**
EN	48.71	**57.93**	55.73	51.53	34.88	48.71	**57.93**
DE	11.70	44.84	47.16	49.69	30.15	45.70	**56.60**
DE&EN	42.33	56.89	55.43	54.99	32.06	47.58	**57.51**
EN	48.71	**57.93**	56.70	52.12	34.88	48.71	**57.93**
FR	6.95	49.46	51.96	54.95	32.23	47.27	**57.02**
FR&EN	21.84	54.02	55.66	55.59	33.53	48.17	**56.64**
EN	48.71	**57.93**	57.35	53.05	34.88	48.71	**57.93**
CLIR	10.07	28.60	35.41	41.42	23.94	44.03	**53.81**
MQIR	36.29	45.89	50.55	51.44	31.05	47.49	**56.84**
MonoIR	48.71	**57.93**	56.60	51.48	34.88	48.71	**57.93**

Table 6: Performance comparison of different retrieval models across multiple language settings for retrieving English documents. This table presents the performance of individual query languages in this scenario. Additionally, XX&EN represents queries mixing the native language and English. The metric used is MAP@100 (%). The best score(s) for each individual language query type (row) are indicated in **bold**. If there is a unique best score, the second best score(s) are underlined.

Query Lang	**BM25**	**Mono-Distill**	**Mixed-Distill**	**Cross-Distill**	**mContriever**
ZH	25.72	46.82	**49.59**	48.61	32.90
ZH & EN	3.67	41.13	**47.46**	45.48	19.54
EN	5.74	38.52	48.73	**48.91**	21.22
FA	34.29	**48.97**	46.06	47.26	15.26
FA & EN	26.02	**48.69**	45.36	45.39	12.97
EN	0.07	46.29	47.28	**47.99**	11.69
RU	36.56	47.97	48.95	**49.46**	36.86
RU & EN	6.14	44.99	**49.71**	48.02	32.74
EN	1.11	44.66	**51.42**	51.30	29.27
MonoIR	32.19	47.92	48.20	**48.44**	28.34
MQIR	11.94	44.94	**47.51**	46.30	21.75
CLIR	2.31	43.16	49.14	**49.40**	20.73

Table 7: Performance comparison of different retrieval models across multiple language settings for retrieving the native documents. This table presents the performance of individual query languages in this scenario. Additionally, XX&EN represents queries mixing the native language and English. The metric used is nDCG@20 (%). The best score(s) for each individual language query type (row) are indicated in **bold**. If there is a unique best score, the second best score(s) are underlined.