NeuCLIRBench: A Modern Evaluation Collection for Monolingual, Cross-Language, and Multilingual Information Retrieval

Dawn Lawrie [ι] **James Mayfield** [ι] **Eugene Yang** [ι] **Andrew Yates** [ι]
Sean MacAvaney [γ] **Ronak Pradeep** [α] **Scott Miller** [β] **Paul McNamee** [ι] **Luca Soldani** [σ]

[ι] HLTCOE, Johns Hopkins University [γ] University of Glasgow [α] University of Waterloo
[γ] Information Sciences Institute, University of Southern California [γ] Allen Institute for AI

{lawrie, mayfield, eugene.yang, andrew.yates}@jhu.edu

Abstract

To measure advances in retrieval, test collections with relevance judgments that can faithfully distinguish systems are required. This paper presents NeuCLIRBench, an evaluation collection for cross-language and multilingual retrieval. The collection consists of documents written natively in Chinese, Persian, and Russian, as well as those same documents machine translated into English. The collection supports several retrieval scenarios including: monolingual retrieval in English, Chinese, Persian, or Russian; cross-language retrieval with English as the query language and one of the other three languages as the document language; and multilingual retrieval, again with English as the query language and relevant documents in all three languages. NeuCLIRBench combines the TREC NeuCLIR track topics of 2022, 2023, and 2024. The 250,128 judgments across approximately 150 queries for the monolingual and cross-language tasks and 100 queries for multilingual retrieval provide strong statistical discriminatory power to distinguish retrieval approaches. A fusion baseline of strong neural retrieval systems is included with the collection so that developers of reranking algorithms are no longer reliant on BM25 as their first-stage retriever. NeuCLIRBench is publicly available.[1]

1 Introduction

This paper presents NeuCLIRBench, a retrieval test collection over newswire documents that enables evaluation of future first-stage retrieval systems and rerankers for several retrieval tasks over four diverse languages, as shown in Figure 1. A retrieval test collection consists of *documents*, *queries*, and *relevance judgments* indicating which documents are relevant to each query. The 10 million documents in NeuCLIRBench are drawn from natively written Chinese, Persian, and Russian articles that appeared in CommonCrawl News, providing a challenging scenario for first-stage retrieval systems to index given its size. To support the evaluation of English monolingual retrieval, all documents have been translated into English using machine translation. Cross-Language Information Retrieval (CLIR) evaluation in this paper uses English queries and documents in the other three languages. The collection includes manual translations of the queries into Chinese, Persian and Russian; these queries support multi-monolingual experiments, as well as cross-language retrieval experiments over other language pairs (*e.g.,* Chinese queries with Russian documents). Finally, the collection supports evaluation of Multilingual Retrieval approaches by combining the documents from all three languages and using queries expressed in English.

NeuCLIRBench is derived from the work of the TREC Neural Cross-Language Information Retrieval (NeuCLIR) track, which ran from 2022 to 2024. NeuCLIRBench combines the queries from the three years of that track. With over 100 queries per task, there is strong statistical discriminatory power when trying to distinguish retrieval approaches. NeuCLIRBench provides deep judgments for each query so that nearly all relevant documents for a query were judged by a human assessor. It includes a fusion baseline for reranking that combines three modern retrieval approaches: a single-vector dual-encoder, a multi-vector dual-encoder, and a learned-sparse approach. Until now, most rerankers have been evaluated using a BM25 first-stage retriever, exposing powerful rerankers only to documents that have lexical matches with the query. This fusion baseline allows rerankers to be tested on their ability to rank highly relevant documents that lack lexical matches with the query.

The remainder of the paper compares NeuCLIRBench to popular test collections used for these tasks (Section 2); summarizes the main aspects of test collection creation (Section 3); and provides results for monolingual English retrieval, multi-monolingual retrieval, cross-language retrieval, and

[1] https://huggingface.co/datasets/neuclir/bench

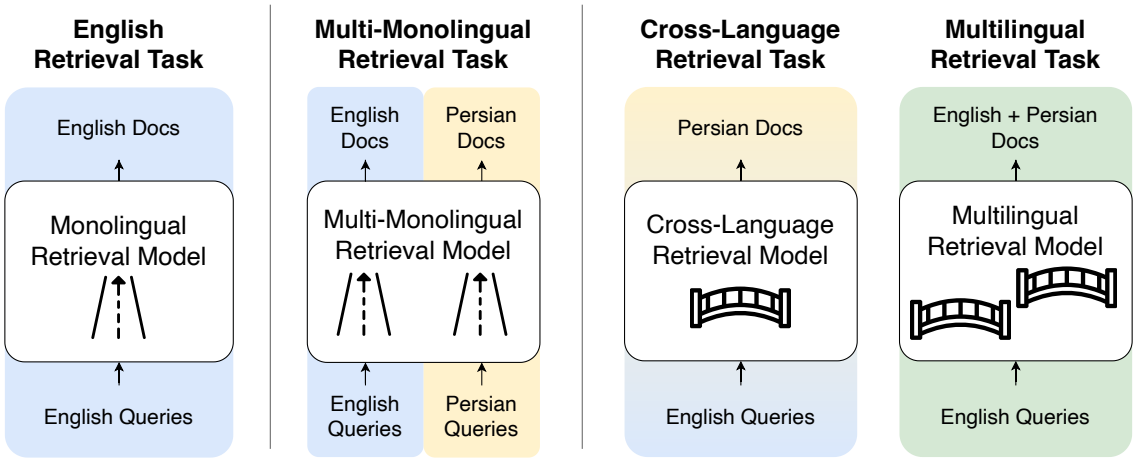

Figure 1: NeuCLIR Tasks

multilingual retrieval (Section 4).

2 Related Work

There has long been an interest in information retrieval systems capable of processing multiple languages, and therefore a longstanding need for corresponding test collections. As an early example, TREC-4 built a Spanish test collection for adhoc retrieval allowing the same system evaluated on both English and Spanish (Harman, 1995). Later, TREC-6 introduced a cross-language retrieval task (Schäuble and Sheridan, 1997) that included three distinct document corpora (English, German, and French) with a unified set of topics translated into each language. This setting enabled benchmarking of both monolingual and cross-language retrieval. Shortly thereafter, benchmarking efforts in non-English monolingual and cross-language retrieval were taken up by other evaluation campaigns, most notably CLEF (Harman et al., 2000), NTCIR (Kando et al., 1999), and FIRE (Majumder et al., 2010). Information retrieval evaluation campaigns eventually focused primarily on monolingual settings and other information access tasks. The cross-language test collections from these campaigns are generally of high-quality, given their use of native bilingual speakers for annotation. However, they still have several problems that motivate our new benchmark: they are comparatively small by modern standards (e.g., TREC-4's Spanish collection contains only 120K documents, compared to the millions typically used in modern evaluations); the labels are not necessarily complete enough to measure today's retrieval systems; and there is a risk of data contamination since Large Language Models are very likely trained on them (Kalal et al., 2024).

A recent trend in information retrieval evaluation is to benchmark monolingual retrieval in many languages at once, without aligning the queries or the documents (Zhang et al., 2021, 2023; Chen et al., 2024). We call such datasets *multi-monolingual*, since they are a collection of many distinct monolingual datasets rather than a true multilingual or cross-language dataset. One example is MIRACL (Zhang et al., 2023), which uses Wikipedia articles in 18 languages and generates queries by asking annotators to create questions that cannot be answered in the first 100 words of an article. Although these benchmarks are useful for measuring how well retrieval systems perform in many individual languages, they cannot be used to evaluate *cross-language* retrieval systems (i.e., with a query written in one language and documents in another).

A common technique relies on machine translation to fill this gap, usually by translating the document corpus of an existing collection from one language into another (Bonifacio et al., 2021). This approach is limited for several reasons. First, machine-translated content is not representative of naturally-occurring content, both in terms of language characteristics ("*translationese*") and in the topicality of the content itself (Graham et al., 2020). Furthermore, translated datasets are not necessarily fit as evaluation data since no human has verified that the translated content is still relevant to the queries. Nevertheless, machine translation plays an important role in training cross-language retrieval systems (Yang et al., 2024c).

Yet another trend is to use "found pairs" of bilingual text to construct cross-language retrieval

benchmarks (Sun and Duh, 2020; Schamoni et al., 2014). For instance, CLIRMatrix (Sun and Duh, 2020) uses cross-language WikiData links between articles to construct queries from article titles in one language and article content from another language. Although this process yields natural text, using document titles as the source of query text causes the dataset to model only basic entity search behavior.

In summary, existing human-annotated cross-language retrieval test collections do not meet modern information retrieval evaluation standards of scale and task complexity, while other techniques (machine translation and found text) face issues with the realism of the text content or the task complexity. Collectively, these limitations motivate us to introduce NeuCLIRBench, a modern cross-language and multilingual retrieval benchmark.

3 Building the Datasets

The NeuCLIRBench test collection has three components: documents, topics, and relevance judgments. We describe each of these components in turn.

3.1 Documents

The document collection, NeuCLIR-1, consists of documents in three languages: Chinese, Persian, and Russian, drawn from CommonCrawl News.[2] The documents were obtained by the CommonCrawl service between August 1, 2016 and July 31, 2021; most of the documents were published within this five-year window. The documents are released as text files without HTML markup. Text was extracted from each source web page using the Python utility *Newspaper*.[3]

The collection is distributed as JSONL, a list of JSON objects, one per line. Each line represents a document. Each document JSON structure consists of the following fields:

id: UUID assigned by CommonCrawl

cc_file: CommonCrawl source file containing the document

timestamp: time of publication (or, if unknown, CommonCrawl download time)

title: article headline or title

text: article body

url: address of the source web page

To ascertain the language of each document, its title and text were independently run through two automatic language identification tools: *cld3*[4], a neural language identification tool; and *VALID* (McNamee, 2016), a compression-based model trained on Wikipedia text. Documents for which the tools agreed on the language, or where one of the tools agreed with the language recorded in the web page metadata, were included in the collection under the language of agreement; all others were removed. This is an imperfect process and some documents comprising text in other languages are included in the collection. The extent of such language pollution is unknown; however, annotators sometimes encountered out-of-language documents in the pools of documents to be assessed. Such documents were always considered to be not relevant. All documents with more than 24,000 characters (approximately ten pages of text) were also removed, as very long documents create assessment challenges. Very short documents were also removed, specifically: Chinese documents containing 75 or fewer characters; Persian documents containing 100 or fewer characters; and Russian documents containing 200 or fewer characters. We observed that such documents are often not genuine news articles, frequently consisting of isolated headlines or commercial advertisements. All remaining documents were deduplicated so that only one instance of each document remained in the collection.

Each collection was limited in size to at most five million documents. After removing duplicates, the Russian collection was significantly above this threshold. Therefore, we used scikit-learn's implementation of random sampling without replacement[5] to downsample the collection. Final collection statistics appear in Table 1.

Since a timestamp is included for each document, the collection can be ordered to support streaming experiments such as in Lawrie et al. (2024a)

3.2 Queries

Queries were developed in batches during the three years that the NeuCLIR Track ran at TREC. All

[2] https://commoncrawl.org/2016/10/news-dataset-available/

[3] https://github.com/codelucas/newspaper

[4] https://pypi.org/project/pycld3/

[5] https://scikit-learn.org/stable/modules/generated/sklearn.utils.random.sample_without_replacement.html

Table 1: Document Collection Statistics for NeuCLIR-1. Token counts from SpaCy (Honnibal et al., 2020).

Language	Document count	Avg. Chars per Document	Median Chars per Document	Avg. Tokens per Document	Median Tokens per Document
Chinese	3,179,209	743	613	427	356
Persian	2,232,016	2032	1427	429	300
Russian	4,627,543	1757	1198	301	204

years used the same general development process, which consisted of an ideation phase, a search phase where documents were examined and judged based on relevance, and a revision phase. There were minor variations in the process over the years based on the language skills of the query creators and the tools available during the search phase.

NeuCLIRBench query creators were instructed to develop traditional TREC-style information needs. These are broader than the queries found in most question-answering evaluation collections, which can be answered with a phrase or sentence. TREC information needs, usually called *topics,* consist of a short title, a sentence-length description, and a paragraph-length narrative. During ideation, query creators write a description indicating what is being searched for, and a narrative that discusses what is and is not considered to be relevant to the topic. During the search phase the query creator uses a search engine to identify a set of documents to judge. The creator counts the number of relevant documents in the first search results. Queries with too many relevant documents are revised or discarded. During the revision phase, the description and narrative may be altered to reflect nuances in document judgment. Finally, the query creator chooses a title that summarizes the topic in three to five words.

All topics coming out of this phase were reviewed by the Track Coordinators; topics that were overly ambiguous, too similar to another topic, or judged by the organizers to be otherwise inappropriate were eliminated. The remaining topics were distributed to track participants. Track submissions were then used during relevance judgment, as discussed in Section 3.3.

As alluded to above, slight changes were made to query creation over the three years of the program. Queries numbered 0 to 199 were each created by a single bilingual query creator. During the search phase, the creator used monolingual search in the language of the collection. The creator counted the number of relevant documents in the top 30 documents ranked using BM25. Any query that had between 1 and 20 relevant documents was added to the set of queries to be reviewed by the coordinators.

Queries numbered 200 to 247 were created by a pair of bilingual query creators, each of whom had language skills in a different document language. In a virtual meeting, the creators brainstormed a topic together. Their instructions were that good topics must "revolve around events, people, and places, and be significant enough to have coverage in more than one language." After exploring the collection with monolingual searches in the two languages, they initiated a single monolingual search in each language. Each creator counted the number of relevant documents in the top 25 documents in their language, ranked using BM25. Creators were instructed to revise the topic if the count was zero or greater than 20 in either of the languages.

Queries numbered 248 to 299 were developed by the Track Coordinators, most of whom lacked language skills in any of the collection's non-English languages. Query creators worked independently to craft topics using the same general three-step process used to create the earlier topics. However, the tools available to them were different. Since these creators entered queries in English rather than in the language of the collection, they were given access to two CLIR search engines: a BM25-based engine that indexed machine translated documents; and a dense retrieval system implemented with ColBERT-X (Nair et al., 2022) that indexed the original documents. The creators' document display included both the original document and the machine translation. The browser could also be used to invoke Google Translate on the original document should the user desire to do so. They could also use HiCAL (Abualsaud et al., 2018) to recommend documents to be judged. Creators made relevance assessments using the same rel-

evance categories:[6] "very valuable," "somewhat valuable," "not that valuable," and "not relevant." They were asked to judge at least thirty documents and find between one and twenty documents in the very or somewhat valuable categories. Document judgments were recorded and added to the judgment pools for final relevance assessments.

Queries 300 to 399 followed a similar approach to that of Queries 200 to 247. A pair of bilingual query creators with skills in different document languages worked together on query ideation. However, rather than judging documents ranked by BM25, their monolingual query returned documents ranked by ColBERT-X (Nair et al., 2022). The switch was made because the BM25 counts did not reliably identify queries with too many relevant documents.

3.3 Relevance Judgments

Each year, TREC NeuCLIR track participants submitted the results of their system runs to NIST. Each query creator with bilingual language skills served as a relevance assessor. The assessors read a subset of the retrieved documents, called the *judgment pool*, in the original language of the document. Thus, each query was assessed by three different assessors when it was judged in all three languages. Pool composition varied slightly over the query sets, but always consisted of at least 20 documents per run. This ensured complete judgments for all runs when using the nDCG@20 metric, the primary metric used to evaluate systems. In the following we describe how the judgment pools were assembled and how relevance was determined.

3.3.1 Creating Judgment Pools

Pools were created from the submitting systems' top-ranked documents. For Queries 0 to 199, the number of documents that a run contributed to a pool was based on whether the submitting team marked the run as a baseline run (see Lawrie et al. (2023) for a list of these runs). Baseline runs contributed their top 25 documents, while non-baseline runs contributed their top 50 documents. For Queries 200 to 299, the top 50 documents for runs that teams prioritized as their top three runs were added to the pools (see Lawrie et al. (2024b) for a list of these runs). For other runs, a depth of twenty was used. For Queries 300 to 399, the top 100 documents for runs that teams prioritized as their top nine runs were included in the pools (see Lawrie et al. (2025) for a list of these runs). For other runs, a depth of 50 was used.

[6] Defined in Section 3.3

3.3.2 Relevance Judgment Process

Assessors used a four-point scale to judge the relevance of each document. Assessors were instructed to provide their assessment as if they were gathering information to write a report on the topic. Relevance was assessed on the most valuable information found in the document; the grades and their associated graded relevance scores were:

Very Valuable: (3 points) information that would appear in the lead paragraph of a report on the topic

Somewhat Valuable: (1 point) information that would appear somewhere in a report on the topic

Not that Valuable: (0 points) information that does not add new information beyond the topic description, or information that would appear only in a report footnote

Not Relevant: (0 points) a document without any information about the topic

3.3.3 Query Analysis

During the assessment periods, 167 Chinese queries, 174 Persian queries, and 171 Russian queries were judged. Within each language some topics had fewer than three relevant documents, while other topics had a concerningly large percentage of relevant documents in the pools. Having topics with fewer than three relevant documents can have undesirable effects on the ability to statistically distinguish systems; they have been removed from NeuCLIRBench. There are 16 Chinese topics, 27 Persian topics, and 16 Russian topics with fewer than three relevant documents. Thus each language has at least 147 topics in NeuCLIRBench.

The requirements for the MLIR task in NeuCLIRBench are slightly different. Instead of requiring at least three relevant documents in each language, the requirement is applied across all three languages; thus, it is possible for a query to have no relevant documents in a particular language as long as the query was assessed in that language and the other two languages combine to have a sufficient number of documents. Only one query was dropped for having too few relevant documents; this resulted in 105 topics for the monolingual English and the multilingual retrieval tasks.

Identifying and eliminating topics likely to contain many unidentified relevant documents is important for future use of the collection. One approach simply calculates the percentage of relevant documents in the pool and sets a cutoff (such as 20% prevalence) above which we cannot be confident that the relevant set is sufficiently identified. Using this cutoff would flag 10 topics in Chinese, 9 topics in Persian, and 14 topics in Russian. Another way of investigating this issue is by examining the occurrence of unjudged documents among those used to calculate the score (in this case those that are in the top 20). The only documents that could possibly affect the score of a system are unjudged documents that are relevant but are treated as non-relevant. As long as the percentage of judged documents in the top results remains high for future systems, those systems will be scored fairly. If that percentage drops significantly, it is not possible to determine whether systems are being scored fairly without additional judgments. We encourage users of the collection to report judged@20 averages when reporting results.

3.4 Resources

NeuCLIRBench relies on the NeuCLIR-1 document set, which comprises documents written in Chinese, Persian, and Russian. The collection also includes updated document translations into English. Documents were re-translated using a vanilla Transformer model that was trained in-house with the *Sockeye* version 2 toolkit (Domhan et al., 2020) using recently released bitext obtained from publicly available corpora.[7] The number of sentences used in training is given in Table 2, along with BLEU scores on the FLORES-200 and NTREX-128 benchmarks for each language (NLLB Team et al., 2022; Federmann et al., 2022).

Canonicalized query files are in tab-separated format with the query ID followed by the query. Queries in NeuCLIRBench are the concatenation of the title and description. Each query file is written in a single language. All queries were first expressed in English, then manually translated into the document languages to produce queries that naturally express the meaning of the original query in those languages. An additional query file is also released in JSONL format containing the narratives written to describe document relevance, as well as translations of the titles, descriptions, and narratives using GoogleTranslate (translated in the summer of the year it appeared at TREC). These translations support CLIR experiments where the underlying retrieval approach expects queries in the same language as the documents; thus, they are a means for crossing the language barrier.

Table 2: MT training data used and BLEU scores on FLORES-200 and NTREX-128 benchmarks.

Language	# Sents	FLORES	NTREX
Chinese	127M	32.7	29.3
Persian	26M	39.4	30.5
Russian	194M	36.2	29.4

The relevance judgments are released in four qrels[8] files, one per language plus one covering all three languages. The non-English ones can be used for CLIR and multi-monolingual experiments; the MLIR file is used for monolingual English and multilingual retrieval experiments.

4 Evaluation on NeuCLIRBench

In this section, we provide a comprehensive set of results from commonly-reported retrieval models in prior works. The full set of model descriptions can be found in Appendix A. These models include dense and sparse neural bi-encoders and rerankers. While this is not a complete list of all publicly available multi-monolingual, cross-language, and multilingual retrieval models, it includes representatives of each type of model as well as the state-of-the-art effectiveness of each as reported in the literature.

4.1 Multi-Monolingual Retrieval

Multi-monolingual retrieval results over the four languages are summarized in Table 3. The top section of the table reports the results for first-stage retrieval, while the bottom section reports results of reranking the strong Fusion baseline. The fusion run is a combination of PLAID-X (Yang et al., 2024b), MILCO (Nguyen et al., 2025), and Qwen3 8B Embed (Zhang et al., 2025) using reciprocal rank fusion (Cormack et al., 2009). These three models are representatives of multi-vector dense, learned-sparse, and single-vector dense retrieval models that are also multilingual. We include other models in the table for reference. The effectiveness of the fusion is the highest reported for the first stage retrievers, indicating that the reranking task is challenging as it starts with a very strong ranking.

[7] https://opus.nlpl.eu

[8] https://trec.nist.gov/data/qrels_eng/

Table 3: Multi-Monolingual Retrieval Results. Since SPLADEv3 is an English-only model we report its results on only the English retrieval task. Models in each group are sorted by the average nDCG@20.

	nDCG@20					Judged@20			
	Chinese	Russian	Persian	English	Avg	Chinese	Russian	Persian	English
BM25	0.391	0.408	0.391	0.349	0.385	1.00	1.00	1.00	0.97
Bi-Encoders									
RepLlama	0.326	0.335	0.110	0.350	0.280	0.73	0.79	0.31	0.91
e5 Large	0.323	0.309	0.348	0.293	0.318	0.69	0.69	0.68	0.80
BGE-M3 Sparse	0.277	0.289	0.386	0.326	0.319	0.71	0.65	0.77	0.85
Qwen3 0.6B Embed	0.481	0.431	0.434	0.408	0.439	0.89	0.88	0.80	0.95
PLAID-X	0.461	0.450	0.490	0.388	0.447	0.88	0.90	0.89	0.77
SPLADEv3	–	–	–	0.420	–	–	–	–	0.97
MILCO	0.420	0.461	0.494	0.423	0.449	0.78	0.87	0.82	0.90
Arctic-Embed Large v2	0.470	0.454	0.493	0.414	0.458	0.89	0.89	0.88	0.96
JinaV3	0.477	0.469	0.495	0.398	0.460	0.89	0.88	0.87	0.94
Qwen3 4B Embed	0.541	0.519	0.550	0.431	0.510	0.94	0.94	0.93	0.96
Qwen3 8B Embed	0.543	0.526	0.565	0.429	0.516	0.95	0.95	0.94	0.96
Fusion	0.577	0.564	0.609	0.483	0.558	0.98	0.98	0.97	0.99
Pointwise Rerankers on *Fusion*									
Qwen3 0.6B Rerank	0.545	0.513	0.507	0.423	0.497	0.98	0.98	0.96	0.99
Jina Reranker	0.565	0.506	0.504	0.468	0.511	0.97	0.97	0.96	0.98
SEARCHER Reranker	0.534	0.519	0.512	0.482	0.512	0.92	0.95	0.92	0.99
Mono-mT5XXL	0.572	0.553	0.592	0.468	0.546	0.98	0.99	0.98	0.99
Qwen3 4B Rerank	0.569	0.580	0.611	0.494	0.563	0.98	0.98	0.97	0.99
Qwen3 8B Rerank	0.595	0.597	0.620	0.511	0.581	0.98	0.98	0.97	0.99
Rank1	0.647	0.604	0.629	0.503	0.596	0.95	0.95	0.95	0.98
Listwise Rerankers on *Fusion*									
RankZephyr 7B	0.522	0.432	0.552	0.435	0.485	0.97	0.95	0.97	0.99
FIRST Qwen3 8B	0.545	0.598	0.615	0.518	0.569	0.97	0.98	0.97	0.99
RankQwen-32B	0.640	0.613	0.657	0.517	0.607	0.98	0.98	0.98	0.99
Rank-K (QwQ)	0.653	0.642	0.659	0.539	0.623	0.97	0.98	0.98	0.99

The reranking results are in the bottom section of the Table, which includes a number of popular and state-of-the-art pointwise and listwise rerankers. Several rerankers, such as Qwen3 0.6B Rerank (Zhang et al., 2025), Jina Reranker (Sturua et al., 2024), and RankZephyr (Pradeep et al., 2023), fail to improve the initial fusion ranking in both English and the multi-monolingual average (the "avg" column). Despite strong effectiveness reported on other benchmarks, which mostly rerank BM25 results, these models fail to improve a very strong initial ranking; this indicates that more development is needed for reranking models to distinguish relevant documents from closely related non-relevant documents.

A majority of these runs have Judged@20 values higher than 0.9; even the zero-shot RankQwen-32B model is above 0.95. In the worst case, we observe these numbers dipping slightly below 0.8. These high judgment rates at the top of the rank indicate strong reusability of NeuCLIRBench on all languages, especially since the majority of the reported models were developed later and were not part of the pooling when judgments were created.

4.2 Cross-Language and Multilingual Retrieval

We report results in Table 4 on the same set of retrieval models for the cross-language and multilingual retrieval tasks. While these results exhibit similar trends, the detailed model ordering is slightly different, providing interesting insights For example, Qwen3 8B Reranker is more effective than its 4B variant in the monolingual task across all four languages, although they are similar in the cross-language tasks. The 4B variant is even higher than the 8B one in the multilingual retrieval task, indicating that the effectiveness gain from using a larger model observed in prior work might not hold when multiple languages are involved in the same task.

Multilingual retrieval is an even more challenging task for the models, as the differences between models are smaller. While Rank1 (Weller et al., 2025) is noticeably more effective than any other pointwise reranker in the cross-language task, it is

Table 4: Cross-Language and Multilingual Retrieval Results. While it is possible to concatenate the translation of the queries from all three languages to support BM25 with query translation (BM25 w/ QT) in the multilingual setting, it would not be a fair comparison as the length of the query is tripled. For consistency, we still report BGE-M3 Sparse here, but it is not designed for scenarios where the queries and documents do not share the same vocabulary. Models in each group are ordered by the nDCG@20 on the multilingual retrieval task.

	Cross-Language Retrieval							**Multilingual Retrieval**	
	nDCG@20				Judged@20			nDCG@20	Judged@20
	Chinese	Russian	Persian	Avg	Chinese	Russian	Persian		
BM25 w/ QT	0.365	0.384	0.380	0.376	1.00	1.00	1.00	–	–
BM25 w/ DT	0.439	0.400	0.447	0.429	0.96	0.97	0.97	0.349	0.97
Bi-Encoders									
BGE-M3 Sparse	0.089	0.051	0.044	0.061	0.24	0.20	0.16	0.054	0.30
e5 Large	0.269	0.296	0.302	0.289	0.59	0.68	0.64	0.205	0.70
RepLlama	0.353	0.391	0.133	0.292	0.78	0.83	0.44	0.255	0.86
Qwen3 0.6B Embed	0.468	0.433	0.366	0.422	0.86	0.86	0.74	0.309	0.92
Arctic-Embed Large v2	0.435	0.445	0.446	0.442	0.83	0.87	0.82	0.352	0.92
JinaV3	0.434	0.468	0.464	0.455	0.82	0.86	0.81	0.355	0.93
MILCO	0.431	0.476	0.494	0.467	0.81	0.88	0.81	0.395	0.93
PLAID-X	0.495	0.463	0.529	0.495	0.93	0.95	0.93	0.396	0.98
Qwen3 4B Embed	0.546	0.526	0.544	0.538	0.93	0.92	0.91	0.415	0.97
Qwen3 8B Embed	0.551	0.545	0.568	0.555	0.94	0.94	0.92	0.423	0.97
Fusion	0.573	0.581	0.617	0.590	0.98	0.99	0.98	0.468	0.99
Pointwise Rerankers on *Fusion*									
Qwen3 0.6B Rerank	0.534	0.481	0.452	0.489	0.97	0.97	0.94	0.368	0.99
Jina Reranker	0.521	0.473	0.415	0.469	0.96	0.96	0.92	0.373	0.98
Mono-mT5XXL	0.550	0.536	0.568	0.551	0.98	0.98	0.97	0.445	0.99
SEARCHER Reranker	0.578	0.544	0.559	0.560	0.95	0.96	0.94	0.468	0.99
Qwen3 8B Rerank	0.599	0.582	0.587	0.590	0.98	0.98	0.97	0.475	0.99
Qwen3 4B Rerank	0.582	0.583	0.605	0.590	0.98	0.98	0.97	0.487	0.99
Rank1	0.649	0.618	0.619	0.628	0.95	0.96	0.94	0.488	0.99
Listwise Rerankers on *Fusion*									
RankZephyr 7B	0.552	0.441	0.568	0.520	0.98	0.96	0.97	0.424	0.99
FIRST Qwen3 8B	0.544	0.591	0.602	0.579	0.98	0.98	0.97	0.449	0.99
RankQwen-32B	0.634	0.625	0.661	0.640	0.98	0.99	0.98	0.480	0.99
Rank-K (QwQ)	0.655	0.636	0.671	0.654	0.98	0.98	0.98	0.506	0.99

barely different from Qwen3 4B Reranker in the multilingual task. The gap between Rank1 (the best pointwise reranker) and Rank-K (the best listwise reranker) is also smaller in the multilingual task than in the cross-language task, also indicating that ranking documents in different languages is more challenging.

Since most multilingual models are trained monolingually on many languages, it is unclear how well these models would perform when the task involves multiple languages, such as cross-language and multilingual retrieval. In fact, based on our reported evaluation results, multilingual retrieval is noticeably harder for multilingual models. We challenge the community to explore more challenging retrieval and multilingual scenarios to further drive the development of multilingual modeling and retrieval.

5 Conclusion

NeuCLIRBench synthesizes the test collection creation work done as part of TREC NeuCLIR 2022-2024. It provides a new test set for evaluating cross-language, multi-monolingual, and multilingual retrieval systems with a large, recent document collection covering three diverse languages, many queries with relevant documents in each of those languages, and relevance judgments we believe to cover almost all relevant documents. It includes high quality machine translations of all documents into English, and manual translations of the English queries into the three document languages. It also includes a new baseline run that fuses three strong first-stage retrievals, providing a strong starting point for rerankers. We believe these features make NeuCLIRBench durable enough to facilitate a wide variety of retrieval experiments for many years to come.

References

Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. 2018. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1317–1320. ACM.

Shantanu Agarwal, Joel Barry, and Scott Miller. 2025. ISI's SEARCHER II System for TREC's 2024 NeuCLIR Track. In *Proceedings of The Thirty-Third Text REtrieval Conference*. NIST.

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.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the MS MARCO passage ranking dataset.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto A. Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Zijian Chen, Ronak Pradeep, and Jimmy Lin. 2025. Accelerating listwise reranking: Reproducing and enhancing FIRST. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3165–3172, New York, NY, USA. Association for Computing Machinery.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. 2022. Patapasco: A Python framework for cross-language information retrieval experiments. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, page 276–280, Berlin, Heidelberg. Springer-Verlag.

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.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The Sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. *IEEE Transactions on Big Data*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 72–81. Association for Computational Linguistics.

Donna Harman. 1995. Overview of the fourth text retrieval conference (TREC-4). In *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*, volume 500-236 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Donna Harman, Martin Braschler, Michael Hess, Michael Kluck, Carol Peters, Peter Schäuble, and Paraic Sheridan. 2000. CLIR evaluation at TREC. In *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers*, volume 2069 of *Lecture Notes in Computer Science*, pages 7–23. Springer.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Nasreen Jaleel, James Allan, W. Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and hard. In *Proceedings of the Thirteenth Text REtrieval Conference*. NIST.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,

Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint 2310.06825*.

Vishakha Suresh Kalal, Andrew Parry, and Sean MacAvaney. 2024. Training on the test model: Contamination in ranking distillation. *CoRR*, abs/2411.02284.

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Soichiro Hidaka. 1999. Overview of IR tasks. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, NTCIR-1, Tokyo, Japan, August 30 - September 1, 1999*. National Center for Science Information Systems (NACSIS).

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. SPLADE-v3: New baselines for SPLADE. *arXiv preprint arXiv:2403.06789*.

Dawn Lawrie, Efsun Kayi, Eugene Yang, James Mayfield, and Douglas W. Oard. 2024a. PLAID SHIRTTT for large-scale streaming dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2574–2578, New York, NY, USA. Association for Computing Machinery.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the TREC 2022 NeuCLIR track. In *Proceedings of The Thirty-First Text REtrieval Conference*. NIST.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2024b. Overview of the TREC 2023 NeuCLIR track. In *Proceedings of the Thirty-Second Text REtrieval Conference*. NIST.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2025. Overview of the TREC 2024 NeuCLIR track. *arXiv preprint arXiv:2509.14355*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2022. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning LLaMA for multi-stage text retrieval. *arXiv:2310.08319*.

Prasenjit Majumder, Dipasree Pal, Ayan Bandyopadhyay, and Mandar Mitra. 2010. Overview of FIRE 2010. In *Multilingual Information Access in South Asian Languages - Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*, volume 7536 of *Lecture Notes in Computer Science*, pages 252–257. Springer.

Paul McNamee. 2016. Language and dialect discrimination using compression-inspired language models. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 195–203, Osaka, Japan. The COLING 2016 Organizing Committee.

Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A language processing toolkit for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*.

Thong Nguyen, Yibin Lei, Jia-Huei Ju, Eugene Yang, and Andrew Yates. 2025. MILCO: Learned sparse retrieval across languages via a multilingual connector. *arXiv preprint 2510.00671*.

Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A unified framework for learned sparse retrieval. In *European Conference on Information Retrieval*, pages 101–116. Springer.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv:2312.02724*.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and

Heng Ji. 2024. FIRST: Faster improved listwise reranking with single token decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652, Miami, Florida, USA. Association for Computational Linguistics.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 488–494. The Association for Computer Linguistics.

Peter Schäuble and Paraic Sheridan. 1997. Cross-language information retrieval (CLIR) track overview. In *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, volume 500-240 of *NIST Special Publication*, pages 31–43. National Institute of Standards and Technology (NIST).

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *arXiv preprint 2409.10173*.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4160–4170. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. In *Second Conference on Language Modeling*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024b. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2368–2373.

Eugene Yang, Dawn J. Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. 2024c. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II*, volume 14609 of *Lecture Notes in Computer Science*, pages 50–65. Springer.

Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. 2025b. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint 2505.14432*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel F Campos. 2025. Arctic-embed 2.0: Multilingual retrieval without compromise. In *Second Conference on Language Modeling*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

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. *arXiv preprint arXiv:2506.05176*.

Appendix

A Model Descriptions

A.1 Sparse Retrieval

Sparse retrieval methods perform lexical matches between the query and the document. Based on the lexical matches, a score is produced for each retrieved document. We conduct experiments with BM25 using the Patapsco framework (Costello et al., 2022). Patapsco is a framework built on top of Pyserini (Lin et al., 2021) specifically designed for experiments across many languages. Given the reliance on lexical matches, the input query and documents must be in the same language. Some experimental settings such as tokenenization and token normalization are language-specific, while others are language agnostic. The language agnostic settings include the BM25 hyperparameter settings where $k1 = 0.9$ and $b = 0.4$ following standard conventions. The RM3 query expansion technique (Jaleel et al., 2004) added ten words based on the top ten documents, weighting terms from the original query and the expansion terms equally.

For the English retrieval task, the spaCy tokenizer and stemmer (Honnibal et al., 2020) are used to process the queries and documents, which have stopwords removed. For the multi-monolingual retrieval task, all characters were normalized. spaCy was used in all languages to perform tokenization. For Chinese, no stemming was applied. For Russian, the spaCy stemmer was used, while for Persian, the Parsivar (Mohtaj et al., 2018) stemmer was used. For the cross-language task, machine translation was used to cross the language barrier. BM25-DT indicates that the documents are translated into English and then the English setting is used. BM25-QT uses machine translation to translate the query into the language of the documents, then the mononlingual setting for that language is followed. In the multilingual task, only the document translation approach is applicable. Query translation produces several versions of the query and scores from the different languages are not comparable; thus these scores cannot be used to rank documents across languages.

A.1.1 Multi-Dense Vector Retrieval

We use PLAID-X (Yang et al., 2024c,b), a cross-language and multilingual variant of the ColBERT model. PLAID-X is fine-tuned from an XLM-RoBERTa Large model with cross-language distillation from an mT5 reranker (described below) using MS-MARCO v1 training queries and passages. Each document is separated into passages using a sliding window of size 180 tokens with a stride of 90. At inference time, document scores are aggregated from the passage scores using MaxP (Dai and Callan, 2019).

A.1.2 Learned Sparse Retrieval

Learned Sparse Retrieval (LSR) methods represent queries and documents as sparse vectors where each dimension is associated with a term in a vocabulary. Some LSR methods produce vectors containing only terms from the input text, while others use a Masked Language Modeling head to also perform term expansion (Nguyen et al., 2023). To efficiently use these representations for first-stage retrieval, they can be stored in an inverted index. We conduct experiments with several LSR methods using Anserini (Yang et al., 2017), which is built on Lucene.

- BGE-M3-Sparse (Chen et al., 2024) is an LSR method trained as part of the multi-granularity M3-Embedding model. It is based on an XLM-RoBERTa backbone with 0.6B parameters. This method does not perform expansion, so it can only be used for monolingual and multi-monolingual retrieval.
- MILCO (Nguyen et al., 2025) is a multilingual LSR method that performs both query expansion and document expansion. MILCO performs multilingual retrieval by producing sparse vectors that are tied to an English vocabulary regardless of the input language, so the model implicitly translates non-English inputs. It is based on an XLM-RoBERTa backbone with 0.6B parameters.
- SPLADE-v3 (Lassance et al., 2024) is a monolingual LSR method that performs both query expansion and document expansion. It is trained on only English MS MARCO data (Bajaj et al., 2018), so it can only be used for monolingual retrieval in English.

A.1.3 Dense Retrieval

Dense Retrieval (DR) methods represent queries and documents as a single dense vector in a latent space (Lin et al., 2022). To efficiently use these representations for first-stage retrieval, they can be stored in a vector index that supports approximate nearest neighbor search, such as FAISS (Douze et al., 2025). We use fast scan product quantization with FAISS where the number of codes is set to half the embedding dimensionality and codes are 4 bits each.

- Arctic Embed v2 (Yu et al., 2025) is a multilingual model based on a XLM-RoBERTa backbone with 0.6B parameters. We use the `snowflake-arctic-embed-l-v2.0` checkpoint, which is the largest available Arctic Embed model.
- E5 (Wang et al., 2024) is a multilingual model based on an XLM-RoBERTa backbone with 0.6B parameters. We use the `multilingual-e5-large-instruct` checkpoint.
- Jina v3 (Sturua et al., 2024) is a multilingual model based on an XLM-RoBERTa backbone with 0.6B parameters.
- Qwen3-Embedding (Zhang et al., 2025) is a family of embedding models trained with synthetic data on top of Qwen3 backbones (Yang et al., 2025a). We use all three model sizes: 0.6B, 4B, and 8B.
- RepLlama (Ma et al., 2023) is a dense retrieval model based on LLaMA-2-7B. We use the `repllama-v1-7b-lora-passage` checkpoint trained on MS MARCO passages.

A.1.4 First-Stage Retrieval Fusion

We use reciprocal rank fusion (RRF) (Cormack et al., 2009) to combine the rankings produced by three strong first-stage retrieval models from the three different families described above: PLAID-X, MILCO, and Qwen-Embedding 8B. Following common practice and Cormack et al. (2009), we set $k = 60$ and do not use per-method weights.

A.1.5 Rerankers

We rerank the top-100 results from the first-stage retrieval fusion using a range of pointwise and listwise reranking models. Pointwise models take a query-document pair as input and output a relevance score. Listwise models take a query and set of documents as input. Some listwise models output one relevance score for each document, whereas others output an ordering of their input documents without producing relevance scores. For models that output an ordering, we use the partial sorting algorithm from RankGPT (Sun et al., 2023) to produce a ranking of all 100 results, where we rerank 20 documents at a time with a sliding window stride of 10 from the bottom of the top 100 candidate documents.

- FIRSTQwen-8B (Chen et al., 2025) is a listwise reranking model built on top of Qwen3 8B (Yang et al., 2025a) and trained with the FIRST objective (Reddy et al., 2024).
- jina-reranker-v3 (Wang et al., 2025) is a reranking model built on top of Qwen3 0.6B that can perform pointwise or listwise reranking. In both configurations, the model outputs one relevance score for each document. We use it in a pointwise configuration.
- mT5 (Bonifacio et al., 2022) is a pointwise reranking model built on top of a multilingual T5 backbone. We use the `mt5-13b-mmarco-100k` checkpoint with 13B parameters.
- Qwen3-Reranker (Zhang et al., 2025) is a family of pointwise reranking models built on top of Qwen3. We use all three model sizes: 0.6B, 4B, and 8B.
- Rank1 (Weller et al., 2025) is a pointwise reranking model fine-tuned to perform reasoning before reranking. Rank1 is built on top of the Qwen2.5 base model (Yang et al., 2024a).
- Rank-K (Yang et al., 2025b) is a listwise reranking model that performs reasoning before reranking using the QwQ 32B reasoning model (Team, 2025).
- RankQwen-32B is a zero-shot listwise reranking system that uses Qwen3-32B[9] without any training to rerank the top-100 documents. This system *does not* use any reasoning.

[9] https://huggingface.co/Qwen/Qwen3-32B

- RankZephyr-7B (Pradeep et al., 2023) is a listwise reranking model built on top of Mistral 7B (Jiang et al., 2023).
- SEARCHER (Agarwal et al., 2025) is a pointwise reranking model built on top of `Mistral-Nemo-Base-2407`.[10] It was trained on Persian, Russian, and Chinese translations of MS MARCO passages (Bajaj et al., 2018), keeping the queries in English. Documents in the training batch were a mixture of all three languages.

[10] https://huggingface.co/mistralai/Mistral-Nemo-Base-2407