# MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages

**Xinyu Zhang**[1*], **Nandan Thakur**[1*], **Odunayo Ogundepo**[1], **Ehsan Kamalloo**[1†], **David Alfonso-Hermelo**[2], **Xiaoguang Li**[3], **Qun Liu**[3], **Mehdi Rezagholizadeh**[2], **Jimmy Lin**[1]

[1]David R. Cheriton School of Computer Science, University of Waterloo, Canada
[2]Huawei Noah's Ark Lab, Canada
[3]Huawei Noah's Ark Lab, China

## Abstract

MIRACL is a multilingual dataset for *ad hoc* retrieval across 18 languages that collectively encompass over three billion native speakers around the world. This resource is designed to support monolingual retrieval tasks, where the queries and the corpora are in the same language. In total, we have gathered over 726k high-quality relevance judgments for 78k queries over Wikipedia in these languages, where all annotations have been performed by native speakers hired by our team. MIRACL covers languages that are both typologically close as well as distant from 10 language families and 13 sub-families, associated with varying amounts of publicly available resources. Extensive automatic heuristic verification and manual assessments were performed during the annotation process to control data quality. In total, MIRACL represents an investment of around five person-years of human annotator effort. Our goal is to spur research on improving retrieval across a continuum of languages, thus enhancing information access capabilities for diverse populations around the world, particularly those that have traditionally been underserved. MIRACL is available at `http://miracl.ai/`.

## 1 Introduction

Information access is a fundamental human right. Specifically, the Universal Declaration of Human Rights by the United Nations articulates that ''everyone has the right to freedom of opinion and expression'', which includes the right ''to seek, receive, and impart information and ideas through any media and regardless of frontiers'' (Article 19). Information access capabilities such as search, question answering, summarization, and recommendation are important technologies for safeguarding these ideals.

With the advent of deep learning in NLP, IR, and beyond, the importance of large datasets as drivers of progress is well understood (Lin et al., 2021b). For retrieval models in English, the MS MARCO datasets (Bajaj et al., 2018; Craswell et al., 2021; Lin et al., 2022) have had a transformative impact in advancing the field. Similarly, for question answering (QA), there exist many resources in English, such as SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), and Natural Questions (Kwiatkowski et al., 2019).

We have recently witnessed many efforts in building resources for non-English languages, for example, CLIRMatrix (Sun and Duh, 2020), XTREME (Hu et al., 2020), MKQA (Longpre et al., 2021), mMARCO (Bonifacio et al., 2021), TyDi QA (Clark et al., 2020), XOR-TyDi (Asai et al., 2021), and Mr. TyDi (Zhang et al., 2021). These initiatives complement multilingual retrieval evaluations from TREC, CLEF, NTCIR, and FIRE that focus on specific language pairs. Nevertheless, there remains a paucity of resources for languages beyond English. Existing datasets are far from sufficient to fully develop information access capabilities for the 7000+ languages spoken on our planet (Joshi et al., 2020). Our goal is to take a small step towards addressing these issues.

To stimulate further advances in multilingual retrieval, we have built the MIRACL dataset on top of Mr. TyDi (Zhang et al., 2021), comprising human-annotated passage-level relevance judgments on Wikipedia for 18 languages, totaling over 726k query–passage pairs for 78k queries. These languages are written using 11 distinct scripts, originate from 10 different language families, and collectively encompass more than three

[*] Equal contribution.
[†] Work done while at Huawei Noah's Ark Lab.



Transactions of the Association for Computational Linguistics, vol. 11, pp. 1114–1131, 2023. https://doi.org/10.1162/tacl_a_00595
Action Editor: Xiaodong He. Submission batch: 4/2023; Revision batch: 4/2023; Published 9/2023.

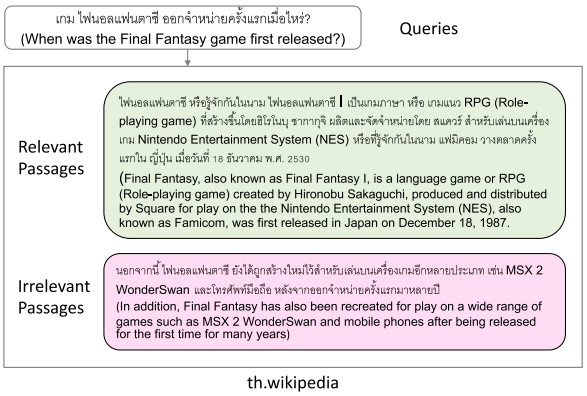


Figure 1: Examples of annotated query–passage pairs in Thai (`th`) from MIRACL.

billion native speakers around the world. They include what would typically be characterized as high-resource languages as well as low-resource languages. Figure 1 shows a sample Thai query with a relevant and a non-relevant passage. In total, the MIRACL dataset represents over 10k hours, or about five person-years, of annotator effort.

Along with the dataset, our broader efforts included organizing a competition at the WSDM 2023 conference that provided a common evaluation methodology, a leaderboard, and a venue for a competition-style event with prizes. To provide starting points that the community can rapidly build on, we also share reproducible baselines in the Pyserini IR toolkit (Lin et al., 2021a).

Compared to existing datasets, MIRACL provides more thorough and robust annotations and broader coverage of the languages, which include both typologically diverse and similar language pairs. We believe that MIRACL can serve as a high-quality training and evaluation dataset for the community, advance retrieval effectiveness in diverse languages, and answer interesting scientific questions about cross-lingual transfer in the multilingual retrieval context.

## 2 Background and Related Work

The focus of this work is the standard *ad hoc* retrieval task in information retrieval, where given a corpus $\mathcal{C}$, the system's task is to return for a given query $q$ an ordered list of top-$k$ passages from $\mathcal{C}$ that maximizes some standard quality metric such as nDCG. A query $q$ is a well-formed natural language question in some language $\mathcal{L}_n$ and the passages draw from the same language $\mathcal{C}_n$. Thus, our focus is *monolingual* retrieval across diverse languages, where the queries and the corpora are in the same language (e.g., Thai queries searching Thai passages), as opposed to *cross-lingual* retrieval, where the queries and the corpora are in *different* languages (e.g., searching a Swahili corpus with Arabic queries).

As mentioned in the Introduction, there has been work over the years on building resources for retrieval in non-English languages. Below, we provide a thorough comparison between MIRACL and these efforts, with an overview in Table 1.

### 2.1 Comparison to Traditional IR Collections

Historically, there have been monolingual retrieval evaluations of search tasks in non-English languages, for example, at TREC, FIRE, CLEF, and NCTIR. These community evaluations typically release test collections built from newswire articles, which typically provide only dozens of queries with modest amounts of relevance judgments, and are insufficient for training neural retrieval models. The above organizations also provide evaluation resources for cross-lingual retrieval, but they cover relatively few language pairs. For example, the recent TREC 2022 NeuCLIR Track (Lawrie et al., 2023) evaluates only three languages (Chinese, Persian, and Russian) in a cross-lingual setting.

### 2.2 Comparison to Multilingual QA Datasets

There are also existing datasets for multilingual QA. For example, XOR-TyDi (Asai et al., 2021) is a cross-lingual QA dataset built on TyDi by annotating answers in English Wikipedia for questions that TyDi considers unanswerable in the original source (non-English) language. This setup, unfortunately, does not allow researchers to examine *monolingual* retrieval in non-English languages.

Another point of comparison is MKQA (Longpre et al., 2021), which comprises 10k question–answer pairs aligned across 26 typologically diverse languages. Questions are paired with exact answers in the different languages, and evaluation is conducted in the open-retrieval setting by matching those answers in retrieved text—thus, MKQA is not a ''true'' retrieval dataset. Furthermore, because the authors translated questions to achieve cross-lingual alignment, the translations may not be ''natural'', as pointed out by Clark et al. (2020).

| Dataset Name | Natural Queries | Natural Passages | Human Labels | # Lang | Avg # Q | Avg # Labels/Q | Total # Labels | Training? |
|---|---|---|---|---|---|---|---|---|
| NeuCLIR (Lawrie et al., 2023) | ✓ | ✓ | ✓ | 3 | 160 | 32.74 | 5.2k | × |
| MKQA (Longpre et al., 2021) | × | ✓ | ✓ | 26 | 10k | 1.35 | 14k | × |
| mMARCO (Bonifacio et al., 2021) | × | × | ✓ | 13 | 808k | 0.66 | 533k | ✓ |
| CLIRMatrix (Sun and Duh, 2020) | × | ✓ | × | 139 | 352k | 693 | 34B | ✓ |
| Mr. TyDi (Zhang et al., 2021) | ✓ | ✓ | ✓ | 11 | 6.3k | 1.02 | 71k | ✓ |
| MIRACL (our work) | ✓ | ✓ | ✓ | 18 | 4.3k | 9.23 | 726k | ✓ |

Table 1: Comparison of select multilingual retrieval datasets. **Natural Queries** and **Natural Passages** indicate whether the queries and passages are ''natural'', i.e., generated by native speakers (vs. human- or machine-translated), and for queries, in natural language (vs. keywords or entities); **Human Labels** indicates human-generated labels (vs. synthetically generated labels); **# Lang** is the number of languages supported; **Avg # Q** is the average number of queries for each language; **Avg # Labels/Q** is the average number of labels provided per query; **Total # Labels** is the total number of human labels (both positive and negative) across all languages (including synthetic labels in CLIRMatrix). **Training?** indicates whether the dataset provides sufficient data for training neural models.

### 2.3 Comparison to Synthetic Datasets

Since collecting human relevance labels is laborious and costly, other studies have adopted workarounds to build multilingual datasets. For example, Bonifacio et al. (2021) automatically translated the MS MARCO dataset (Bajaj et al., 2018) from English into 13 other languages. However, translation is known to cause inadvertent artifacts such as ''translationese'' (Clark et al., 2020; Lembersky et al., 2012; Volansky et al., 2015; Avner et al., 2016; Eetemadi and Toutanova, 2014; Rabinovich and Wintner, 2015) and may lead to training data of questionable value.

Alternatively, Sun and Duh (2020) built synthetic bilingual retrieval datasets in a resource called CLIRMatrix based on the parallel structure of Wikipedia that covers 139 languages. Constructing datasets automatically by exploiting heuristics has the virtue of not requiring expensive human annotations and can be easily scaled up to cover many languages. However, such datasets are inherently limited by the original resource they are built from. For instance, in CLIRMatrix, the queries are the titles of Wikipedia articles, which tend to be short phrases such as named entities. Also, multi-degree judgments in the dataset are directly converted from BM25 scores, which creates an evaluation bias towards lexical approaches.

### 2.4 Comparison to Mr. TyDi

Since MIRACL inherits from Mr. TyDi, it makes sense to discuss important differences between the two: Mr. TyDi (Zhang et al., 2021) is a human-labeled retrieval dataset built atop TyDi QA (Clark et al., 2020), covering 11 typologically diverse languages. While Mr. TyDi enables the training and evaluation of monolingual retrieval, it has three shortcomings we aimed to address in MIRACL.

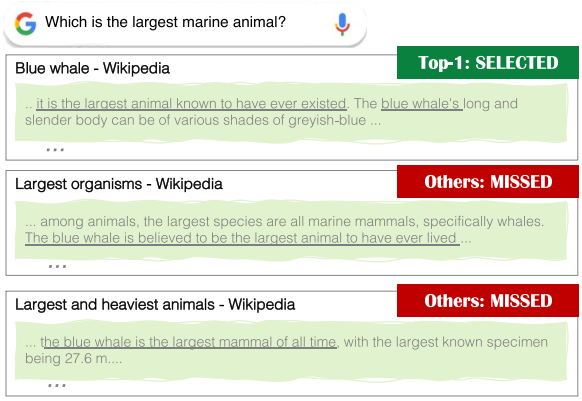


Figure 2: Examples of missing relevant passages in TyDi QA (and thus Mr. TyDi) for the query ''*Which is the largest marine animal?*'' Since only relevant passages in the top Wikipedia article are included, other relevant passages are missed.

**Limited Positive Passages.** In TyDi QA, and thus Mr. TyDi, candidate passages for annotation are selected only from the top-ranked Wikipedia article based on a Google search. Consequently, a considerable number of relevant passages that exist in other Wikipedia articles are ignored; Figure 2 illustrates an instance of this limitation. In contrast,

MIRACL addresses this issue by sourcing candidate passages from all of Wikipedia, ensuring that our relevant passages are diverse.

Moreover, in MIRACL, we went a step further and asked annotators to assess the top-10 candidate passages from an ensemble model per query, resulting in richer annotations compared to those of Mr. TyDi, which were mechanistically generated from TyDi QA as no new annotations were performed. Furthermore, we believe that explicitly judged negative examples are quite valuable, compared to, for example, implicit negatives in MS MARCO sampled from BM25 results, as recent work has demonstrated the importance of so-called ''hard negative'' mining (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021; Santhanam et al., 2021; Formal et al., 2021). As the candidate passages come from diverse models, they are particularly suitable for feeding various contrastive techniques.

**Inconsistent Passages.** Since passage-level relevance annotations in Mr. TyDi were derived from TyDi QA, it retained exactly those same passages. However, as TyDi QA was not originally designed for retrieval, it did not provide consistent passage segmentation for all Wikipedia articles. Thus, the Mr. TyDi corpora comprised a mix of TyDi QA passages and custom segments that were heuristically adjusted to ''cover'' the entire raw Wikipedia dumps. This inconsistent segmentation is a weakness of Mr. TyDi that we rectified in MIRACL by re-segmenting the Wikipedia articles provided by TyDi QA and re-building the relevance mapping for these passages (more details in Section 4.2).

**No Typologically Similar Languages.** The 11 languages included in TyDi QA encompass a broad range of linguistic typologies by design, belonging to 11 different sub-families from 9 different families. However, this means that they are all quite distant from one another. In contrast, new languages added to MIRACL include typologically similar languages. The inclusion of these languages is crucial because their presence can better foster research on cross-lingual comparisons. For example, cross-lingual transfer can be more effectively studied when more similar languages are present. We explore this issue further in Section 5.3.

**Summary.** Overall, not only is MIRACL an order of magnitude larger than Mr. TyDi, as shown in Table 1, but MIRACL goes beyond simply scaling up the dataset to correcting many shortcomings observed in Mr. TyDi. The result is a larger, higher-quality, and richer dataset to support monolingual retrieval in diverse languages.

## 3 Dataset Overview

MIRACL is a multilingual retrieval dataset that spans 18 different languages, focusing on the monolingual retrieval task. In total, we have gathered over 726k manual relevance judgments (i.e., query–passage pairs) for 78k queries across Wikipedia in these languages, where all assessments have been performed by native speakers. There are sufficient examples in MIRACL to train and evaluate neural retrieval models. Detailed statistics for MIRACL are shown in Table 2.

Among the 18 languages in MIRACL, 11 are existing languages that are originally covered by Mr. TyDi, where we take advantage of the Mr. TyDi queries as a starting point. Resources for the other 7 new languages are created from scratch. We generated new queries for all languages, but since we inherited Mr. TyDi queries, fewer new queries were generated for the 11 existing languages as these languages already had sizable training and development sets. For all languages, we provide far richer annotations with respect to Wikipedia passages. That is, compared to Mr. TyDi, where each query has on average only a single positive (relevant) passage (Zhang et al., 2021), MIRACL provides far more positive as well as negative labels for all queries. Here, we provide an overview of the data splits; details of fold construction will be introduced in Section 4.2.

- **Training and development sets:** For the 11 existing Mr. TyDi languages, the training (development) sets comprise subsets of the queries from the Mr. TyDi training (development) sets. The main difference is that MIRACL provides richer annotations for more passages. For the new languages, the training (development) data consist of entirely of newly generated queries.

- **Test-A sets:** Similar to the training and development sets, the test-A sets align with the test sets in Mr. TyDi. However, the test-A sets exist *only* for the existing languages.

| Lang | ISO | Train | | Dev | | Test-A | | Test-B | | # Passages | # Articles | Avg. Q Len | Avg. P Len |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | # Q | # J | # Q | # J | # Q | # J | # Q | # J | | | | |
| Arabic | ar | 3,495 | 25,382 | 2,896 | 29,197 | 936 | 9,325 | 1,405 | 14,036 | 2,061,414 | 656,982 | 6 | 54 |
| Bengali | bn | 1,631 | 16,754 | 411 | 4,206 | 102 | 1,037 | 1,130 | 11,286 | 297,265 | 63,762 | 7 | 56 |
| English | en | 2,863 | 29,416 | 799 | 8,350 | 734 | 5,617 | 1,790 | 18,241 | 32,893,221 | 5,758,285 | 7 | 65 |
| Finnish | fi | 2,897 | 20,350 | 1,271 | 12,008 | 1,060 | 10,586 | 711 | 7,100 | 1,883,509 | 447,815 | 5 | 41 |
| Indonesian | id | 4,071 | 41,358 | 960 | 9,668 | 731 | 7,430 | 611 | 6,098 | 1,446,315 | 446,330 | 5 | 49 |
| Japanese | ja | 3,477 | 34,387 | 860 | 8,354 | 650 | 6,922 | 1,141 | 11,410 | 6,953,614 | 1,133,444 | 17 | 147 |
| Korean | ko | 868 | 12,767 | 213 | 3,057 | 263 | 3,855 | 1,417 | 14,161 | 1,486,752 | 437,373 | 4 | 38 |
| Russian | ru | 4,683 | 33,921 | 1,252 | 13,100 | 911 | 8,777 | 718 | 7,174 | 9,543,918 | 1,476,045 | 6 | 46 |
| Swahili | sw | 1,901 | 9,359 | 482 | 5,092 | 638 | 6,615 | 465 | 4,620 | 131,924 | 47,793 | 7 | 36 |
| Telugu | te | 3,452 | 18,608 | 828 | 1,606 | 594 | 5,948 | 793 | 7,920 | 518,079 | 66,353 | 5 | 51 |
| Thai | th | 2,972 | 21,293 | 733 | 7,573 | 992 | 10,432 | 650 | 6,493 | 542,166 | 128,179 | 42 | 358 |
| Spanish | es | 2,162 | 21,531 | 648 | 6,443 | – | – | 1,515 | 15,074 | 10,373,953 | 1,669,181 | 8 | 66 |
| Persian | fa | 2,107 | 21,844 | 632 | 6,571 | – | – | 1,476 | 15,313 | 2,207,172 | 857,827 | 8 | 49 |
| French | fr | 1,143 | 11,426 | 343 | 3,429 | – | – | 801 | 8,008 | 14,636,953 | 2,325,608 | 7 | 55 |
| Hindi | hi | 1,169 | 11,668 | 350 | 3,494 | – | – | 819 | 8,169 | 506,264 | 148,107 | 10 | 69 |
| Chinese | zh | 1,312 | 13,113 | 393 | 3,928 | – | – | 920 | 9,196 | 4,934,368 | 1,246,389 | 11 | 121 |
| German | de | – | – | 305 | 3,144 | – | – | 712 | 7,317 | 15,866,222 | 2,651,352 | 7 | 58 |
| Yoruba | yo | – | – | 119 | 1,188 | – | – | 288 | 2,880 | 49,043 | 33,094 | 8 | 28 |
| Total | | 40,203 | 343,177 | 13,495 | 130,408 | 7,611 | 76,544 | 17,362 | 174,496 | 106,332,152 | 19,593,919 | 10 | 77 |

Table 2: Descriptive statistics for all languages in MIRACL, organized by split. **# Q**: number of queries; **# J**: number of labels (relevant and non-relevant); **# Passages**: number of passages; **# Articles**: number of Wikipedia articles from which the passages were drawn; **# Avg. Q Len**: average number of tokens per query; **# Avg. P Len**: average number of tokens per passage. Tokens are based on characters for th, ja, and zh, otherwise delimited by whitespace. Underlined are the new languages in MIRACL.

- **Test-B sets:** For all languages, the test-B sets are composed entirely of new queries that have never been released before (compared to test-A sets, whose queries ultimately draw from TyDi QA and thus have been publicly available for quite some time now). These queries can be viewed as a *true* held-out test set.

Although the training, development, and test-A sets of MIRACL languages that overlap with Mr. TyDi align with Mr. TyDi, in some cases there are fewer queries in MIRACL than in their corresponding Mr. TyDi splits. This is because our annotators were unable to find relevant passages for some queries from Mr. TyDi; we call these ''invalid'' queries and removed them from the corresponding MIRACL splits (details in Section 5.2).

To evaluate the quality of system outputs, we use standard information retrieval metrics, nDCG@10 and Recall@100, which are defined as follows:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2(i+1)} \tag{1}$$

$$\text{nDCG@}k = \frac{\text{DCG@}k}{\text{iDCG@}k} \tag{2}$$

$$\text{Recall@}k = \frac{\sum_{i=1}^{k} \text{rel}_\text{i}}{k} \tag{3}$$

where $\text{rel}_i = 1$ if $d_i$ is relevant to the query and 0 otherwise, and iDCG@$k$ is the DCG@$k$ of the ideal ranked list (i.e., with binary judgments, all relevant documents appear before non-relevant documents). We set $k$ to 10 and 100 for the two metrics, respectively, to arrive at nDCG@10 and Recall@100. These serve as the official metrics of MIRACL.

In addition to releasing the evaluation resources, we organized a WSDM Cup challenge at the WSDM 2023 conference to encourage participation from the community, where the test-B sets were used for final evaluation. Among the 18 languages, German (de) and Yoruba (yo), the bottom block in Table 2, were selected as *surprise languages*, while the rest are *known languages*. This distinction was created for the official WSDM competition. Whereas data from the known languages were released in October 2022, the identity of the surprise languages was concealed until two weeks before the competition deadline in

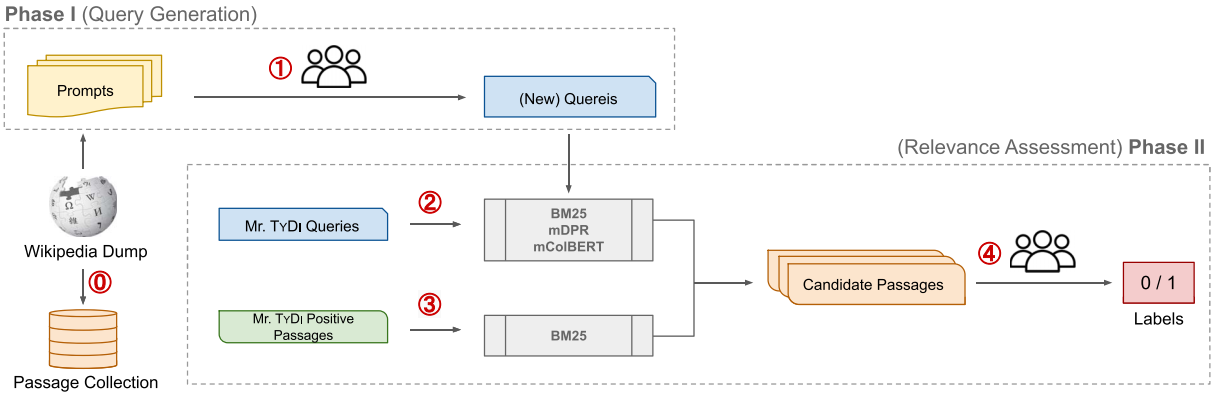


Figure 3: Diagram illustrating the MIRACL annotation workflow.

January 2023. For the known languages, participants were given ample data and time to train language-specific models. On the other hand, for the surprise languages, no training splits were provided to specifically evaluate retrieval under a limited data and time condition. However, since the WSDM Cup challenge has concluded, the distinction between surprise and known languages is no longer relevant.

## 4 Dataset Construction

To build MIRACL, we hired native speakers as annotators to provide high-quality queries and relevance judgments. At a high level, our workflow comprised two phases: First, the annotators were asked to generate well-formed queries based on ‘‘prompts’’ (Clark et al., 2020) (details in Section 4.2). Then, they were asked to assess the relevance of the top-$k$ query–passage pairs produced by an ensemble baseline retrieval system.

An important feature of MIRACL is that our dataset was *not* constructed via crowd-sourced workers, unlike other previous efforts such as SQuAD (Rajpurkar et al., 2016). Instead, we hired 31 annotators (both part-time and full-time) across all languages. Each annotator was interviewed prior to being hired and was verified to be a native speaker of the language they were working in. Our team created a consistent onboarding process that included carefully crafted training sessions. We began interviewing annotators in mid-April 2022; dataset construction began in late April 2022 and continued until the end of September 2022.

Throughout the annotation process, we constantly checked randomly sampled data to monitor annotation quality (see Section 4.3). Whenever issues were detected, we promptly communicated with the annotators to clarify the problems so that they were resolved expeditiously. Constant interactions with the annotators helped minimize errors and misunderstandings. We believe that this design yielded higher quality annotations than what could have been obtained by crowd-sourcing.

In total, MIRACL represents over 10k hours of assessor effort, or around five person-years. We offered annotators the hourly rate of $18.50 per hour (converted into USD). For reference, the local minimum wage is $11.50 USD/hr.

### 4.1 Corpora Preparation

For each MIRACL language, we prepared a pre-segmented passage corpus from a raw Wikipedia dump. For the existing languages in Mr. TyDi, we used exactly the same raw Wikipedia dump as Mr. TyDi and TyDi QA from early 2019. For the new languages, we used releases from March 2022. We parsed the Wikipedia articles using WikiExtractor[1] and segmented them into passages based on natural discourse units using two consecutive newlines in the wiki markup as the delimiter. Query–passage pairs formed the basic annotation unit (Step ⓪ in Figure 3).

### 4.2 Annotation Workflow

MIRACL was created using a two-phase annotation process adapted from TyDi QA, which was in turn built on best practices derived from Clark et al. (2020). The two phases are *query generation* and *relevance assessment*.

**Query Generation.** In Phase I, annotators were shown ‘‘prompts’’, comprising the first 100 words

[1] https://github.com/attardi/wikiextractor.

of randomly selected Wikipedia articles that provide contexts to elicit queries. Following Clark et al. (2020), the prompts are designed to help annotators write queries for which they ''seek an answer.''

To generate high-quality queries, we asked annotators to avoid generating queries that are directly answerable by the prompts themselves. They were asked to generate well-formed natural language queries likely (in their opinion) to have precise, unambiguous answers. Using prompts also alleviates the issue where the queries may be overly biased towards their personal experiences. We also gave the annotators the option of skipping any prompt that did not ''inspire'' any queries. This process corresponds to Step ① in Figure 3.

Note that in this phase, annotators were asked to generate queries *in batch* based on the prompts, and did not proceed to the next phase before finishing tasks in this phase. Thus, the annotators had not yet examined any retrieval results (i.e., Wikipedia passages) at this point. Moreover, we did not suggest to annotators during the entire process that the queries should be related to Wikipedia in order to prevent the annotators from writing oversimplified or consciously biased queries. Therefore, it could be the case that their queries cannot be readily answered by the information contained in the corpora, which we discuss in Section 5.2.

**Relevance Assessment.** In the second phase, for each query from the previous phase, we asked the annotators to judge the binary relevance of the top-$k$ candidate passages ($k = 10$) from an ensemble retrieval system that combines three separate models, which corresponds to Step ② in Figure 3:

- **BM25** (Robertson and Zaragoza, 2009), a traditional retrieval algorithm based on lexical matching, which has been shown to be a robust baseline when evaluated zero-shot across domains and languages (Thakur et al., 2021; Zhang et al., 2021). We used the implementation in Anserini (Yang et al., 2018), which is based on the open-source Lucene search library, with default parameters and the corresponding language-specific analyzer in Lucene if it exists. If not, we simply used the white space tokenizer.
- **mDPR** (Karpukhin et al., 2020; Zhang et al., 2021), a single-vector dense retrieval method that has proven to be effective for many retrieval tasks. This model was trained using the Tevatron toolkit (Gao et al., 2023) starting from an mBERT checkpoint and then fine-tuned using the training set of MS MARCO Passage. Retrieval was performed in a zero-shot manner.
- **mColBERT** (Khattab and Zaharia, 2020; Bonifacio et al., 2021), a multi-vector dense retrieval model that has been shown to be effective in both in-domain and out-of-domain contexts (Thakur et al., 2021; Santhanam et al., 2021). This model was trained using the authors' official repository.[2] Same as mDPR, the model was initialized from mBERT and fine-tuned on MS MARCO Passage; retrieval was also performed zero shot.

For each query, we retrieved the top 1000 passages using each model, then performed ensemble fusion by first normalizing all retrieval scores to the range $[0, 1]$ and then averaging the scores. A final ranked list was then generated from these new scores. Based on initial experiments, we found that annotating the top 10 passages per query yielded a good balance in terms of obtaining diverse passages and utilizing annotator effort.

As mentioned in Section 2.4, we rectified the inconsistent passage segmentation in Mr. TyDi by re-segmenting the Wikipedia articles of TyDi QA. The downside, unfortunately, is that annotated passages from Mr. TyDi may no longer exist in the MIRACL corpora. Thus, to take advantage of existing annotations, for queries from the 11 existing languages in Mr. TyDi, we augmented the set of passages to be assessed with ''projected'' relevant passages. This allowed us to take advantage of existing annotations transparently in our workflow. To accomplish this, we used relevant passages from Mr. TyDi as queries to search the corresponding MIRACL corpus using BM25. As the passages in both corpora differ in terms of segmentation but not content, the top retrieved passages are likely to have substantial overlap with the annotated passage in Mr. TyDi (and hence are also likely to be relevant). This is shown as Step ③ in Figure 3.

[2] https://github.com/stanford-futuredata/ColBERT#colbertv1.

We used a simple heuristic to determine how many of these results to re-assess. If the score of the top retrieved passage from MIRACL is 50% higher than the score of the passage ranked second, we have reasonable confidence (based on initial explorations) that the top passage is a good match for the original relevant passage. In this case, we only add the top passage to the set of candidates that the assessor considers. Otherwise, we add the top 5 passages.

Once the candidate passages are prepared, the annotators are asked to provide a binary label for each query–passage pair (1 = ''relevant'' and 0 = ''not relevant''). This is shown in Step ④ in Figure 3. Note that the ''projected'' passages prepared from Step ③ are not identified to the annotators. In all cases, they simply receive a set of passages to label per query, without any explicit knowledge of where the passages came from.

In our design, each query–passage pair was only labeled by one annotator, which we believe is a better use of limited resources, compared to the alternative of multiple judgments over fewer queries. The manual assessment proceeded in batches of approximately 1000 query–passage pairs. Our annotators differed greatly in speed, but averaged over all languages and individuals, each batch took roughly 4.25 hours to complete.

**Fold Creation and Data Release.** At a high-level, MIRACL contains two classes of queries: those inherited from Mr. TyDi and those that were created from scratch, from which four splits were prepared. During the annotation process, all queries followed the same workflow. After the annotation process concluded, we divided MIRACL into training sets, development sets, test-A sets, and test-B sets, as described in Section 3. For existing languages, the training, development, and test-A sets align with the training, development, and test sets in Mr. TyDi, and the test-B sets are formed by the newly generated queries. For the new (known) languages, all generated queries were split into training, development, and test-B sets with a split ratio of 50%, 15%, and 35%. Note that there are no test-A sets for these languages.

### 4.3 Quality Control

To ensure data quality, we implemented two processes: automatic heuristic verification that includes simple daily checks and manual periodic assessment executed by human reviewers on a random subset of annotations.

#### 4.3.1 Automatic Verification

Automatic heuristic verification was applied to both phases daily, using language-specific heuristics to flag ''obviously bad'' annotations. In Phase I, we flagged the generated query if it was: (1) left empty; (2) similar to previous queries in the same language;[3] (3) missing interrogative indicators;[4] or (4) overly short or long.[5] In Phase II, we flagged the label if it was left empty or an invalid value (i.e., values that are not 0 or 1).

The most extreme cases of bad annotations (e.g., empty or duplicate queries) were removed from the dataset, while minor issues were flagged but the data were retained. For both cases, whenever the metrics dropped below a reasonable threshold,[6] we scheduled a meeting with the annotator in question to discuss. This allowed us to minimize the number of obvious errors while giving and receiving constructive feedback.

#### 4.3.2 Manual Assessment

Manual assessment was also applied to both phases. To accomplish this, we hired another group of native speakers of each language as reviewers (with minor overlap). Similarly, we provided consistent onboarding training to each reviewer.

**Phase I.** In this phase, reviewers were given both the prompts and the generated queries, and asked to apply a checklist to determine whether the queries met our requirements. Criteria include the examination of the query itself (spelling or syntax errors, fluency, etc.) and whether the query could be answered directly by the prompt, which we wished to avoid (see Section 4.2).

*How fluent are the queries?* Review results showed that approximately 12% of the questions

[3]We measured similarity using Levenshtein distance. A query was flagged if its similarity score to any previous query was greater than 75%.

[4]For example, writing system- and language-specific interrogative punctuation and particles.

[5]For `zh`, `ja`, and `th`, length is measured in characters, where the expected range is $[6, 40]$ for `zh` and `ja`, and $[6, 120]$ for `th`. Otherwise, length is measured in tokens delimited by white space, with expected range $[3, 15]$.

[6]In practice, we set this threshold to 0.8 in terms of the percentage of questionable annotations in each submission.

had spelling or syntax errors, or ‘‘sound artificial’’. However, almost all these flaws (over 99%) did not affect the understanding of the question itself. We thus retained these ‘‘flawed’’ queries to reflect real-world distributions.

*How related are the queries to the prompts?* We measured the lexical overlap between the queries and their corresponding prompts to understand their connections. Our analysis shows approximately two words of overlap on average, which is consistent with statistics reported in Clark et al. (2020). Overlaps primarily occur in entities or stopwords. We thus conclude that the generated queries are reasonably different from the given prompts.

**Phase II.** In this phase, reviewers were provided the same guidance as annotators performing the relevance assessment. They were asked to (independently) label a randomly sampled subset of the query–passage pairs. The degree of agreement on the overlapping pairs is used to quantify the quality of the relevance labels. As a high-level summary, we observe average agreements of over 80% on query–passage relevance.

*Why are there disagreements?* As is well-known from the IR literature dating back many decades (Voorhees, 1998), real-world annotators disagree due to a number of ‘‘natural’’ reasons. To further understand why, we randomly sampled five query–passage pairs per language for closer examination. After some iterative sense-making, we arrived at the conclusion that disagreements come from both mislabeling and partially relevant pairs. We show examples of the two cases in Figure 4a and Figure 4b.

Figure 4a depicts a common mislabeling case, where a passage starting with misleading information is labeled as non-relevant. We observe more mislabeling errors from the reviewers compared to the annotators, which may be attributed to the annotators’ greater commitment to the task and more experience in performing it. That is, in these cases, the reviewers are more often ‘‘wrong’’ than the original assessors.

Partially relevant annotations can take on various forms. The most frequent case is exemplified by Figure 4b, where a passage is related to the query but does not provide a precise answer. Differences in ‘‘thresholds’’ for relevance are inevitable under a binary relevance system. Furthermore, disagreements often arise when an annotator labels a passage as non-relevant, but a reviewer labels it as relevant. These cases suggest that annotators exercise greater caution and are more ‘‘strict’’ when they are uncertain. Overall, these analyses convinced us that the generated queries and annotations are of high quality.

**Query:**
ইন্দোনেশিয়ার রাজধানী কোথায় ?

Where is the capital of Indonesia?

**Candidate Passage:**
সুরিনামের রাজধানী পারামারিবোতে ইন্দোg gনেশিয়ার একটি দূতাবাস রয়েছে। এই দূতাবাসের মাধ্যমে, ইন্দোনেশিয়ার সাথে গায়ানার বিভিন্ন বিষয়াবলী এবং কার্যক্রম সম্পন্ন করা হয়। অপরদিকে, ইন্দোনেশিয়ার রাজধানী জাকার্তায় সুরিনামের একটি দূতাবাস রয়েছে। ইন্দোনেশিয়া এবং সুরিনাম উভয় দেশই বিশ্ব বাণিজ্য সংস্থা (ডব্লিউটিও) এবং ফোরাম অব ইস্ট এশিয়া-লাতিন আমেরিকা কো-অপারেশন এর সদস্য।.

Indonesia has an embassy in Paramaribo, the capital of Suriname. Through this embassy, Guyana's various affairs and activities with Indonesia are carried out. On the other hand, Suriname has an embassy in **Jakarta, the capital of Indonesia**. Both Indonesia and Suriname are members of the World Trade Organization (WTO) and the Forum of East Asia-Latin America Cooperation.

(a) Example of an annotation mistake: the relevant information (highlighted in red) appears in the middle of a passage containing mostly non-relevant material.

**Query:**
هل إسرائيل هي دولة دينية ؟
Is Israel a religious state?

**Candidate Passage:**
أصدرت إسرائيل منذ تأسيسها العديد من القوانين التي تعكس الهوية اليهودية وقيم الأغلبية (حوالي 75٪ في عام 2016 (من مواطنيها .ومع ذلك فإن النقاش الدائر في إسرائيل إزاء كون الدولة علمانية أم دينية جعل النقاش يتركز حول طبيعة الدولة اليهودية على وجه التحديد . كما ركز جانب آخر من النقاش على وضع الأقليات في إسرائيل وعلى وجه الخصوص السكان العرب الإسرائيليين.

Since its founding, Israel has passed many laws that reflect the Jewish identity and the values of the majority (about 75% in 2016) of its citizens. However, the debate in Israel about whether the state is secular or religious has made the debate focus on the nature of the Jewish state specifically. Another aspect of the discussion focused on the situation of minorities in Israel and in particular the Israeli Arab population.

(b) Example of a partially relevant query–passage pair.

Figure 4: Examples of disagreements between annotators and reviewers.

## 5 MIRACL Analysis

To better characterize MIRACL, we present three separate analyses of the dataset we have built.

### 5.1 Question Words

We first analyze question types, following Clark et al. (2020). Table 3 shows the distribution of the question word in the English dataset of MIRACL, along with SQuAD (Rajpurkar et al., 2016) as a point of reference. The MIRACL statistics are grouped by splits in Table 3. The test-B column shows the distribution of the new queries generated with the workflow described in Section 4.2, whereas the other columns correspond to queries derived from Mr. TyDi.

| | MIRACL | | | | SQuAD |
|---|---|---|---|---|---|
| | Train | Dev | Test-A | Test-B | |
| HOW | 15% | 17% | 19% | 6% | 12% |
| WHAT | 26% | 27% | 31% | 34% | 51% |
| WHEN | 26% | 25% | 20% | 6% | 8% |
| WHERE | 6% | 7% | 4% | 6% | 5% |
| WHICH | 1% | 1% | 2% | 25% | 5% |
| WHO | 13% | 11% | 9% | 19% | 11% |
| WHY | 1% | 1% | 1% | 1% | 2% |
| YES/NO | 6% | 7% | 4% | 6% | <1% |

Table 3: Query distribution of each split in English MIRACL, compared to SQuAD.

We observe a more balanced distribution of question words in MIRACL compared to SQuAD. In addition, the question words highlight a distribution shift between Mr. TyDi and the new queries. More specifically, the test-B split contains more WHICH and WHO queries, while containing fewer HOW and WHEN queries, compared to the other splits inherited from Mr. TyDi. This also shows that exactly replicating a query generation process is challenging, even when closely following steps outlined in prior work.

### 5.2 Queries with No Relevant Passages

Recall that since the query generation and relevance assessment phases were decoupled, for some fraction of queries in each language, annotators were not able to identify any relevant passage in the pairs provided to them. Because we aimed to have at least one relevant passage per query, these queries—referred to as ''invalid'' for convenience—were discarded from the final dataset.

For each language, we randomly sampled five invalid queries and spot-checked them using a variety of tools, including interactive searching on the web. From this, we identified a few common reasons for the lack of relevant passages in Wikipedia: (1) query not asking for factual information, for example, ''*Kuinka nuoret voivat osallistua suomalaisessa yhteiskunnassa?*'' (Finnish, ''*How can young people participate in Finnish society?*''); (2) query being too specific, for example, ''华盛顿 2021年新增人口多少?'' (Chinese, ''*What is the new population of Washington in 2021?*''); (3) inadequate Wikipedia content, where the query seems reasonable, but the relevant information could not be found in Wikipedia due to the low-resource nature of the language (even after interactive search with Google), for example, ''*Bawo ni aaye Tennis se tobi to?*'' (Yoruba, ''*How big is the tennis court?*'').

We note that just because no relevant passage was found in the candidates presented to the annotators, it is not necessarily the case that no relevant passage exists in the corpus. However, short of exhaustively assessing the corpus (obviously impractical), we cannot conclude this with certainty. Nevertheless, for dataset consistency we decided to discard ''invalid'' queries from MIRACL.

### 5.3 Discussion of MIRACL Languages

We provide a discussion of various design choices made in the selection of languages included in MIRACL, divided into considerations of linguistic diversity and demographic diversity.

**Linguistic Diversity.** One important consideration in the selection of languages is diversity from the perspective of linguistic characteristics. This was a motivating design principle in TyDi QA, which we built on in MIRACL. Our additional languages introduced *similar* pairs as well, opening up new research opportunities. A summary can be found in Table 4.

Families and sub-families provide a natural approach to group languages based on historical origin. Lineage in this respect explains a large amount of resemblance between languages, for example, *house* in English and *haus* in German (Greenberg, 1960). The 18 languages in MIRACL are from 10 language families and 13 sub-families, shown in the first two columns of Table 4.

The notion of *synthesis* represents an important morphological typology, which categorizes languages based on how words are formed from *morphemes*, the smallest unit of meaning. The languages in MIRACL are balanced across the three synthesis types (*analytic*, *agglutinative*, and *fusional*), which form a spectrum with increasing complexity in word formation when moving from analytic to fusional languages. See more discussion in Plank (1999), Dawson et al. (2016), and Nübling (2020). Additional features of the languages are summarized in Table 4, including the written scripts, word order, use of white space for delimiting tokens, and grammatical gender.

| Language Familiy | Language Sub-Family | Language | Script | Synthesis | Word Order | White Space | Gender | # Speakers | Wikipedia Size |
|---|---|---|---|---|---|---|---|---|---|
| Dravidian | South-Central | Telugu | Telugu | agglutinative | SOV | ✓ | ✓ | 96M | 66,353 |
| Koreanic | – | Korean | Hangul | agglutinative | SOV | ✓ | × | 80M | 437,373 |
| Japonic | – | Japanese | Japanese | agglutinative | SOV | × | × | 128M | 1,133,444 |
| Austronesian | Malayo-Polynesian | Indonesian | Latin | agglutinative | SVO | ✓ | × | 300M | 446,330 |
| Uralic | Finno-Ugric | Finnish | Latin | agglutinative | SVO | ✓ | × | 6M | 447,815 |
| Niger–Congo | Atlantic-Congo | Swahili | Latin | agglutinative | SVO | ✓ | × | 83M | 47,793 |
| | | Yoruba | Latin | analytic | SVO | ✓ | × | 52M | 33,094 |
| Sino-Tibetan | Sinitic | Chinese | Chinese | analytic | SVO | × | ✓ | 1350M | 1,246,389 |
| Kra–Dai | Tai | Thai | Thai | analytic | SVO | × | ✓ | 72M | 128,179 |
| Indo-European | Germanic | English | Latin | analytic | SVO | ✓ | ✓ | 1130M | 5,758,285 |
| | | German | Latin | fusional | SVO, SOV | ✓ | ✓ | 178M | 2,651,352 |
| | Italic | French | Latin | fusional | SVO | ✓ | ✓ | 398M | 2,325,608 |
| | | Spanish | Latin | fusional | SVO | ✓ | ✓ | 592M | 1,669,181 |
| | Balto-Slavic | Russian | Cyrillic | fusional | SVO | ✓ | ✓ | 260M | 1,476,045 |
| | Indo-Iranian | Bengali | Bengali | fusional | SOV | ✓ | ✓ | 337M | 63,762 |
| | | Hindi | Devanagari | fusional | SOV | ✓ | ✓ | 592M | 148,107 |
| | | Persian | Arabic | agglutinative | SOV | ✓ | × | 110M | 857,827 |
| Afro-Asiatic | Semitic | Arabic | Arabic | fusional | VSO | ✓ | ✓ | 630M | 656,982 |

Table 4: Characteristics of languages in MIRACL, including language (sub-)family, script, linguistic typologies (synthesis, word order, and gender), number of speakers (L1 and L2), and number of Wikipedia articles. Under **White Space**, ✓ indicates the language uses white space as the token delimiter; under **Gender**, ✓ indicates that gender is evident in the language. Under **# Speakers** and **Wikipedia Size**, cells are highlighted column-wise with the color gradient ranging from **green** (high values) to **red** (low values), where the **Wikipedia Size** column is identical to the **# Articles** column in Table 2. Underlined are the new languages not included in Mr. TyDi.

In our context, linguistic diversity is critical to answering important questions about the transfer capabilities of multilingual language models (Pires et al., 2019), including whether certain typological characteristics are intrinsically challenging for neural language models (Gerz et al., 2018) and how to incorporate typologies to improve the effectiveness of NLP tasks (Ponti et al., 2019; Jones et al., 2021). While researchers have examined these research questions, most studies have not been in the context of retrieval specifically.

In this respect, MIRACL can help advance the state of the art. Consider language family, for example: Our resource can be used to compare the transfer capacity between languages at different ''distances'' in terms of language kinship. While Mr. TyDi provides some opportunities to explore this question, the additional languages in MIRACL enrich the types of studies that are possible. For example, we might consider both contrastive pairs (i.e., those that are very typologically different) as well as similar pairs (i.e., those that are closer) in the context of multilingual language models.

Similar questions abound for synthesis characteristics, written script, word order, etc. For example, we have three exemplars in the Indo-Iranian sub-family (Bengali, Hindi, and Persian): Despite lineal similarities, these languages use different scripts. How does ''cross-script'' relevance transfer work in multilingual language models? We begin to explore some of these questions in Section 6.2.

**Demographic Diversity.** The other important consideration in our choice of languages is the demographic distribution of language speakers. As noted in the Introduction, information access is a fundamental human right that ought to extend to *every* inhabitant of our planet, regardless of the languages they speak.

We attempt to quantify this objective in the final two columns of Table 4, which presents statistics of language speakers and Wikipedia articles. We count both L1 and L2 speakers.[7] Both columns are highlighted column-wise based on the value, from **green** (high values) to **red** (low values). We can use the size of Wikipedia in that language as a proxy for the amount of language resources that are available. We see that in many

[7]L1 are the native speakers; L2 includes other speakers who learned the language later in life.

| ISO | nDCG@10 | | | | | | Recall@100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BM25 | mDPR | Hyb. | mCol. | mCon. | in-L. | BM25 | mDPR | Hyb. | mCol. | mCon. | in-L. |
| ar | 0.481 | 0.499 | 0.673 | 0.571 | 0.525 | 0.649 | 0.889 | 0.841 | 0.941 | 0.908 | 0.925 | 0.904 |
| bn | 0.508 | 0.443 | 0.654 | 0.546 | 0.501 | 0.593 | 0.909 | 0.819 | 0.932 | 0.913 | 0.921 | 0.917 |
| en | 0.351 | 0.394 | 0.549 | 0.388 | 0.364 | 0.413 | 0.819 | 0.768 | 0.882 | 0.801 | 0.797 | 0.751 |
| fi | 0.551 | 0.472 | 0.672 | 0.465 | 0.602 | 0.649 | 0.891 | 0.788 | 0.895 | 0.832 | 0.953 | 0.907 |
| id | 0.449 | 0.272 | 0.443 | 0.298 | 0.392 | 0.414 | 0.904 | 0.573 | 0.768 | 0.669 | 0.802 | 0.823 |
| ja | 0.369 | 0.439 | 0.576 | 0.496 | 0.424 | 0.570 | 0.805 | 0.825 | 0.904 | 0.895 | 0.878 | 0.880 |
| ko | 0.419 | 0.419 | 0.609 | 0.487 | 0.483 | 0.472 | 0.783 | 0.737 | 0.900 | 0.722 | 0.875 | 0.807 |
| ru | 0.334 | 0.407 | 0.532 | 0.477 | 0.391 | 0.521 | 0.661 | 0.797 | 0.874 | 0.866 | 0.850 | 0.850 |
| sw | 0.383 | 0.299 | 0.446 | 0.358 | 0.560 | 0.644 | 0.701 | 0.616 | 0.725 | 0.692 | 0.911 | 0.909 |
| te | 0.494 | 0.356 | 0.602 | 0.462 | 0.528 | 0.781 | 0.831 | 0.762 | 0.857 | 0.830 | 0.961 | 0.957 |
| th | 0.484 | 0.358 | 0.599 | 0.481 | 0.517 | 0.628 | 0.887 | 0.678 | 0.823 | 0.845 | 0.936 | 0.902 |
| es | 0.319 | 0.478 | 0.641 | 0.426 | 0.418 | 0.409 | 0.702 | 0.864 | 0.948 | 0.842 | 0.841 | 0.783 |
| fa | 0.333 | 0.480 | 0.594 | 0.460 | 0.215 | 0.469 | 0.731 | 0.898 | 0.937 | 0.910 | 0.654 | 0.821 |
| fr | 0.183 | 0.435 | 0.523 | 0.267 | 0.314 | 0.376 | 0.653 | 0.915 | 0.965 | 0.730 | 0.824 | 0.823 |
| hi | 0.458 | 0.383 | 0.616 | 0.470 | 0.286 | 0.458 | 0.868 | 0.776 | 0.912 | 0.884 | 0.646 | 0.777 |
| zh | 0.180 | 0.512 | 0.526 | 0.398 | 0.410 | 0.515 | 0.560 | 0.944 | 0.959 | 0.908 | 0.903 | 0.883 |
| de | 0.226 | 0.490 | 0.565 | 0.334 | 0.408 | – | 0.572 | 0.898 | 0.898 | 0.803 | 0.841 | – |
| yo | 0.406 | 0.396 | 0.374 | 0.561 | 0.415 | – | 0.733 | 0.715 | 0.715 | 0.917 | 0.770 | – |
| K. Avg | 0.394 | 0.415 | 0.578 | 0.441 | 0.433 | 0.535 | 0.787 | 0.788 | 0.889 | 0.828 | 0.855 | 0.856 |
| S. Avg | 0.316 | 0.443 | 0.470 | 0.448 | 0.412 | – | 0.653 | 0.807 | 0.807 | 0.860 | 0.806 | – |
| Avg | 0.385 | 0.418 | 0.566 | 0.441 | 0.431 | – | 0.772 | 0.790 | 0.880 | 0.832 | 0.849 | – |

Table 5: Baseline results on the MIRACL dev set, where ''K. Avg.'' and ''S. Avg'' indicate the average scores over the known (ar–zh) and surprise languages (de–yo). **Hyb.**: Hybrid results of BM25 and mDPR; **mCol.**: mColBERT; **mCon.**: mContriever; **in-L**: in-language fine-tuned mDPR.

cases, there are languages with many speakers, but are poor in resources. Particularly noteworthy examples include Telugu, Indonesian, Swahili, Yoruba, Thai, Bengali, and Hindi. We hope that the inclusion of these languages will catalyze interest in multilingual retrieval and in turn benefit large populations for whom the languages have been historically overlooked by the mainstream IR research community.

## 6 Experiments Results

### 6.1 Baselines

As neural retrieval models have gained in sophistication in recent years, the ''software stack'' for end-to-end systems has grown more complex. This has increased the barrier to entry for ''newcomers'' who wish to start working on multilingual retrieval. We believe that the growth of the diversity of languages introduced in MIRACL should be accompanied by an increase in the diversity of participants.

To that end, we make available in the popular Pyserini IR toolkit (Lin et al., 2021a) several baselines to serve as foundations that others can build on. Baseline scores for these retrieval models are shown in Table 5 in terms of the two official retrieval metrics of MIRACL. The baselines include the three methods used in the ensemble system introduced in Section 4.2 (BM25, mDPR, mColBERT) plus the following approaches:

- **Hybrid** combines the scores of BM25 and mDPR results. For each query–passage pair, the hybrid score is computed as $s_{\text{Hybrid}} = \alpha \cdot s_{\text{BM25}} + (1 - \alpha) \cdot s_{\text{mDPR}}$, where we set $\alpha = 0.5$ without tuning. Scores of BM25 and mDPR ($s_{\text{BM25}}$ and $s_{\text{mDPR}}$) are first normalized to $[0, 1]$.

- **mContriever** (Izacard et al., 2022) adopts additional pretraining with contrastive loss based on unsupervised data prepared from CCNet (Wenzek et al., 2020), which demonstrates improved effectiveness in downstream IR tasks. We used the authors' released multilingual checkpoint, where the model was further fine-tuned on English MS MARCO after additional pretraining.[8]

[8] https://huggingface.co/facebook/mcontriever-msmarco.

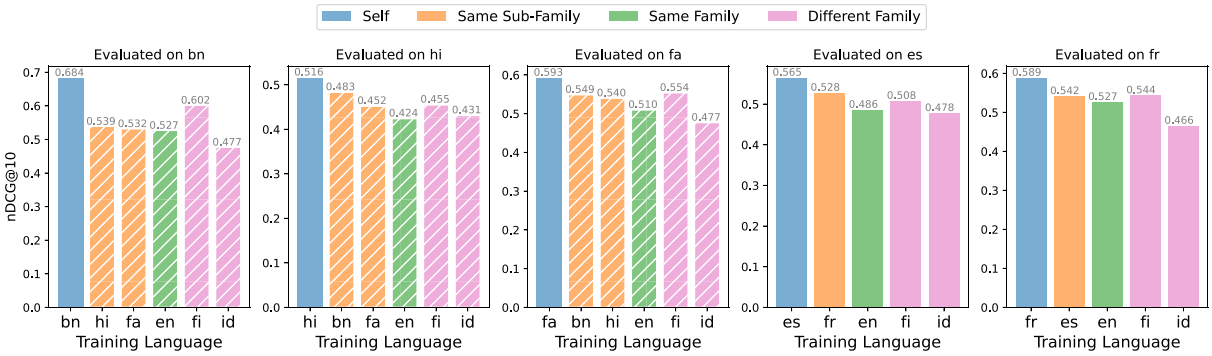


Figure 5: Case study of how language kinship affects cross-lingual transfer. Results are grouped according to the target language and each bar indicates a source language (used for fine-tuning mDPR). Within each panel, the relationship between the source and target languages moves from closer to more distant from left to right (self, same language sub-family, same language family, different language family). Striped bars denote that the source language has a different script from the target language. The exact nDCG@10 score is shown at the top of each bar.

- **In-language fine-tuned mDPR** follows the same model configuration as the mDPR baseline, but we fine-tuned each model on the MIRACL training set of the target language rather than MS MARCO. Here, the negative examples are sampled from the labeled negatives and the unlabeled top-30 candidates from BM25.

Note that all languages in MIRACL are included in the pretraining corpus of mBERT, which provides the backbone of the mDPR and mColBERT models. However, three languages (`fa`, `hi`, `yo`) are not included in CCNet (Wenzek et al., 2020), the dataset used by mContriever for additional pretraining. Code, documentation, and instructions for reproducing these baselines have been released together with the MIRACL dataset and can be found on the MIRACL website.

These results provide a snapshot of the current state of research. We see that across these diverse languages, mDPR does not substantially outperform decades-old BM25 technology, although BM25 exhibits much lower effectiveness in French and Chinese. Nevertheless, the BM25–mDPR hybrid provides a strong zero-shot baseline that outperforms all the other individual models on average nDCG@10. Interestingly, the BM25–mDPR hybrid even outperforms in-language fine-tuned mDPR on most of the languages, except for the ones that are comparatively under-represented in mBERT pretraining (e.g., `sw`, `th`, `hi`). Overall, these results show that plenty of work remains to be done to advance multilingual retrieval.

## 6.2 Cross-Lingual Transfer Effects

As suggested in Section 5.3, MIRACL enables the exploration of the linguistic factors influencing multilingual transfer; in this section, we present a preliminary study. Specifically, we evaluate cross-lingual transfer effectiveness on two groups of *target* languages: (A) {`bn`, `hi`, `fa`} and (B) {`es`, `fr`}, where the models are trained on different source languages drawn from (with respect to the target language): (1) different language families, (2) same language family but different sub-families, and (3) same sub-family. We add the ‘‘self’’ condition as an upper-bound, where the model is trained on data in the target language itself.

The evaluation groups are divided based on the script of the language: in group (A), all the source languages are in a different script from the target languages, whereas in group (B), all the source languages are in the same script as the target language (Latin script in this case). In these experiments, we reused checkpoints from the ‘‘in-language fine-tuned mDPR’’ condition in Section 6.1. For example, when the source language is `hi` and the target language is `bn`, we encode the `bn` query and the `bn` corpus using the `hi` checkpoint; this corresponds to the `hi` row, **in-L.** column in Table 5.

The results are visualized in Figure 5, where each panel corresponds to the target language indicated in the header. Each bar within a panel represents a different source language and the $y$-axis shows the zero-shot nDCG@10 scores. Within each panel, from left to right, the kinship

relationship between the source and target languages moves from close to distant, where the source languages that are at the same distance to the target languages are colored the same (self, same language sub-family, same language family, different language family). Striped bars denote that the source language is in a different script from the target language.

We find that languages from the same sub-families show better transfer capabilities in general, regardless of the underlying script. Among the five languages studied, those from the same sub-families achieve the best transfer scores (i.e., the orange bars are the tallest). The only exception appears to be when evaluating bn using fi as the source language (the leftmost panel). Interestingly, the source language being in the same family (but a different sub-family) does not appear to have an advantage over the languages in different families (i.e., the green bars versus the pink bars). This suggests that transfer effects only manifest when the languages are closely related within a certain degree of kinship.

We also observe that transfer effects do not appear to be symmetrical between languages. For example, the model trained on bn achieves 0.483 and 0.549 on hi and fa, respectively, which are over 90% of the ''self'' score on hi and fa (the first orange bar in the second and third panels). However, models trained on hi and fa do not generalize on bn to the same degree, scoring less than 80% of the ''self'' score on bn (the two orange bars in the first panel).

We emphasize that this is merely a preliminary study on cross-lingual transfer effects with respect to language families and scripts, but our experiments show the potential of MIRACL for further understanding multilingual models.

## 7 Conclusion

In this work, we present MIRACL, a new high-quality multilingual retrieval dataset that represents approximately five person-years of annotation effort. We provide baselines and present initial explorations demonstrating the potential of MIRACL for studying interesting scientific questions. Although the WSDM Cup challenge associated with our efforts has ended, we continue to host a leaderboard to encourage continued participation from the community.

While MIRACL represents a significant stride towards equitable information access, further efforts are necessary to accomplish this objective. One obvious future direction is to extend MIRACL to include more languages, especially low-resource ones. Another possibility is to augment MIRACL with cross-lingual retrieval support. However, pursuing these avenues demands additional expenses and manual labor.

Nevertheless, MIRACL already provides a valuable resource to support numerous research directions: It offers a solid testbed for building and evaluating multilingual versions of dense retrieval models (Karpukhin et al., 2020), late-interaction models (Khattab and Zaharia, 2020), as well as reranking models (Nogueira and Cho, 2019; Nogueira et al., 2020), and will further accelerate progress in multilingual retrieval research (Shi et al., 2020; MacAvaney et al., 2020; Nair et al., 2022; Zhang et al., 2022). Billions of speakers of languages that have received relatively little attention from researchers stand to benefit from improved information access.

## Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable feedback. We would also like to thank our annotators, without whom MIRACL could not have been built. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, a gift from Huawei, and Cloud TPU support from Google's TPU Research Cloud (TRC).

## References

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. https://doi.org/10.18653/v1/2021.naacl-main.46

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, 31(1):30–54. https://doi.org/10.1093/llc/fqu047

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. *arXiv: 1611.09268v3*.

Luiz Henrique Bonifacio, Israel Campiotti, Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv:2108.13897*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Nick Craswell, Bhaskar Mitra, Daniel Campos, Emine Yilmaz, and Jimmy Lin. 2021. MS MARCO: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 1566–1576. https://doi.org/10.1145/3404835.3462804

Hope Dawson, Antonio Hernandez, and Cory Shain. 2016. Morphological types of languages. In *Language Files: Materials for an Introduction to Language and Linguistics, 12th Edition*. Department of Linguistics, The Ohio State University. https://doi.org/10.26818/9780814252703

Sauleh Eetemadi and Kristina Toutanova. 2014. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 159–164, Doha, Qatar. https://doi.org/10.3115/v1/D14-1018

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, New York, NY, USA.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Tevatron: An efficient and flexible toolkit for Neural Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3120–3124.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1029

Joseph Harold Greenberg. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26:178–194. https://doi.org/10.1086/464575

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Alexander Jones, William Yang Wang, and Kyle Mahowald. 2021. A massively multilingual analysis of cross-linguality in shared embedding space. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.471

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver,

Canada. https://doi.org/10.18653/v1/P17-1147

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. https://doi.org/10.18653/v1/2020.acl-main.560

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 39–48. https://doi.org/10.1145/3397271.3401075

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the TREC 2022 NeuCLIR track. In *Proceedings of the 31st Text REtrieval Conference*.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825. https://doi.org/10.1162/COLI_a_00111

Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2022. Fostering coopetition while plugging leaks: The design and implementation of the MS MARCO leaderboards. In *Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022)*, pages 2939–2948, Madrid, Spain. https://doi.org/10.1145/3477495.3531725

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. 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)*, pages 2356–2362. https://doi.org/10.1145/3404835.3463238

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021b. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02181-7

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406. https://doi.org/10.1162/tacl_a_00433

Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In *Proceedings of the 42nd European Conference on Information Retrieval, Part II (ECIR 2020)*, pages 246–254. https://doi.org/10.1007/978-3-030-45442-5_31

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 *Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022), Part I*, pages 382–396, Stavanger, Norway. https://doi.org/10.1007/978-3-030-99736-6_26

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.63

Damaris Nübling. 2020. Inflectional morphology. In *The Cambridge Handbook of Germanic Linguistics*. Cambridge University Press. https://doi.org/10.1017/9781108378291.011

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. https://doi.org/10.18653/v1/P19-1493

Frans Plank. 1999. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 3:279–340. https://doi.org/10.1515/lity.1999.3.3.279

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601. https://doi.org/10.1162/coli_a_00357

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847. https://aclanthology.org/2021.naacl-main.466

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432. https://doi.org/10.1162/tacl_a_00148

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. https://doi.org/10.18653/v1/D16-1264

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389. https://doi.org/10.1561/1500000019

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. https://doi.org/10.18653/v1/2022.naacl-main.272

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.249

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)*, pages 4160–4170, Online. https://doi.org/10.18653/v1/2020.emnlp-main.340

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Neural Information Processing Systems: Datasets and Benchmarks Track*.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118. https://doi.org/10.1093/llc/fqt031

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the*

*21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 315–323, Melbourne, Australia. https://doi.org/10.1145/290941.291017

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16. https://doi.org/10.1145/3239571

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. https://doi.org/10.18653/v1/2021.mrl-1.12

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards best practices for training multilingual dense retrieval models. *arXiv:2204.02363*.