# Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval

**Nandan Thakur**[*†§], **Jianmo Ni**[†♡], **Gustavo Hernández Ábrego**[◊]
**John Wieting**[♡], **Jimmy Lin**[§], **Daniel Cer**[†◊]

[◊]Google Research, [♡]Google DeepMind, [§]University of Waterloo

## Abstract

There has been limited success for dense retrieval models in multilingual retrieval, due to uneven and scarce training data available across multiple languages. Synthetic training data generation is promising (e.g., InPars or Promptagator), but has been investigated only for English. Therefore, to study model capabilities across both cross-lingual and monolingual retrieval tasks, we develop **SWIM-IR**, a synthetic retrieval training dataset containing 33 (high to very-low resource) languages for fine-tuning multilingual dense retrievers without requiring any human supervision. To construct SWIM-IR, we propose SAP (*summarize-then-ask prompting*), where the large language model (LLM) generates a textual summary prior to the query generation step. SAP assists the LLM in generating informative queries in the target language. Using SWIM-IR, we explore synthetic fine-tuning of multilingual dense retrieval models and evaluate them robustly on three retrieval benchmarks: XOR-Retrieve (cross-lingual), MIRACL (monolingual) and XTREME-UP (cross-lingual). Our models, called SWIM-X, are competitive with human-supervised dense retrieval models, e.g., mContriever-X, finding that SWIM-IR can cheaply substitute for expensive human-labeled retrieval training data. SWIM-IR dataset and SWIM-X models are available at: https://github.com/google-research-datasets/SWIM-IR.

## 1 Introduction

Dense retrieval models have demonstrated impressive performance in ad-hoc information retrieval (IR) tasks, e.g., web search, outperforming traditional retrieval systems such as BM25 (Karpukhin et al., 2020; Lin et al., 2021; Ni et al., 2022; Neelakantan et al., 2022, *inter alia*). A major reason for its success lies in the availability of large-scale supervised training datasets in English, such as MS MARCO (Nguyen et al., 2016) or NQ (Kwiatkowski et al., 2019), and coupled with effective training strategies, such as custom hard-negative mining (Xiong et al., 2021; Lin et al., 2023), or teacher distillation (Hofstätter et al., 2021; Ren et al., 2021).

However, there is a limited exploration of dense retrieval models in multilingual retrieval,[1] due to

| Dataset | Q Gen. | Cross. | Mono. | # L | Domain | # Train |
|---|---|---|---|---|---|---|
| NeuCLIR | Human | EN→L | L→L | 3 | News (hc4) | × |
| MKQA | Human | L→EN | × | 26 | Wikipedia | 10K |
| mMARCO | Translate | × | L→L | 13 | MS MARCO | 533K |
| Mr.TyDI | Human | × | L→L | 11 | Wikipedia | 49K |
| MIRACL | Human | × | L→L | 18 | Wikipedia | 726K |
| JH-POLO | GPT-3 | EN→L | × | 3 | News (hc4) | 78K |
| **SWIM-IR** | **PaLM 2** | **L→EN** | **L→L** | **33** | **Wikipedia** | **28M** |

Table 1: We construct SWIM-IR, a "synthetic" multilingual dataset with 28 million PaLM 2 generated training pairs across 33 languages in our work; (Q Gen.) denotes the query generation technique; (Cross. and Mono.) denotes the retrieval task and (query→document) language pair; (# L and # Train) denotes the language count and available training pairs.

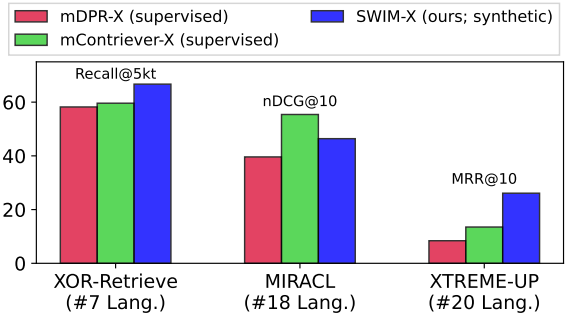


Figure 1: Summary of the quantitative results across three multilingual retrieval benchmarks evaluated in our work. SWIM-X is fine-tuned on SWIM-IR (PaLM 2 generated synthetic training data) without any human supervision. All scores are macro-averaged.

[*]Work done while Nandan was a student researcher at Google Research. [†]Correspondence to: Nandan Thakur <nandan.thakur@uwaterloo.ca>, Jianmo Ni <jianmon@google.com>, Daniel Cer <cer@google.com>.

[1]Throughout the paper, we use "multilingual retrieval" to col-



*Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7699–7724
June 16-21, 2024 

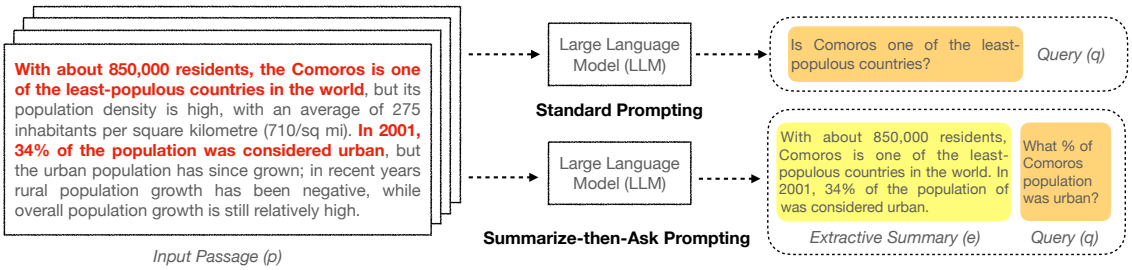


Figure 2: An illustration of SAP (*Summarize-then-Ask Prompting*) versus standard prompting for English query generation on English Wikipedia. SAP assists the LLM in improving the query generation quality (orange box) by identifying the relevant sections of the input passage (highlighted in red) via the extractive summarization (yellow box) as an intermediate reasoning step.

uneven and low distribution of human-supervised training data for other languages apart from English (Reimers and Gurevych, 2020; Feng et al., 2022; Wieting et al., 2023). Collecting human annotations for training data generation is not scalable, as it is cumbersome to search and hire native speakers, check their language proficiency, and teach them. Additionally, human annotators are expensive, thereby requiring a large annotation budget for generating a sufficient amount of training pairs (cf. Figure 6).

Multilingual query generation is a complex task (Wang et al., 2021). It requires understanding of semantic mappings of words across languages, similar to machine translation (Forcada, 2002; Tan et al., 2019; Zhu et al., 2023). Recently, utilizing LLMs for query generation has been popular in English (Bonifacio et al., 2022; Dai et al., 2023). But as illustrated in Figure 2, standard prompt templates can lead the LLM to generate either extractive or uninformative[2] queries across languages.

To improve the quality of the generated query, we propose SAP (*Summarize-then-Ask Prompting*), where we optimize the prompt to break down the query generation with LLM in two stages: (i) *summary extraction*, which identifies the relevant information from the long input passage and extracts the best representative sentences as the summary, and (ii) *query generation*, which generates a multilingual query relevant for the input passage, using the extracted summary (first stage) as the intermediate step. SAP highlights the relevant information within the passage and produces difficult (i.e., informative) queries in the target language.

In our work, we utilize PaLM 2 (Anil et al., 2023), a recent multilingual LLM (successor of PaLM 540B (Chowdhery et al., 2023)) for query generation. The generated query paired with the original passage from Wikipedia is used to construct the SWIM-IR dataset. SWIM-IR provides synthetic training (query-passage) pairs for improving dense retrieval models without requiring any human supervision. The dataset spans across 33 diverse languages, including both high and very-low resource languages and is one of the largest multilingual synthetic training dataset with 28 million training pairs (cf. Table 1).

We develop synthetic multilingual (both monolingual and cross-lingual) dense retrieval models called SWIM-X, using mT5 (base) (Xue et al., 2021) as the backbone and fine-tune on SWIM-IR. We compare SWIM-X against models fine-tuned with human supervision by changing only the training dataset while keeping other, i.e., model parameters and training settings unchanged. We evaluate on three standard multilingual retrieval benchmarks (two cross-lingual and one monolingual). As shown in Figure 1, on XOR-Retrieve (Asai et al., 2021a), SWIM-X outperforms the best-supervised baseline (mContriever-X) by 7.1 points at Recall@5kt. On MIRACL (Zhang et al., 2023b), a monolingual retrieval benchmark, SWIM-X is inferior to mContriever-X by 9.0 points at nDCG@10, which shows room for future improvement. On XTREME-UP (Ruder et al., 2023), a challenging benchmark containing 20 underrepresented Indo-European languages, SWIM-X outperforms mContriever-X by 11.7 points at MRR@10. We publicly open-source SWIM-IR dataset and SWIM-X models at `https://github.com/google-research-datasets/SWIM-IR`.

## 2 SWIM-IR Dataset Overview

In our dataset overview, we first describe the SAP design formulation for multilingual query generation (§2.1), data construction details (§2.2), and fi-

lectively denote both cross-language, i.e., cross-lingual and within language, i.e., monolingual retrieval tasks.

[2] *Uninformative* denotes a query that can be easily answered using the first (or last) few words in the passage.

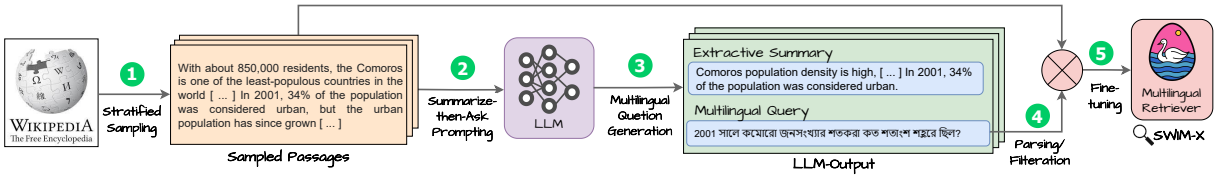


Figure 3: An illustration of the cross-lingual SWIM-IR dataset construction procedure. Steps are as follows: (1) Sample N passages from the English Wikipedia using stratified sampling for each language out of the L languages; (2) Feed a sampled passage along with the few-shot exemplars to the LLM with SAP; (3 & 4) Parse the LLM output to receive the synthetic query in the target language (above in Bengali); (5) Fine-tune a multilingual dense retriever model (SWIM-X) with training pairs combined for all languages, i.e., N×L pairs.

nally discuss about human validation and content filtration (§2.3).

### 2.1 SAP Design Formulation

Multilingual query generation is not a trivial task as it requires a deep understanding of the passage content and its own translations across different languages (Wang et al., 2021). Also, passages can often be lengthy and contain information on multiple topics. Using the entire passage can potentially cause hallucinations in models by generating non-meaningful queries, which affects the retrieval performance (Gospodinov et al., 2023).

To break down the task complexity of multilingual query generation and improve the query quality, we implement summarize-then-ask prompting (SAP). As shown above in Figure 2, we identify the relevant information within a passage by asking the LLM to generate an extractive summary and use it as an intermediate step for generating informative queries (Wei et al., 2023). The procedure is described in more detail below:

**(i) Summary extraction.** The LLM constructs an extractive summary $e_s$ of the input passage $p_s$, where $s$ denotes the source language. The summary captures the highly relevant information contained within the passage $p_s$ (which occasionally may be long) acting as an useful intermediate signal for the LLM to generate a multilingual query in the later stage. We denote the first stage as $e_s = \text{LLM}(p_s; \theta^1, \cdots, \theta^k)$, where $(\theta^1, \cdots, \theta^k)$ denotes the $k$ few-shot prompt exemplars[3] containing the passage, summary in the source language $s$ and the query in the target language $t$.[4]

**(ii) Query generation.** Next, the LLM combines the summary $e_s$ generated previously with the original input passage $p_s$, highlighting the relevant information required for composing the query $q_t$ in the target language $t$. We denote this stage as $q_t = \text{LLM}(e_s, p_s; \theta^1, \cdots, \theta^k)$, where extractive summary $e_s$, input passage $p_s$ and $k$-shot exemplars all appear from the first stage.

### 2.2 SWIM-IR Dataset Construction

For constructing SWIM-IR, we only require an unlabeled corpus of passages and few-shot exemplars. An overview of the cross-lingual generation procedure is shown in Figure 3. Prompt examples are provided in the Appendix (§C.3).

**Cross-lingual.** The goal is to generate a query in the target language $t$ using the input passage in English (source language $s$). We use a stratified sampling algorithm (for more details, refer to §E.4 in the Appendix) to sample a maximum of one million passages for each target language $t$ from the English Wikipedia corpus used in XOR-Retrieve (Clark et al., 2020; Asai et al., 2021a) or XTREME-UP (Ruder et al., 2023). Next, we construct five prompt exemplars and manually construct both the summary and query for the exemplar in English. Further, we use Google Translate[5] to translate the exemplar queries across other languages. Finally, we construct the prompt, where we explain our query generation task as an instruction, include the target language, and the 5-shot exemplars as an input to the LLM with SAP.

**Monolingual.** The goal is to generate a query in the same language as the input passage ($s = t$). We follow the setting similar to the cross-lingual task. We first sample one million passages (if available) for each language-specific Wikipedia corpus in MIRACL (Zhang et al., 2023b).[6] Next, we carefully select three training pairs as our

[3]Multilingual query generation requires few-shot prompt exemplars. As our experiments show in (§4), zero-shot prompting often generates unparseable outputs with PaLM 2.

[4]In our work, we did not use abstractive summarization, as LLMs have notoriously been shown to hallucinate and generate factual inconsistencies in their output generations (Maynez et al., 2020; Liu et al., 2023).

[5]Google Translate: translate.google.com

[6]For 16 out of the 18 languages, MIRACL contains a training split except for two: German (de) and Yoruba (yo).

| Benchmark | Retrieval Task | Evaluation Metric | Query → Passage | # L | ISO | Languages | Train Split # Q | Train Split HNeg. | Dev/Test Split # Q | Dev/Test Split # Passages |
|---|---|---|---|---|---|---|---|---|---|---|
| XOR-Retrieve (Asai et al., 2021a) | Cross-lingual | Recall@5kt | L → English | 7 | ar, bn, fi, ja, ko, ru, te | Arabic, Bengali, Finnish, Japanese, Korean, Russian, Telugu | 15,250 | Yes (1 each) | 2,110 | 18,003,200 |
| MIRACL (Zhang et al., 2023b) | Monolingual | nDCG@10 | L → L | 18 | ar, bn, de, en, es, fa, fi, fr, hi, id, ja, ko, ru, sw, te, th, yo, zh | Arabic, Bengali, German, English, Spanish, Farsi, Finnish, French, Hindi, Indonesian, Japanese, Korean, Russian, Swahili, Telugu, Thai, Yoruba, Chinese | 88,288 | Yes (max 4) | 13,495 | 106,332,152 |
| XTREME-UP (Ruder et al., 2023) | Cross-lingual | MRR@10 | L → English | 20 | as, bho, brx, gbm, gom, gu, hi, hne, kn, mai, ml, mni, mr, mwr, or, pa, ps, sa, ta, ur | Assamese, Bhojpuri, Boro, Garhwali, Konkani, Gujarati, Hindi, Chhattisgarhi, Kannada, Maithili, Malayalam, Manipuri, Marathi, Marwari, Odia, Punjabi, Pashto, Sanskrit, Tamil, Urdu | 13,270 | No | 5,300 | 112,426 |

Table 2: Overview of the multilingual retrieval evaluation benchmarks used in our work: (i) XOR-Retrieve (Dev) (Asai et al., 2021a), (ii) MIRACL (Dev) (Zhang et al., 2023b) and (iii) XTREME-UP (Test) (Ruder et al., 2023); (HNeg.) denotes availability of hard negatives for fine-tuning; (# L) denotes the number of languages covered by the benchmark; (# Q) denotes the number of queries in each dataset split.

| Lang. (ISO) | fluency (↑) | | | adequacy (↑) | | | language (↑) | | |
|---|---|---|---|---|---|---|---|---|---|
| Rating (→) | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| English (en) | 2% | 3% | 95% | 2% | 13% | 85% | 0% | 0% | 100% |
| Spanish (es) | 1% | 10% | 89% | 14% | 12% | 74% | 1% | 0% | 99% |
| Chinese (zh) | 7% | 19% | 74% | 7% | 30% | 63% | 0% | 0% | 100% |
| Hindi (hi) | 12% | 5% | 83% | 6% | 19% | 75% | 0% | 0% | 100% |
| Bengali (bn) | 6% | 4% | 90% | 10% | 14% | 76% | 1% | 0% | 99% |

Table 3: Human validation statistics on SWIM-IR. Annotators evaluate the quality of the generated query on a three-level rating scale (0/1/2) based on thee factors: (i) fluency, (ii) adequacy and (iii) language.

prompt exemplars.[7] For languages with no training split, we manually construct our prompt exemplars. Further, we use Google Bard[8] to generate exemplar summaries in the target language. Finally, we construct the prompt, where we explain our query generation task as an instruction, and the 5-shot exemplars with SAP.

### 2.3 Human Validation & Content Filtration

**Human validation.** The goal of our query generation is to generate an adequate and fluent query according to a given passage (Qiu and Xiong, 2019). To evaluate the intrinsic query quality, we conduct a validation study in SWIM-IR on a subset of five languages.[9] Within the five evaluated languages, three are high-resource, one medium-resource and one low-resource. For each language, we randomly sample a fixed amount of query-passage pairs resulting in a overall sum of 500 evaluation pairs to be human validated across all languages.

We compute the query quality on a three-level rating scheme (0/1/2) based on three evaluation criteria: fluency, adequacy, and language. (i) *fluency*, measures the coherence of the generated query, i.e., whether the query is understandable and readable by the user and contains no spelling or grammatical mistakes. (ii) *adequacy*, measures the relevancy of the query with passage (used for query generation) (iii) *language*, detects the language of the generated query, or whether code-switching occurs in the generated query.

**Validation statistics.** Table 3 reports the human validation statistics. For fluency, major mistakes are observed in Hindi (12%) and Chinese (7%), where the passage sampled in MIRACL (Zhang et al., 2023b) can be too short (only 2–3 words long), this leads to the exact duplication of the exact text in the query. For adequacy, we observe that in Chinese (30%) of the queries are not relevant to the passage. Similar to fluency, a low adequacy is observed in cases when either query is generated for a short passage or when the query is about a related topic which is not directly referenced within the passage. Finally for language, annotators achieve between 99–100% for all languages indicating PaLM 2 is likely to generate the query in the correct language.

**Content filtration**. LLMs have been shown to generate undesirable content, particularly under conditions that prime the model with material targeted at drawing out any negative patterns or associations in the training data (Gehman et al., 2020; Bender et al., 2021). To avoid this, we use the Google Cloud Natural Language content classification categories[10] to filter out harmful content present within the SWIM-IR training pairs. We discard samples with a high content classification of either `/Adult` or any of the `/Sensitive Subjects` labels. For more details on content filtration, refer to (§D) in the Appendix.

[7] As language-specific passages consume more tokens, e.g., Telugu, to save computational budget, we rely only on 3-shot exemplars (instead of five) for the monolingual task.

[8] Google Bard: bard.google.com

[9] The authors in the paper are native speakers of the five languages used for evaluation: English (en), Bengali (bn), Spanish (es), Chinese (zh) and Hindi (hi).

[10] cloud.google.com/natural-language/docs/categories

| Model | PLM | PT | Finetune (Datasets) | Recall@5kt | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Avg. | Ar | Bn | Fi | Ja | Ko | Ru | Te |
| *Existing Supervised Baselines (Prior work)* | | | | | | | | | | | |
| Dr. DECR (Li et al., 2022) | XLM-R | WikiM | NQ + XOR* | 73.1 | 70.2 | 85.9 | 69.4 | 65.1 | 68.8 | 68.8 | 83.2 |
| mDPR (Asai et al., 2021a) | mBERT | — | XOR | 50.2 | 48.9 | 60.2 | 59.2 | 34.9 | 49.8 | 43.0 | 55.5 |
| mBERT + xQG (Zhuang et al., 2023) | mBERT | — | XOR | 53.5 | 42.4 | 54.9 | 54.1 | 33.6 | 52.3 | 33.8 | 52.5 |
| Google MT + DPR (Asai et al., 2021a) | BERT | — | NQ | 69.6 | 69.6 | 82.2 | 62.4 | 64.7 | 68.8 | 60.8 | 79.0 |
| OPUS MT + DPR (Asai et al., 2021a) | BERT | — | NQ | 50.6 | 52.4 | 62.8 | 61.8 | 48.1 | 58.6 | 37.8 | 32.4 |
| *Zero-shot baselines (English-only supervision)* | | | | | | | | | | | |
| mContriever | mT5 | mC4 | — | 38.9 | 35.9 | 33.9 | 43.6 | 34 | 35.1 | 45.1 | 44.5 |
| mDPR-EN | mT5 | — | MS MARCO | 39.3 | 34.3 | 35.5 | 45.2 | 40.2 | 36.5 | 43.9 | 39.5 |
| mContriever-EN | mT5 | mC4 | MS MARCO | 44.0 | 37.5 | 38.2 | 50.6 | 41.1 | 37.2 | 49.8 | 53.8 |
| *Supervised Baselines (Cross-lingual supervision)* | | | | | | | | | | | |
| mDPR-X | mT5 | — | XOR | 53.6 | 51.5 | 63.5 | 52.5 | 45.6 | 52.3 | 43.0 | 66.8 |
| mContriever-X | mT5 | mC4 | XOR | 55.3 | 52.1 | 68.1 | 54.5 | 47.7 | 50.5 | 50.2 | 64.3 |
| mDPR-X | mT5 | — | MS MARCO + XOR | 58.2 | 55.3 | 70.1 | 56.7 | 49.8 | 55.8 | 50.6 | 69.3 |
| mContriever-X | mT5 | mC4 | MS MARCO + XOR | 59.6 | 54.7 | 73.4 | 57.0 | 53.1 | 56.5 | 51.5 | 71.0 |
| *Synthetic Baselines (Our work)* | | | | | | | | | | | |
| SWIM-X (500K) | mT5 | — | SWIM-IR | 59.0 | 54.0 | 67.4 | 59.2 | 52.7 | 55.1 | 54.4 | 70.2 |
| SWIM-X (500K) | mT5 | mC4 | SWIM-IR | 63.0 | 57.0 | 71.1 | 61.8 | 56.8 | 60.7 | 63.3 | 70.2 |
| SWIM-X (7M) | mT5 | — | SWIM-IR | 65.1 | 57.9 | 75.0 | 65.6 | 59.3 | 58.9 | 64.6 | 74.4 |
| SWIM-X (7M) | mT5 | mC4 | SWIM-IR | 66.7 | 61.2 | 77.0 | 65.0 | 62.2 | 62.8 | 65.4 | 73.5 |

Table 4: Experimental results showing Recall@5kt for cross-lingual retrieval on XOR-Retrieve dev (Asai et al., 2021a); (PLM) denotes the pre-trained language model; (PT) denotes the pre-training dataset; (*) Dr.DECR is fine-tuned in a complex training setup across more datasets (§3.3); WikiM denotes WikiMatrix (Schwenk et al., 2021); XOR denotes XOR-Retrieve; SWIM-X (ours) is fine-tuned on 500K and 7M synthetic data.

# 3 Experiments

## 3.1 Datasets and Metrics

We evaluate on three multilingual retrieval benchmarks: (i) **XOR-Retrieve** (Asai et al., 2021a), (ii) **MIRACL** (Zhang et al., 2023b) and (iii) **XTREME-UP** (Ruder et al., 2023). XOR-Retrieve and XTREME-UP are cross-lingual and MIRACL is monolingual. Following prior work, we evaluate models at Recall@5kt on XOR-Retrieve, nDCG@10 on MIRACL and MRR@10 on XTREME-UP. An overview of the evaluation dataset statistics is available in Table 2. For additional details, refer to the Appendix (§F).

## 3.2 Experimental Methods

**Baseline categories.** Following common practice across all datasets, we evaluate three range of baselines: (i) *Zero-shot baselines*: where the model denoted by "EN" (model-EN) is fine-tuned using supervised English-only training data such as MS MARCO (Nguyen et al., 2016) or NQ (Kwiatkowski et al., 2019). (ii) *Supervised baselines*: where the model denoted by "X" (model-X) is fine-tuned on human-supervised, i.e., multilingual training data. (iii) *Synthetic baselines*: where the model denoted by "SWIM-X" is fine-tuned without any supervision, relying purely on synthetic multilingual training data. Additionally, we report the amount of synthetic pairs, e.g., SWIM-X (500K) is fine-tuned on 500K training pairs.

**Model choices.** For our dense retrieval models, we adapt DPR (Karpukhin et al., 2020) to the multilingual setting with the mT5-base (Xue et al., 2021) language model with 580M parameters. Next, we include mContriever (Izacard et al., 2022) which adopts an additional pre-training stage with contrastive loss based on unsupervised data prepared from pairwise sentence cropping in mC4 (Xue et al., 2021). For query generation, we use PaLM 2 (S) (Anil et al., 2023) for efficient generation due to its low-cost and inference latency.

**Existing baselines.** For XOR-Retrieve, we include Dr. DECR (Li et al., 2022), a cross-lingual ColBERT (Khattab and Zaharia, 2020) fine-tuned on a large amount of supervised data in a computationally expensive setup involving knowledge distillation with English ColBERTv2 (Santhanam et al., 2022). xQG (Zhuang et al., 2023) involves cross-language query generation and concatenating the queries along with the passage representation. We also include two-stage translation baselines, Google Translate and Opus-MT from Asai et al. (2021a). For MIRACL, we include the official BM25, mDPR and Hybrid (combining BM25, mDPR and mColBERT) baselines (Zhang et al., 2023b), and Cohere-API is used as a reranker with top-100 BM25 results (Kamalloo et al., 2023).

## 3.3 Training Methodology

**Zero-shot & supervised baselines.** We replicate mContriever and mDPR zero-shot baselines by ini-

| Model | Avg. | ar | bn | en | es | fa | fi | fr | hi | id | ja | ko | ru | sw | te | th | zh | de | yo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Existing Supervised Baselines (Prior work)* | | | | | | | | | | | | | | | | | | | |
| BM25 | 38.5 | 48.1 | 50.8 | 35.1 | 31.9 | 33.3 | 55.1 | 18.3 | 45.8 | 44.9 | 36.9 | 41.9 | 33.4 | 38.3 | 49.4 | 48.4 | 18.0 | 22.6 | 40.6 |
| mDPR | 41.8 | 49.9 | 44.3 | 39.4 | 47.8 | 48.0 | 47.2 | 43.5 | 38.3 | 27.2 | 43.9 | 41.9 | 40.7 | 29.9 | 35.6 | 35.8 | 51.2 | 49.0 | 39.6 |
| Hybrid | 56.6 | 67.3 | 65.4 | 54.9 | 64.1 | 59.4 | 67.2 | 52.3 | 61.6 | 44.3 | 57.6 | 60.9 | 53.2 | 44.6 | 60.2 | 59.9 | 52.6 | 56.5 | 37.4 |
| Cohere-API | 54.2 | 66.7 | 63.4 | 50.1 | 50.7 | 48.4 | 67.5 | 44.3 | 57.3 | 50.5 | 51.6 | 54.6 | 47.7 | 54.3 | 63.8 | 60.6 | 38.9 | 41.4 | 62.9 |
| *Zero-shot baselines (English-only supervision)* | | | | | | | | | | | | | | | | | | | |
| mDPR-EN | 39.8 | 49.7 | 50.1 | 35.4 | 35.3 | 39.3 | 48.2 | 31.3 | 37.4 | 35.6 | 38.9 | 44.1 | 36.1 | 33.8 | 49.2 | 50.6 | 34.7 | 32.1 | 34.4 |
| mContriever-EN | 37.8 | 49.1 | 48.4 | 32.7 | 33.3 | 37.1 | 48.4 | 27.0 | 35.9 | 32.7 | 34.1 | 40.2 | 35.1 | 44.5 | 46.2 | 45.0 | 27.5 | 29.7 | 33.7 |
| *Supervised Baselines (Monolingual supervision)* | | | | | | | | | | | | | | | | | | | |
| mDPR-X | 39.6 | 52.8 | 57.1 | 30.2 | 24.7 | 37.6 | 46.1 | 26.4 | 27.8 | 37.3 | 42.9 | 38.3 | 34.9 | 53.7 | 68.4 | 58.2 | 34.9 | 19.2 | 22.2 |
| mContriever-X | 55.4 | 66.4 | 68.4 | 44.2 | 42.8 | 48.9 | 65.2 | 46.2 | 45.0 | 45.8 | 56.8 | 58.8 | 51.2 | 67.7 | 79.0 | 70.7 | 49.4 | 42.3 | 48.4 |
| *Synthetic Baselines (Our work)* | | | | | | | | | | | | | | | | | | | |
| SWIM-X (180K) | 46.4 | 60.2 | 57.1 | 34.7 | 33.4 | 36.3 | 40.6 | 64.3 | 33.0 | 39.5 | 40.8 | 43.3 | 49.7 | 40.0 | 55.9 | 56.3 | 63.3 | 50.2 | 36.5 |

Table 5: Experimental results for monolingual retrieval on MIRACL dev (Zhang et al., 2023b). All scores denote **nDCG@10**; (Hyb.) denotes Hybrid retriever with ranked fusion of three retrievers: mDPR, mColBERT and BM25; BM25, mDPR and Hybrid scores taken from (Zhang et al., 2023b); Cohere-API is used as a reranker on top of 100 BM25 results, taken from (Kamalloo et al., 2023). SWIM-X (ours) is fine-tuned on 180K synthetic training pairs.

tializing from an mT5-base checkpoint (Xue et al., 2021) and further fine-tuning on MS MARCO, following a setup similar to Ni et al. (2022). Similarly, mContriever-X and mDPR-X have been additionally fine-tuned on training split available for each dataset. For additional technical details on supervised baselines, refer to the Appendix (§E.2). As mContriever includes an additional pre-training stage, we set the batch size to 8192, learning rate to $1e^{-3}$ and pre-train for 600K steps with mC4 (Xue et al., 2021). For more details on pre-training, refer to the Appendix (§E.1).

**Synthetic baselines.** For SWIM-X, we pre-train the mT5-base checkpoint on mC4 (Xue et al., 2021) for 600K steps using a contrastive loss function objective, similar to Contriever (Izacard et al., 2022). Next, we fine-tune the pre-trained mT5-base model on SWIM-IR with in-batch negatives and a contrastive loss function (van den Oord et al., 2018). During fine-tuning, we set the batch size to 4096, learning rate to $1e^{-3}$ and fine-tune between 5K to 50K training steps, depending upon the size of the training dataset. For technical details on synthetic baselines, refer to the Appendix (§E.3).

### 3.4 Experimental Results

**XOR-Retrieve.** Table 4 shows that SWIM-X (7M), fine-tuned on 7M synthetic pairs (max. of 1M per language) outperforms the best supervised baseline, mContriever-X, by 7.1 points Recall@5kt. Without mC4 pre-training, SWIM-X (7M) performance drops by only 1.6 points. We also evaluate SWIM-X (500K), a limited-budget baseline fine-tuned on 500K training pairs, which outperforms mContriever-X by 3.6 points. Few existing baselines outperform SWIM-X, however, the comparison is not fair. For instance, Dr. DECR is a multilingual ColBERT (Khattab and Zaharia, 2020) model, which is computationally expensive at inference (Thakur et al., 2021). Similarly, Google MT + DPR relies on a Google Translate system for the translation of queries to English.

**MIRACL.** Table 5 shows that the SWIM-X (180K) model is competitive on MIRACL. SWIM-X (180K) outperforms the best zero-shot model by 6.6 points nDCG@10. However, SWIM-X underperforms mContriever-X on MIRACL, fine-tuned on 90K human-labeled training pairs with up to four hard negatives available in MIRACL by 9.0 points nDCG@10. This highlights the difficulty in the monolingual retrieval task, as models need to rely on human-supervision for improvement. Few existing baselines outperform SWIM-X, however the comparison is not fair. The Hybrid baseline relies on information based on aggregation of three models, and for Cohere-API, the underlying model information is unknown.

**XTREME-UP.** Table 6 shows the results on XTREME-UP. SWIM-X (120K) is fine-tuned by randomly selecting 5 exemplars from the XTREME-UP training dataset (human-labeled queries) for all languages, whereas the MT variant reuses XOR-Retrieve prompt exemplars with translated summaries and queries for 15 languages.[11] SWIM-X (120K)$^{MT}$ outperforms the best supervised baseline, mContriever-X$^{\heartsuit}$ (fine-tuned without MS MARCO) by a huge margin of 12.6 points MRR@10, but performs minimally worse than the MT version by 0.9 points. Interest-

[11] We were unable to translate our prompt exemplars for 5 languages due to language unavailability in Google Translate: Boro (brx), Garhwali (gbm), Chattisgarhi (hne) and Marwari (mwr). Manipuri (mni) is available in Google Translate in "Meitei" script instead of the "Bengali-Assamese" script present in the XTREME-UP dataset.

| Model | Avg. | as | bho | brx | gbm | gom | gu | hi | hne | kn | mai | ml | mni | mr | mwr | or | pa | ps | sa | ta | ur |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Zero-shot baselines (English-only supervision)* | | | | | | | | | | | | | | | | | | | | | |
| mDPR-EN | 6.3 | 2.6 | 6.4 | 0.4 | 7.2 | 1.3 | 8.6 | 13.3 | 5.2 | 10.4 | 6.4 | 12.3 | 0.2 | 8.9 | 5.8 | 0.4 | 6.0 | 5.6 | 5.2 | 10.2 | 10.0 |
| mContriever-EN | 7.9 | 7.9 | 3.2 | 7.8 | 0.3 | 9.7 | 2.2 | 11.1 | 15.2 | 8.2 | 10.6 | 8.6 | 15.6 | 0.4 | 10.7 | 8.5 | 1.1 | 10.3 | 3.3 | 5.7 | 12.9 |
| *Supervised Baselines (Cross-lingual supervision)* | | | | | | | | | | | | | | | | | | | | | |
| mDPR-X | 8.4 | 6.7 | 9.9 | 4.8 | 10.0 | 8.7 | 8.8 | 9.1 | 9.4 | 9.0 | 10.0 | 10.5 | 4.8 | 7.8 | 9.6 | 6.9 | 8.6 | 7.4 | 8.5 | 8.1 | 9.1 |
| mContriever-X | 12.4 | 9.8 | 15.7 | 6.7 | 14.0 | 11.7 | 13.3 | 15.5 | 13.9 | 13.6 | 13.9 | 16.9 | 6.5 | 12.0 | 13.8 | 7.5 | 13.4 | 9.8 | 12.4 | 13.0 | 14.1 |
| mContriever-X$^{\heartsuit}$ | 13.5 | 11.6 | 15.4 | 8.0 | 16.9 | 12.3 | 15.2 | 16.7 | 15.7 | 14.7 | 15.6 | 17.4 | 7.0 | 14.2 | 14.7 | 9.1 | 13.2 | 10.1 | 14.8 | 12.1 | 14.9 |
| *Synthetic Baselines (Our work)* | | | | | | | | | | | | | | | | | | | | | |
| SWIM-X (120K)$^{MT}$ | 26.1 | 25.2 | 29.5 | 2.1 | 30.8 | 22.1 | 31.5 | 35.8 | 31.5 | 28.7 | 32.2 | 34.6 | 2.2 | 32.7 | 27.7 | 14.8 | 30.7 | 21.0 | 28.2 | 30.6 | 29.2 |
| SWIM-X (120K) | 25.2 | 24.4 | 27.7 | 4.3 | 28.3 | 25.4 | 29.4 | 32.4 | 28.8 | 30.1 | 31.8 | 34.4 | 5.1 | 30.7 | 25.7 | 15.8 | 29.6 | 20.6 | 26.1 | 27.9 | 26.1 |

Table 6: Experimental results for cross-lingual retrieval on XTREME-UP test (Ruder et al., 2023). ($^{\heartsuit}$) denotes the mContriever-X model fine-tuned without MS MARCO (Nguyen et al., 2016); Two variants of SWIM-X considered, both fine-tuned on 120K synthetic data: (1) SWIM-X (120K)$^{MT}$ fine-tuned using Google Translate, i.e., translated prompt exemplars for 15 languages, whereas (2) SWIM-X (120K) is fine-tuned using prompt exemplars sampled from XTREME-UP training split for all languages.

ingly, none of the evaluated baselines perform well on two extremely low-resource languages, Boro (brx) and Manipuri (mni).

### 3.5 Effectiveness of Summarization in SAP

In our work, we utilize SAP, where we employ extractive summarization as a rationale for PaLM 2 to generate informative multilingual queries. To evaluate the effectiveness of summarization, we assess both models (i.e., contrasting with and without summarization) on cross-lingual retrieval using Recall@5kt on XOR-Retrieve. We additionally evaluate different PaLM 2 model sizes to observe a correlation between retrieval model performance and changes in LLM size, i.e., model parameters. To ensure consistency, we adopt the experimental setup utilized in SWIM-X (500K) for all models.

Our results are shown in Figure 4 (left). we infer two insights: (i) an increase in the LLM size provides diminishing returns in terms of Recall@5kt on XOR-Retrieve. (ii) SAP outperforms standard prompting by at least 0.6 points consistently with all various PaLM-2 generators on XOR-Retrieve, with a maximum improvement of up to 3.2 points Recall@5kt. We observe that PaLM 2 with large sizes (sizes $>$ S) are inherently able to generate coherent queries, leading to diminishing improvements in SAP versus standard prompting.

### 3.6 How much Synthetic Data to Generate?

We analyze the optimal amount of synthetic training data required for fine-tuning SWIM-X. Figure 6 depicts the relative improvement in Recall@5kt on XOR-Retrieve. SWIM-X performance (gradually increasing) starts to saturate after 500K synthetic training pairs. The first observation is that with only 2K training pairs, SWIM-X (2K) achieves 49.1 Recall@5kt on XOR-Retrieve, already outperforming the best zero-shot (English-only) baseline. The break-even point occurs at 200K pairs, where SWIM-X (250K) achieves 60.5, outperforming mContriever-X, which achieves a 59.6 Recall@5kt on XOR-Retrieve.

### 3.7 Indo-European Language Transferability

We investigate the language transfer capabilities of the available Indic split (Indo-European language family) in SWIM-IR. We fine-tune individual SWIM-X models for eight selected languages and evaluate them on XTREME-UP. From Figure 5, we observe that SWIM-X models fine-tuned for Konkani (gom) or Hindi (hi) transfers best on all languages in XTREME-UP (rows 3 and 4), whereas fine-tuning for Tamil (ta) transfers worst overall (row 8). Assamese (as), Konkani (gom), Odia (or), Pashto (pa) and Sanskrit (sa) exhibit the lowest zero-shot capabilities with SWIM-X, thereby highlighting the importance of in-language synthetic data. Hindi (hi), Kannada (kn) and Malayalam (ml) demonstrate good zero-shot transfer capabilities with all Indic languages.

## 4 Ablation Studies

**Optimal value of k-shot exemplars.** We investigate the optimal value of few-shot exemplars required by PaLM 2 and the variation in the retrieval performance on XOR-Retrieve.[12] From Figure 4 (right), we observe a linear improvement in Recall@5kt with increase in K. Best Recall@5kt is observed with K = 5. SAP is unable to perform well zero-shot (i.e., K = 0) due to the complex nature of the multilingual query generation task which requires few-shot exemplars to understand and generate a summary and a query.

[12] We limit K = 5 to fit within a context length of 4096 tokens. For additional exemplars, PaLM 2 would need a longer context length increasing the computational cost significantly.

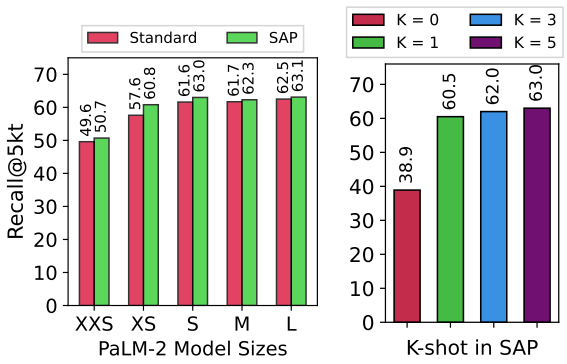


Figure 4: (Left) SAP (*Summarize-then-Ask Prompting*) (green) versus standard prompting (red) for various PaLM 2 model sizes. (Right) Varying K-shot prompt exemplars. SWIM-X is fine-tuned on 500K SWIM-IR training pairs and evaluated on XOR-Retrieve.

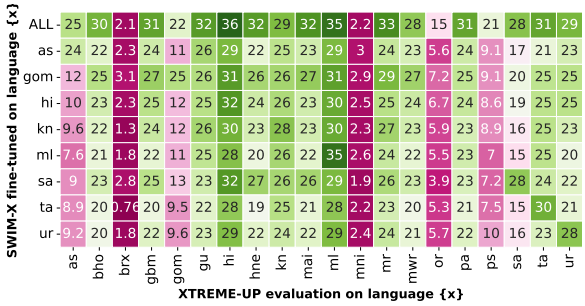


Figure 5: Heatmap showing MRR@10 denoting language-based transfer ability of SWIM-X (120K) across Indo-European languages available in XTREME-UP (Ruder et al., 2023). (ALL) denotes SWIM-X fine-tuned on all XTREME-UP languages.

**ByT5 tokenizer.** We evaluate whether the poor performance of SWIM-X on low-resource languages in XTREME-UP can be attributed to low-quality language tokenization. We replicate SWIM-X using a ByT5-base model as backbone, which contains a language independent tokenizer extension (Xue et al., 2022). From our results in Table 7, ByT5 models underperform by up to 9.8 points MRR@10 on XTREME-UP, in contrast to mT5-base. Additionally, the performance of SWIM-X on both `mni` and `brx` does not improve with ByT5. We leave it as future work to investigate the low performance on `mni` and `brx`.

**Training split query replacement.** Next, we evaluate the impact of human-generated versus LLM-generated queries on retrieval performance on XTREME-UP. We replace all human-generated queries in the XTREME-UP training split with only synthetic queries generated using PaLM 2 (S). From Table 7, the performance drops by 2.0 points at MRR@10. This confirms that human-generated queries are of better quality, which correlates with an improvement in MRR@10 on XTREME-UP. However, SWIM-X can be fine-tuned efficiently using few synthetic training pairs, by only marginally dropping in retrieval performance.

## 5 Cost Comparison

Generating synthetic training data is relatively inexpensive; however, it is not free. The cost is dependent upon the length of the prompt, input, and output generated from the LLM. The costs also linearly increase with each additional language pair. At the time of writing, PaLM 2 and similar LLMs cost about 0.0005 USD for 1000 characters in the input and output text.[13] Our prompts on average contain about 8–9K characters in the prompt input and generate about 1–2K characters in the output. The relative performance improvement associated with annotation cost in XOR-Retrieve is shown in Figure 6. Generating 200K synthetic training pairs in SWIM-IR will roughly cost $1K USD. SWIM-X (200K) performs comparably to the best supervised baseline (mContriever-X), trained on 15.2K human-annotated pairs, requiring roughly 14 times more, i.e., $14.1K USD to annotate, if we pay an hourly rate of $18.50 USD per hour for the annotator (local minimum wages is $11.50 USD/hr) following (Zhang et al., 2023b), assuming an estimated annotation cost of 3.0 minutes per example (Ruder et al., 2023).

## 6 Background and Related Work

The development of pre-trained multilingual LMs has contributed toward recent progress in multilingual retrieval (Asai et al., 2021a; Izacard et al., 2022; Asai et al., 2021b; Li et al., 2022; Ruder et al., 2023; Zhang et al., 2023b,a). Notable baselines in this field include mDPR and mContriever. mDPR (Asai et al., 2021a,b; Zhang et al., 2023a) extends English DPR (Karpukhin et al., 2020) to the multilingual setting, while mContriever (Izacard et al., 2022) adopts an unsupervised pre-training objective using the contrastive loss function and data prepared from mC4 (Xue et al., 2021), and is fine-tuned on MS MARCO.

**Synthetic data generation.** Traditionally, docT5query (Nogueira and Lin, 2019) for query generation has been prominent for generating synthetic training data in English (Ma et al., 2021;

[13] PaLM 2 pricing: cloud.google.com/vertex-ai/pricing

| Model | PLM | Query Gen. | brx | mni | MRR@10 |
|---|---|---|---|---|---|
| *1. Models with Byte-level (UTF-8) tokenizer* | | | | | |
| mContriever-X$^{\heartsuit}$ | ByT5 | Human | 1.8 | 1.0 | 2.1 |
| SWIM-X (120K)$^{MT}$ | ByT5 | PaLM 2 | 2.1 | 4.9 | 13.3 |
| SWIM-X (120K) | ByT5 | PaLM 2 | 5.1 | 5.8 | 15.4 |
| *2. Human-generated query replacement in XTREME-UP* | | | | | |
| mContriever-X$^{\heartsuit}$ | mT5 | Human | - | - | 13.5 |
| SWIM-X (≈10K) | mT5 | PaLM 2 | - | - | 11.5 |

Table 7: XTREME-UP ablation studies. First, we replace mT5 pre-trained model with ByT5 (Xue et al., 2022). Next, we replace the human-generated queries in the training dataset with PaLM-2 synthetic queries; MRR@10 scores are macro-averaged for all 20 languages; `brx` denotes Boro and `mni` denotes Manipuri.

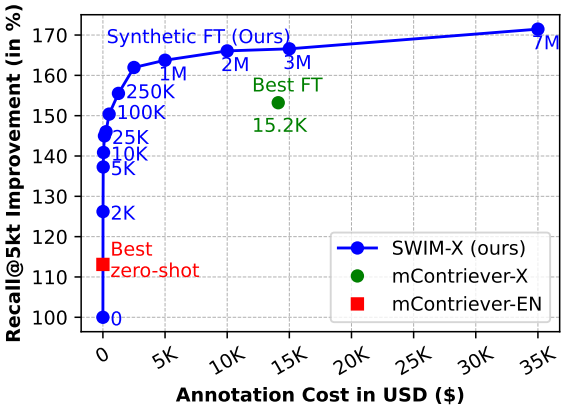


Figure 6: Recall@5kt improvement (in %) on XOR-Retrieve versus annotation cost in USD ($) to construct the training dataset. The amount of generated training pairs (human-generated marked in red and green; LLM-generated marked in blue) is provided with each marked data point in the graph.

Thakur et al., 2021; Wang et al., 2022; Thakur et al., 2022). Recently, using LLMs for query generation has gained interest. Bonifacio et al. (2022) proposed InPars, where they few-shot prompt GPT-3 (Brown et al., 2020) to generate synthetic queries. Similarly, complementary works (Sachan et al., 2022; Jeronymo et al., 2023; Boytsov et al., 2024; Saad-Falcon et al., 2023; Dua et al., 2023) all follow a similar setup as in Bonifacio et al. (2022). Dai et al. (2023) proposed Promptagator, which studied task-dependent few-shot LLM prompting and used the synthetic data for both retrieval and ranking models. Similarly, HyDE (Gao et al., 2023) and GenRead (Yu et al., 2023) generate synthetic documents instead of queries. However, prior work has focused on English, with the exception of HyDE. In our work, we robustly investigate how LLMs can be used for improving multilingual retrieval systems.

**Multilingual datasets.** Prior work investigates techniques to build multilingual datasets for better fine-tuning or evaluation of dense retrieval models. Datasets such as NeuCLIR (Lawrie et al., 2023), MKQA (Longpre et al., 2021) have been constructed using human annotators. Similarly, mMARCO (Bonifacio et al., 2021) has been generated using machine translation of MS MARCO (Nguyen et al., 2016). However, as translated documents are not written by native speakers, mMARCO and similar datasets suffer from artifacts such as "Translationese" (Clark et al., 2020). A concurrent work, JH-POLO (Mayfield et al., 2023), prompts GPT-3 to generate English queries from language specific passages in NeuCLIR.

## 7 Discussion and Future Work

A large-scale construction of SWIM-IR is challenging. Conducting SAP-based LLM generation at a large scale would require an efficient solution. Currently, we support a total of 33 languages. Extending naively to 100 languages would lead to at least 3 times the cost (fixed cost with every language). Hence, naively increasing more languages is not feasible. Instead, in the future, we can focus on generating synthetic data for diverse languages present within groups or clusters, based on linguistic characteristics within a language family or subfamily (Rijhwani et al., 2019) and rely on cross-lingual transfer for the remaining languages.

## 8 Conclusion

In this work, we present SWIM-IR, a synthetic multilingual retrieval training dataset with 28 million training pairs across 33 diverse languages. SWIM-IR allows synthetic fine-tuning of multilingual dense retrieval models cheaply without human supervision. SWIM-IR is constructed using SAP, which stands for *summarize-then-ask prompting*, assisting the LLM to identify the relevant sections of the input passage, improving the quality of the generated multilingual query.

Our rigorous evaluation across three multilingual retrieval benchmarks assesses our dataset quality. We find that SWIM-X, fine-tuned on SWIM-IR (keeping model training parameters unchanged) outperforms the best supervised cross-lingual baseline by 7.1 points Recall@5kt on XOR-Retrieve and 11.7 points MRR@10 on XTREME-UP, while remaining competitive in monolingual retrieval on MIRACL.

## 9 Limitations of SWIM-IR

SWIM-IR, like any other dataset, is not perfect and has limitations. These limitations do not directly affect the downstream multilingual retrieval task, where dense retrieval models learn how to match relevant passages to queries. The dataset has been created for the "sole" purpose of training multilingual retrieval models. We describe below a few noted limitations:

**1. Decontextualization.** PaLM 2 captures the salient information from the paragraph, but can generate the query in a reduced context, which cannot be answered without the Wikipedia paragraph.

**2. Code-switching.** PaLM 2 can occasionally generate a code-switched query with words combined from English and the target language. Code-switching is more frequently observed for cross-lingual generation in low-resource languages.

**3. Passage quality and length.** A good quality passage contains relevant information about a topic, which PaLM 2 uses to generate a synthetic query. However, if the passage is really short with little or no information, or contains noisy information, this can likely generate a subpar query.

**4. Factual inconsistencies in LLM generation.** LLMs have been found to generate text lacking sufficient grounding to knowledge sources (Dziri et al., 2022; Ji et al., 2023), thereby posing risks of misinformation and hallucination in their generated outputs (Maynez et al., 2020; Raunak et al., 2021; Muller et al., 2023). Queries in SWIM-IR are relevant for the input passage, but are not human-verified, thereby queries may contain factual inconsistencies. We leave it as future work to investigate techniques to improve factual consistency of generated queries (Sun et al., 2021; Huang et al., 2023).

## Acknowledgements

We would like to thank Jinhyuk Lee and other internal reviewers from Google for reviewing our paper and giving feedback on the draft.

## References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. PaLM 2 Technical Report. *CoRR*, abs/2305.10403.

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 547–564. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 7547–7560.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2387–2392. ACM.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2021. mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. *CoRR*, abs/2108.13897.

Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2024. InPars-Light: Cost-Effective Unsupervised Training of Efficient Rankers. *Transactions on Machine Learning Research*. Reproducibility Certification.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Emma Strubell, Sameer Singh, and Pat Verga. 2023. To Adapt or to Annotate: Challenges and Interventions for Domain Adaptation in Open-Domain Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14429–14446. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Mikel L. Forcada. 2002. Explaining real MT to translators: between compositional semantics and word-for-word. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, Manchester, England. European Association for Machine Translation.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1762–1777. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2Query-: When Less is More. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, volume 13981 of *Lecture Notes in Computer Science*, pages 414–422. Springer.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *SIGIR '21: The*

*44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 113–122. ACM.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023. Zero-shot Faithful Factual Error Correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5660–5676, Toronto, Canada. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research*.

Vitor Jeronymo, Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, Roberto de Alencar Lotufo, Jakub Zavrel, and Rodrigo Frassetto Nogueira. 2023. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval. *CoRR*, abs/2301.01820.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-Hermelo, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Evaluating Embedding APIs for Information Retrieval. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 518–526. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 3948, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Dawn J. Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the TREC 2022 NeuCLIR Track. *CoRR*, abs/2304.12367.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022. Learning Cross-Lingual IR from an English Retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States. Association for Computational Linguistics.

Jimmy Lin, Rodrigo Frassetto Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for

Multilingual Open Domain Question Answering. *Trans. Assoc. Comput. Linguistics*, 9:1389–1406.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

James Mayfield, Eugene Yang, Dawn J. Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. 2023. Synthetic Cross-language Information Retrieval Training Data. *CoRR*, abs/2305.00331.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and Modeling Attribution for Cross-Lingual Question Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and Code Embeddings by Contrastive Pre-Training. *CoRR*, abs/2201.10005.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR*, abs/1611.09268.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9844–9855. Association for Computational Linguistics.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 17761826, New York, NY, USA. Association for Computing Machinery.

Jiazuo Qiu and Deyi Xiong. 2019. Generating Highly Relevant Questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5983–5987, Hong Kong, China. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2825–2835. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5919–5930. Association for Computational Linguistics.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Ifeoluwa Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, R. Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages. *CoRR*, abs/2305.11938.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. 2023. UDAPDR: Unsupervised Domain Adaptation via LLM Prompting and Distillation of Rerankers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11265–11279, Singapore. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving Passage Retrieval with Zero-Shot Question Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3781–3797. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. 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. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Si Sun, Yingzhuo Qian, Zhenghao Liu, Chenyan Xiong, Kaitao Zhang, Jie Bao, Zhiyuan Liu, and Paul Bennett. 2021. Few-Shot Text Ranking with Meta Adapted Synthetic Weak Supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5030–5043, Online. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual Neural Machine Translation with Language Clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, and Jimmy Lin. 2022. Injecting Domain Adaptation with Learning-to-hash for Effective and Efficient Zero-shot Dense Retrieval. *CoRR*, abs/2205.11498.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Bingning Wang, Ting Yao, Weipeng Chen, Jingfang Xu, and Xiaochuan Wang. 2021. Multi-Lingual Question Generation with Language Agnostic Language Model. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2262–2272. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

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 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association.

John Wieting, Jonathan Clark, William Cohen, Graham Neubig, and Taylor Berg-Kirkpatrick. 2023. Beyond Contrastive Learning: A Variational Generative Model for Multilingual Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12044–12066, Toronto, Canada. Association for Computational Linguistics.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than Retrieve: Large Language Models are Strong Context Generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023a. Toward Best Practices for Training Multilingual Dense Retrieval Models. *ACM Trans. Inf. Syst.*, 42(2).

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *CoRR*, abs/2304.04675.

Shengyao Zhuang, Linjun Shou, and Guido Zuccon. 2023. Augmenting Passage Representations with Query Generation for Enhanced Cross-Lingual Dense Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1827–1832. ACM.

# A Appendix

The following supplementary sections in SWIM-IR are arranged as follows:

- Appendix B provides information on the SWIM-IR dataset release.
- Appendix C provides the additional material with SWIM-IR, including the data card, examples, and prompts.
- Appendix D provides details on SWIM-IR content filtration.
- Appendix E provides information in detail on hyperparameter tuning and training methodology for baseline models, including multilingual pre-training, synthetic fine-tuning, and passage sampling strategies.
- Appendix F provides statistics for three multilingual retrieval evaluation datasets: XOR-Retrieve, MIRACL, and XTREME-UP.
- Appendix G contains additional experimental results on XOR-Retrieve and MIRACL.

# B Details on SWIM-IR Dataset Release

**Dataset release format.** The SWIM-IR dataset will be released and available in multiple formats. Officially, the dataset is released within the Google Cloud Storage (GCS) cloud storage bucket.[14] Later, for longer term preservation, the dataset will be maintained through a TensorFlow Dataset (TFDS). To enable a wider audience within the research community, we plan to release an official copy of SWIM-IR as a Hugging Face dataset (Lhoest et al., 2021).

**High quality check.** The SWIM-IR dataset has undergone a high-quality check and a thorough review internally at Google to avoid inaccurate or misleading conclusions drawn from the dataset. High-quality checks are integral to the scientific process to enable researchers to address errors, inconsistencies and identify potential sources of bias within datasets (Pushkarna et al., 2022). This enables a robust and trustworthy scientific analysis within the community.

**Long term preservation.** SWIM-IR will be available for a long time by continually updating the Tensorflow dataset (TFDS) and Hugging Face dataset. The authors will be responsible for maintaining the dataset and extending the work in the future to support more languages (Joshi et al., 2020). Another useful feature is (EN→L) cross-language retrieval setting, i.e., English query retrieves language-specific passages within a corpus.

**Licensing.** The SWIM-IR corpora is based on multilingual Wikipedia. Therefore for licensing SWIM-IR, we follow the same license as Wikipedia: Creative Commons Attribution-ShareAlike 4.0 Unported License (CC BY-SA 4.0).[15] The license allows both researchers and industry alike to access the SWIM-IR dataset, copy, and redistribute it for future work.

# C SWIM-IR Extra Material

## C.1 SWIM-IR Data Card

We release the data card associated with the SWIM-IR. The data card was generated using the template provided by the Data Cards Playbook (Pushkarna et al., 2022). It has been formatted using Markdown.[16] The SWIM-IR data card is provided along with our dataset release on the GitHub repository: `https://github.com/google-research-datasets/SWIM-IR`.

## C.2 SWIM-IR Dataset Statistics

The languages covered and the amount of training pairs available in SWIM-IR are provided in Table 8. The majority of the training pairs (sampled for a maximum of 1M per language pair) are provided for 18 languages in MIRACL, which overlap with the 7 languages in XOR-Retrieve. An additional 100K training pairs come from the rest of the 15 Indo-European languages from XTREME-UP. Two examples from SWIM-IR for each task, cross-lingual and monolingual retrieval, are provided in Figure 8. The cross-lingual example is from Chinese (`zh`) and the monolingual is from Spanish (`es`).

Each SWIM-IR training data point has six associated text fields. We describe each field below: (i) `_id`: denotes the unique identifier of the training pair. (ii) `title`: denotes the title of the Wikipedia article. (iii) `text`: denotes the passage extracted from the Wikipedia article. (iv) `query`: denotes the synthetic multilingual query generated using PaLM 2 (Anil et al., 2023). (v) `lang`: denotes the target language in which the query was generated. (vi) `code`: denotes the ISO code of the generated query language.

[14] storage.googleapis.com/gresearch/swim-ir/swim_ir_v1.tar.gz

[15] https://creativecommons.org/licenses/by-sa/4.0

[16] The Markdown format and the template are available here: https://github.com/pair-code/datacardsplaybook

### C.3 SWIM-IR Prompts

All prompts and their templates (across all 33 languages) used to develop SWIM-IR are available in the GitHub repository.[17] We provide a few individual prompt examples for all three datasets in the Appendix: (1) XOR-Retrieve (English passage; synthetic Bengali query) in Figure 9, (2) MIRACL (Chinese passage; synthetic Chinese query) in Figure 10, and (3) XTREME-UP (English passage; synthetic Hindi query) in Figure 11.

## D Content Filtration

LLMs have been shown to generate undesirable content, particularly when primed with material aimed at eliciting negative patterns or associations from the model's training data (Gehman et al., 2020; Bender et al., 2021). Initially, we expected that the sampled Wikipedia passages would predominantly contain safe material suitable for prompting LLMs. However, after examination, we discovered that between 6–10% of the pairs contained sensitive subjects and adult content (i.e., weapons; violence and abuse; accidents and disasters; death and tragedy; war and conflict). To address this issue, we used the Google Cloud Natural Language content classification categories[18] to identify and remove pairs where either the original sampled passage or the resulting LLM generated query has a content classification of either `/Adult` or any of the `/Sensitive Subjects` labels.

## E Additional Technical Details

### E.1 mContriever Pre-training

In the original implementation of mContriever (Izacard et al., 2022), the authors initialized the model using the mBERT (Devlin et al., 2019) pre-trained language model (PLM). Subsequently, the model was jointly pre-trained on 29 languages covering the CCNet dataset (Wenzek et al., 2020) with a contrastive pre-training objective.

In our adaptation of mContriever, we initialize using the mT5-base model checkpoint (Xue et al., 2021). Next, we jointly pre-train the model on 101 languages[19] available in mC4 dataset (Xue et al., 2021). For each mC4 document, we sample two random non-overlapping texts with a maximum text span size of 256 tokens. Similar to the mT5 pre-training objective (Xue et al., 2021), examples were not uniformly sampled over languages; instead, the probability of selecting a training sample from a specific language is directly proportional to the amount of training data available in the mC4 dataset. We randomly sample a maximum of 20K samples per language and use them as a validation subset.

We optimize our mContriever model with the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e^{-3}$, batch size of 8192, and for 600K pre-training steps. During the first 500K pre-training steps, we use a language-mixed training objective, where a single training batch can contain examples across multiple languages. For the subsequent 100K training steps, we use a language-unmixed training objective, where a single training batch contains all examples from only a single language, i.e., no mixing of different language pairs within a training batch. We internally conducted a brief evaluation of the mContriever pre-trainining strategies using language-mixing (500K) and with both language-mixing and unmixing (600K) checkpoints. Notably on XOR-Retrieve, we observed a significant performance improvement with the additional language-unmixed pre-training, resulting in an improvement of 7.3 points Recall@5kt.

### E.2 Supervised Baselines

**XOR-Retrieve.** For the zero-shot baseline model, we fine-tune on the English-only MS MARCO (Nguyen et al., 2016) dataset using our base initialization model, mT5 (Xue et al., 2021). We use in-batch negatives, AdamW optimizer (Loshchilov and Hutter, 2019) and with a learning rate of $1e^{-3}$. The query sequence length is set to a maximum sequence length of 64 tokens, whereas the document is limited to a maximum sequence length of 256 tokens. On MS MARCO, models are fine-tuned with a batch size of 4096 and for an additional 50K training steps.

For our supervised baselines, we fine-tune on the XOR-Retrieve training dataset containing 15,250 training pairs. Each training pair in XOR-Retrieve is accompanied by one hard negative (Asai et al., 2021a). We fine-tune our baseline models on XOR-Retrieve using triplets containing the query, relevant passage and a single hard negative. We use the AdamW optimizer (Loshchilov and Hutter, 2019), a learning rate of $1e^{-3}$, a batch size of 4096 and fine-tune the model for 15K train-

[17] https://github.com/google-research-datasets/SWIM-IR
[18] cloud.google.com/natural-language/docs/categories
[19] The list of all 101 languages in mC4 can be found at: www.tensorflow.org/datasets/catalog/c4

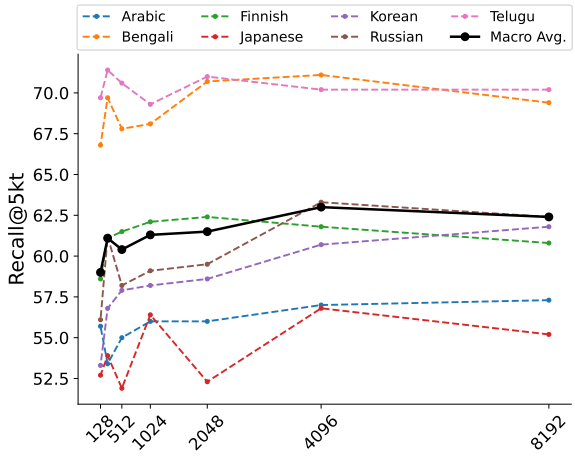


Figure 7: Training batch size ablation of SWIM-X (500K) on XOR-Retrieve (Asai et al., 2021a). The best Recall@5kt is achieved with 4096 training batch size. To avoid overfitting, we fine-tune all SWIM-X variants on 500K SWIM-IR training pairs with decreasing training steps of {40K, 40K, 30K, 30K, 20K, 20K, 15K} for increasing batch sizes of {128, 256, 512, 1024, 2048, 4096, 8192} respectively.

ing steps.

**MIRACL.** For the zero-shot baseline model, we first fine-tune on the MS MARCO (Nguyen et al., 2016) dataset. We use the same fine-tuning setup as described for XOR-Retrieve. For monolingual supervised models, we use the MIRACL training data. MIRACL authors provides between one to nine hard negatives for each training query. We randomly sample up to a maximum of four hard negatives for each query and use the AdamW optimizer (Loshchilov and Hutter, 2019), learning rate of $1e^{-3}$, a batch size of 4096 and fine-tune the model for 15K training steps.

**XTREME-UP.** For the zero-shot baseline model, we fine-tune on the MS MARCO (Nguyen et al., 2016) dataset. For the supervised baselines, we use the XTREME-UP training data containing 13,270 training pairs and fine-tune with only in-batch negatives (i.e., no hard negatives). We use the AdamW optimizer (Loshchilov and Hutter, 2019), a learning rate of $1e^{-3}$, a batch size of 1024, and fine-tune the model for 5K training steps.

### E.3 Synthetic Baselines

We fine-tune all SWIM-X models using in-batch negatives (no hard negatives), AdamW optimizer (Loshchilov and Hutter, 2019) and with a learning rate of $1e^{-3}$. The pre-trained language model for SWIM-X is the mT5-base model with 580M parameters (Xue et al., 2021). The batch size and the training steps varies for each dataset. An ablation for batch size is provided in Figure 7. Training data is evenly distributed across all languages present. For example, if there are 100K pairs with 5 different languages, each language contains around 20K training pairs.

**XOR-Retrieve.** SWIM-X is fine-tuned with a batch size of 4096 and with a maximum of 50K training steps on SWIM-IR cross-lingual pairs. For the 500K training pairs, we fine-tune for 20K steps, and for the maximum of 7M pairs, we fine-tune for 50K training steps. The training pairs within a single batch include language-mixing, i.e., one or more language-specific training pairs are sampled within a single training batch.

**MIRACL.** SWIM-X is fine-tuned for a batch-size of 4096 and for a maximum of 15K training steps on SWIM-IR monolingual pairs. As shown in (Roy et al., 2020; Zhang et al., 2023a), language-unmixed training setup is shown to work well for monolingual retrieval. Following prior work, SWIM-X training pairs include language unmixing, i.e., all pairs are from a single language. The examples are uniformly sampled across all languages, i.e., probability that a training sample comes from a specific language is equal for all languages, unlike during mC4 pre-training.

**XTREME-UP.** SWIM-X is fine-tuned for a batch size of 1024 and for a maximum of 15K training steps on SWIM-IR cross-lingual (Indic) pairs. Similar to XOR-Retrieve, the training pairs include language-mixing within a single batch during SWIM-X fine-tuning.

### E.4 Stratified Sampling in SWIM-IR

In our work, we use a stratified sampling technique to select a subset of passages from the Wikipedia corpus, aiming for a relatively uniform distribution of training samples across all languages. Our Wikipedia corpus contains entities which are sorted alphabetically (A-Z). We then compute inclusion threshold $I_{th}$, which is defined as $I_{th} = D_{sample}/D_{total}$, where $(D_{sample})$ is number of passages required to sample and $(D_{total})$ is the total numbers of passages in corpus. Next, for each passage $(p_i)$ in the corpus, we randomly generate an inclusion probability $\hat{p}_i \in [0, 1]$. We select the passage $(p_i)$ if $p_i \leq I_{th}$. This approach ensures a uniform sampling of passages with Wikipedia en-

tities between all letters (A-Z).[20]

## F Evaluation Dataset Information

We evaluate on three multilingual retrieval benchmarks: (i) **XOR-Retrieve** (Asai et al., 2021a), (ii) **MIRACL** (Zhang et al., 2023b) and (iii) **XTREME-UP** (Ruder et al., 2023). We excluded NeuCLIR (Lawrie et al., 2023) from our evaluation as it contained a fewer subset of languages namely, Chinese (zh), Farsi (fa) and Russian (ru). Although MKQA (Longpre et al., 2021) contained a wider variety of languages, it is primarily used for question-answering (QA) rather than multilingual retrieval. All three selected evaluation datasets contain a training split. Only XTREME-UP has released their test split publicly, which we use for evaluation in the paper. However, for both XOR-Retrieve and MIRACL, we evaluate on the development split.

**XOR-Retrieve** (Asai et al., 2021a) is a cross-lingual open retrieval training and evaluation task within TyDi-QA (Clark et al., 2020). XOR-Retrieve contains 15K human annotated relevant passage-query pairs in the training set with one hard negative and 2K passage-answer pairs in the *dev* set. The corpus $C$ contains 18.2M passages with a maximum of 100 word tokens from the English Wikipedia. The queries are multilingual and cover seven languages. We evaluate our models using recall at m kilo-tokens, i.e., Recall@mkt, which computes the fraction of queries for which the minimal answer is contained within the top $m$ thousand tokens of the retrieved passages. Following prior work in Asai et al. (2021a), we evaluate our models at Recall@5kt and Recall@2kt.

**MIRACL** (Zhang et al., 2023b) is a monolingual open retrieval evaluation task containing 18 languages. MIRACL was developed on top of Mr. TyDi (Zhang et al., 2021), and covers more languages and provides denser judgments by human annotators. The test set is not publicly released, hence in this paper we evaluate using the dev set. The training set contains 88,288 pairs, with the exception of Yoruba (yo) and German (de) which do not have any training data available. The authors also provide labeled hard negatives for the training query-passage pairs. The dev set contains around 13,495 query-passage pairs. The corpus $C$ in MIRACL are language-specific Wikipedia articles with various sizes starting from smallest, Yoruba (yo) with 49K passages, till the largest, English (en) with 39.2M passages. Following prior work in Zhang et al. (2023b) and Kamalloo et al. (2023), we evaluate our models at nDCG@10 and Recall@100.

**XTREME-UP** Ruder et al. (2023) contains diverse information-access and user-centric tasks focused on under-represented languages. In our work, we evaluate a cross-lingual retrieval task containing 5,280 query-passage pairs in the training set. The corpus $C$ contains 112,426 passages sampled from TyDi-QA (Clark et al., 2020). The test set contains 10,705 query-passage pairs for evaluation. The cross-language retrieval for the question-answering (QA) task contains 20 under-represented Indic languages. Following prior work in Ruder et al. (2023), we evaluate our models at MRR@10.

## G Additional Results

**XOR-Retrieve.** In Table 9, we report the Recall@2kt scores across all multilingual retrievers on XOR-Retrieve. We find similar trends for improvement, SWIM-X (7M) outperforms the best supervised model, mContriever-X, by 3.9 points at Recall@2kt. The SWIM-X (7M) without mC4 pre-training is a strong baseline outperforming SWIM-X (7M) with mC4 pre-training on 4 out of the 7 languages evaluated in XOR-Retrieve.

**MIRACL.** In Table 10, we report the Recall@100 scores across all multilingual retrievers on MIRACL. mContriever-X achieves the highest Recall@100 score of 86.5, SWIM-X on the other hand achieves 78.9 at Recall@100, which is competitive and outperforms both the zero-shot baselines, i.e., mDPR-EN and mContriever-EN. For Yoruba, Our SWIM-X outperforms the supervised mContriever-X which shows the importance of synthetic training data for low-resource languages, as the MIRACL supervised training dataset does not contain training pairs in Yoruba (i.e., no human-labeled training pairs).

[20] All Wikipedia entities starting with a non-alphabet are included in the beginning of the Wikipedia corpus.

| Cross-Lingual (18) | | Monolingual (18) | | Cross-Lingual (15) | |
|---|---|---|---|---|---|
| Q-P Lang. | # Train Pairs | Q-P Lang. | # Train Pairs | Q-P Lang. | # Train Pairs |
| MIRACL (Zhang et al., 2023b) | | | | XTREME-UP (Ruder et al., 2023) | |
| ar-en | 901,363 | ar-ar | 890,389 | as-en | 5,899 |
| bn-en | 909,748 | bn-bn | 257,327 | bho-en | 5,763 |
| de-en | 909,145 | de-de | 943,546 | gom-en | 5,755 |
| en-en | - | en-en | 936,481 | gu-en | 5,870 |
| es-en | 905,771 | es-es | 947,340 | kn-en | 5,763 |
| fa-en | 910,295 | fa-fa | 973,409 | mai-en | 5,768 |
| fi-en | 906,429 | fi-fi | 967,139 | ml-en | 5,907 |
| fr-en | 911,694 | fr-fr | 977,900 | mni-en | 5,604 |
| hi-en | 919,729 | hi-hi | 466,272 | mr-en | 5,977 |
| id-en | 907,826 | id-id | 837,459 | or-en | 5,837 |
| ja-en | 906,862 | ja-ja | 893,520 | pa-en | 5,840 |
| ko-en | 905,669 | ko-ko | 941,459 | ps-en | 5,694 |
| ru-en | 904,933 | ru-ru | 915,693 | sa-en | 5,779 |
| sw-en | 905,242 | sw-sw | 123,099 | ta-en | 5,930 |
| te-en | 902,190 | te-te | 220,431 | ur-en | 5,816 |
| th-en | 914,610 | th-th | 451,540 | | |
| yo-en | 902,467 | yo-yo | 43,211 | | |
| zh-en | 921,701 | zh-zh | 946,757 | | |
| Overall Training Pairs = 28,265,848 | | | | | |

Table 8: Dataset Statistics of SWIM-IR across both cross-lingual and monolingual settings; (Q-P Lang.) denotes the language code of the query-passage training pair in SWIM-IR; (# Train Pairs) denotes the count of the relevant training pairs containing the synthetic query and original passage pair.

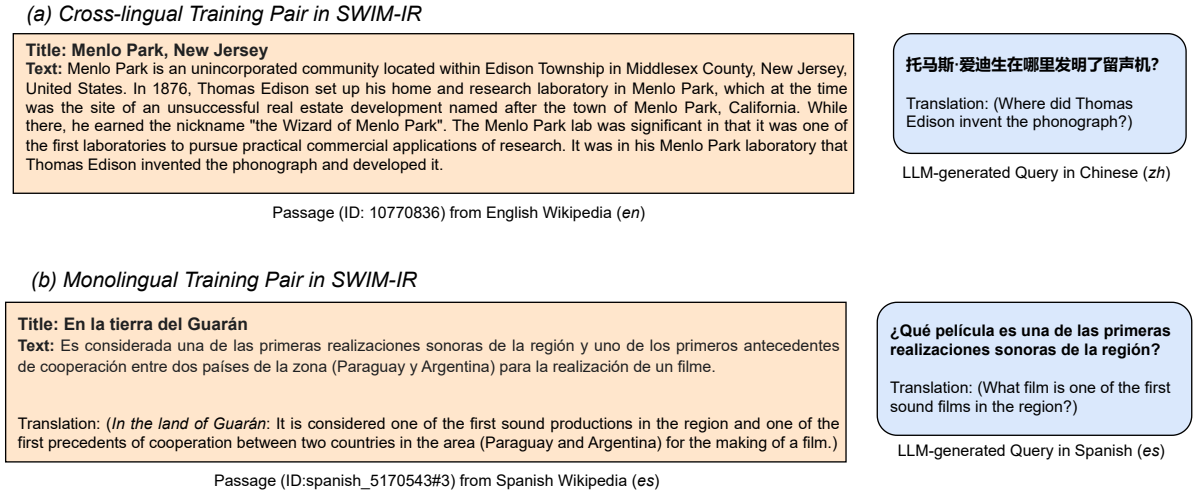


Figure 8: Dataset examples showing both (a) cross-lingual and (b) monolingual training pairs in the SWIM-IR dataset. The passage is selected from English Wikipedia, and PaLM 2 generates the query. A detailed description of all the dataset column headers are provided in Appendix (§C.2). All translations in the figure above have been provided using Google Translate (translate.google.com) for illustration purposes.

| Model | PLM | PT | Finetune (Datasets) | Recall@2kt | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Avg. | Ar | Bn | Fi | Ja | Ko | Ru | Te |
| *Existing Supervised Baselines (Prior work)* | | | | | | | | | | | |
| Dr. DECR (Li et al., 2022) | XLM-R | WikiM | NQ + XOR* | 66.0 | – | – | – | – | – | – | – |
| mDPR (Asai et al., 2021a) | mBERT | — | XOR | 40.5 | 38.8 | 48.4 | 52.5 | 26.6 | 44.2 | 33.3 | 39.9 |
| mBERT + xQG (Zhuang et al., 2023) | mBERT | — | XOR | 46.2 | 42.4 | 54.9 | 54.1 | 33.6 | 52.3 | 33.8 | 52.5 |
| Google MT + DPR (Asai et al., 2021a) | BERT | — | NQ | 62.2 | 62.5 | 74.7 | 57.3 | 55.6 | 60.0 | 52.7 | 72.3 |
| OPUS MT + DPR (Asai et al., 2021a) | BERT | — | NQ | 42.7 | 43.4 | 53.9 | 55.1 | 40.2 | 50.5 | 30.8 | 20.2 |
| *Zero-shot baselines (English-only supervision)* | | | | | | | | | | | |
| mContriever | mT5 | mC4 | — | 29.9 | 27.2 | 23.0 | 35.0 | 27.0 | 27.7 | 35.0 | 34.0 |
| mDPR-EN | mT5 | — | MS MARCO | 30.6 | 26.2 | 26.0 | 37.9 | 32.8 | 24.6 | 34.6 | 32.4 |
| mContriever-EN | mT5 | mC4 | MS MARCO | 33.8 | 27.8 | 24.3 | 42.4 | 29.9 | 31.2 | 40.5 | 40.3 |
| *Supervised Baselines (Cross-lingual supervision)* | | | | | | | | | | | |
| mDPR-X | mT5 | — | XOR | 43.6 | 43.7 | 50.0 | 44.6 | 36.1 | 41.1 | 35.9 | 54.2 |
| mContriever-X | mT5 | mC4 | XOR | 46.6 | 40.1 | 62.5 | 47.1 | 38.2 | 44.2 | 38.4 | 55.5 |
| mDPR-X | mT5 | — | MS MARCO + XOR | 49.5 | 46.0 | 63.8 | 49.0 | 39.0 | 48.4 | 43.9 | 56.3 |
| mContriever-X | mT5 | mC4 | MS MARCO + XOR | 53.0 | 47.6 | 65.1 | 51.6 | 47.3 | 50.2 | 44.3 | 65.1 |
| *Synthetic Baselines (Our work)* | | | | | | | | | | | |
| SWIM-X (500K) | mT5 | — | SWIM-IR | 49.2 | 46.3 | 57.2 | 49.0 | 42.7 | 45.6 | 44.7 | 58.8 |
| SWIM-X (500K) | mT5 | mC4 | SWIM-IR | 53.3 | 46.6 | 61.8 | 51.9 | 46.5 | 49.1 | 55.3 | 61.8 |
| SWIM-X (7M) | mT5 | — | SWIM-IR | 56.6 | 50.8 | 65.1 | 56.1 | 48.1 | 54.0 | 55.7 | 66.4 |
| SWIM-X (7M) | mT5 | mC4 | SWIM-IR | 56.9 | 53.4 | 67.8 | 55.1 | 49.4 | 52.6 | 55.3 | 64.7 |

Table 9: Experimental results showing Recall@2kt for cross-lingual retrieval on XOR-Retrieve dev (Asai et al., 2021a); (PLM) denotes the pre-trained language model; (PT) denotes the pre-training dataset; (*) Dr.DECR is fine-tuned in a complex training setup across more datasets (§3.3); WikiM denotes WikiMatrix (Schwenk et al., 2021); XOR denotes XOR-Retrieve; SWIM-X (ours) is fine-tuned on 500K and 7M synthetic data.

| Model | Avg. | ar | bn | en | es | fa | fi | fr | hi | id | ja | ko | ru | sw | te | th | zh | de | yo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Existing Supervised Baselines (Prior work)* | | | | | | | | | | | | | | | | | | | |
| BM25 | 77.2 | 88.9 | 90.9 | 81.9 | 70.2 | 73.1 | 89.1 | 65.3 | 86.8 | 90.4 | 80.5 | 78.3 | 66.1 | 70.1 | 83.1 | 88.7 | 56.0 | 57.2 | 73.3 |
| mDPR | 79.0 | 84.1 | 81.9 | 76.8 | 86.4 | 89.8 | 78.8 | 91.5 | 77.6 | 57.3 | 82.5 | 73.7 | 79.7 | 61.6 | 76.2 | 67.8 | 94.4 | 89.8 | 79.5 |
| Hybrid | 88.0 | 94.1 | 93.2 | 88.2 | 94.8 | 93.7 | 89.5 | 96.5 | 91.2 | 76.8 | 90.4 | 90.0 | 87.4 | 72.5 | 85.7 | 82.3 | 95.9 | 88.9 | 80.7 |
| Cohere-API | 76.9 | 85.4 | 85.6 | 74.6 | 71.7 | 77.1 | 80.9 | 81.6 | 72.4 | 68.3 | 81.6 | 77.1 | 76.7 | 66.6 | 89.8 | 86.9 | 76.9 | 72.5 | 57.6 |
| *Zero-shot baselines (English-only supervision)* | | | | | | | | | | | | | | | | | | | |
| mDPR-EN | 76.9 | 85.5 | 85.9 | 72.4 | 66.8 | 79.7 | 86.0 | 71.4 | 74.2 | 67.0 | 80.1 | 77.1 | 77.4 | 80.2 | 91.9 | 84.8 | 68.5 | 70.9 | 58.6 |
| mContriever-EN | 76.6 | 73.5 | 80.8 | 52.1 | 49.5 | 61.7 | 66.0 | 51.8 | 50.3 | 63.5 | 65.6 | 56.3 | 58.9 | 73.5 | 85.9 | 76.6 | 58.2 | 36.3 | 30.2 |
| *Supervised Baselines (Monolingual supervision)* | | | | | | | | | | | | | | | | | | | |
| mDPR-X | 60.6 | 73.5 | 80.8 | 52.1 | 49.5 | 61.7 | 66.0 | 51.8 | 50.3 | 63.5 | 65.6 | 56.3 | 58.9 | 73.5 | 85.9 | 76.6 | 58.2 | 36.3 | 30.2 |
| mContriever-X | 86.5 | 92.0 | 95.3 | 80.6 | 78.8 | 84.0 | 93.1 | 86.0 | 82.1 | 83.7 | 89.5 | 87.7 | 86.7 | 93.3 | 96.7 | 94.3 | 85.9 | 79.3 | 68.8 |
| *Synthetic Baselines (Our work)* | | | | | | | | | | | | | | | | | | | |
| SWIM-X (180K) | 78.9 | 89.2 | 87.8 | 72.9 | 70.0 | 76.3 | 91.6 | 75.8 | 72.5 | 74.3 | 77.6 | 76.8 | 77.9 | 87.8 | 84.9 | 92.9 | 69.9 | 72.4 | 69.3 |

Table 10: Experimental results for monolingual retrieval on MIRACL dev (Zhang et al., 2023b). All scores denote Recall@100; Hybrid denotes a hybrid retriever with ranked fusion of three retrievers: mDPR, mColBERT and BM25; BM25, mDPR and Hybrid scores (Zhang et al., 2023b); Cohere-API is used as a reranker on top of 100 BM25 results (Kamalloo et al., 2023). SWIM-X is fine-tuned on 180K synthetic data.

**5-shot Summarize-then-Ask Prompting for XOR-Retrieve**

Read the following article and write a factual summary. Your summary will act as a surrogate for asking a question based on the article. Finally, translate the question to Bengali.

Article: Long Lost Family is a BAFTA award winning British television series that has aired on ITV since 21 April 2011. The programme, which is presented by Davina McCall and Nicky Campbell, aims to reunite close relatives after years of separation. It is made by the production company Wall to Wall. "Long Lost Family" is based on the Dutch series "Spoorloos" (), airing on NPO 1 since February 1990 and it is made by KRO-NCRV. Presented by Davina McCall and Nicky Campbell, the series offers a last chance for people who are desperate to find long lost relatives.
Summary: Long Lost Family is a BAFTA award winning British television series aired since 2011. The series aim to reunite close relatives after years of seperation which is presented by Davina McCall and Nicky Campbell.
Question [Bengali]: ব্রিটিশ টেলিভিশন সিরিজ লং লস্ট ফ্যামিলি কোন পুরস্কার জিতেছে?

Article: Muscular activity accounts for much of the body's energy consumption. All muscle cells produce adenosine triphosphate (ATP) molecules which are used to power the movement of the myosin heads. Muscles have a short-term store of energy in the form of creatine phosphate which is generated from ATP and can regenerate ATP when needed with creatine kinase. Muscles also keep a storage form of glucose in the form of glycogen. Glycogen can be rapidly converted to glucose when energy is required for sustained, powerful contractions. Within the voluntary skeletal muscles, the glucose molecule can be metabolized anaerobically in a process.
Summary: All muscle cells produce adenosine triphosphate (ATP) molecules for movement of myosin heads. A short term store of energy is generated from ATP in the form of cratine phosphate and can regenerate ATP when needed with creatine kinase.
Question [Bengali]: কীভাবে পেশী কোষগুলি মায়োসিন মাথার নড়াচড়ার জন্য শক্তিকে শক্তি দেয়?

Article: The 1960s brought anime to television and in America. The first anime film to be broadcast was "Three Tales" in 1960. The following year saw the premiere of Japan's first animated television series, "Instant History", although it did not consist entirely of animation. Osamu Tezuka's "Tetsuwan Atom" ("Astro Boy") is often miscredited as the first anime television series, premiering on January 1, 1963. "Astro Boy" was highly influential to other anime in the 1960s, and was followed by a large number of anime about robots or space.
Summary: First anime movie broadcast on TV was 'Three Tales' in 1960. First anime TV series was 'Instant History' in 1961. 'Astro Boy' first aired in 1963 was a highly influential anime about robots or space.
Question [Bengali]: ১৯৬০ সালে টিভিতে সম্প্রচারিত প্রথম অ্যানিমে ছবি কোনটি?

Article: Łęczna is a town in eastern Poland with 19,780 inhabitants (2014), situated in Lublin Voivodeship. It is the seat of Łęczna County and the smaller administrative district of Gmina Łęczna. The town is located in northeastern corner of historic province of Lesser Poland. Łęczna tops among the hills of the Lublin Upland, at the confluence of two rivers —the Wieprz, and the Świnka. On December 31, 2010, the population of the town was 20,706. Łęczna does not have a rail station, the town has been placed on a national Route 82 from Lublin to Włodawa. And shall be considered as a
Summary: Łęczna is a town in eastern Poland with 19,780 inhabitants. It is a hill located in the Lublin Upland, at the confluence of two rivers - Wieprz and Świnka. It is a road hub, and has no rail station.
Question [Bengali]: লিচেনা পোল্যান্ডের কোন দুটি নদীর সঙ্গমস্থলে অবস্থিত?

Article: The μ-law algorithm (sometimes written "mu-law", often approximated as "u-law") is a companding algorithm, primarily used in 8-bit PCM digital telecommunication systems in North America and Japan. It is one of two versions of the G.711 standard from ITU-T, the other version being the similar A-law, used in regions where digital telecommunication signals are carried on E-1 circuits, e.g. Europe. Companding algorithms reduce the dynamic range of an audio signal. In analog systems, this can increase the signal-to-noise ratio (SNR) achieved during transmission; in the digital domain, it can reduce the quantization error (hence increasing signal to quantization noise ratio).
Summary: The μ-law algorithm is a companding algorithm, which is used to reduce the dynamic range of audio signals. In analog systems, this can increase the signal-to-noise ratio (SNR) achieved during transmission.
Question [Bengali]: μ-আইন অ্যালগরিদম কীভাবে অ্যানালগ সিস্টেমে সংক্রমণকে প্রভাবিত করে?

Article: **{Input Wikipedia Article in English}**
Summary:

Figure 9: 5-shot SAP (*Summarize-then-Ask Prompting*) for XOR-Retrieve (Asai et al., 2021a) is shown for Bengali (bn). There are five exemplars (5-shot) in our cross-lingual query generation task. The passages are randomly selected from the XOR-Retrieve Wikipedia corpus. A summary and a query for all above exemplars is manually written in English by the authors. Finally, the English written query is translated to Bengali (bn) for all above exemplars using Google Translate (translate.google.com).

**3-shot Summarize-then-Ask Prompting for MIRACL**

Read the following article in Chinese and write a factual summary in Chinese. Your summary will act as a surrogate for asking a question in Chinese based on the article.

Article: 四川各地小吃通常也被看作是川菜的组成部分。由于重庆地区小吃相对较少，除重庆麻辣小面外，川菜小吃主要以成都小吃为主。 主要有担担面、川北凉粉、麻辣小面、酸辣麵、 酸辣粉、叶儿粑、酸辣豆花、三合泥、红油抄手等以及用创始人姓氏命名的赖汤圆、龙抄手、钟水饺、吴抄手等。 甜品方面，以原产四川眉山的冰粉和四川宜宾长宁县的凉糕最有名。
Summary: 四川美食种类繁多，小吃也非常有名，主要有担担面、川北凉粉、麻辣小面、酸辣粉、叶儿粑、酸辣豆花、三合泥、红油抄手、赖汤圆、龙抄手、钟水饺、吴抄手等。甜品方面，以原产四川眉山的冰粉和四川宜宾长宁县的凉糕最有名。
Question [Chinese]: 四川美食有哪些?

Article: 獅子座流星雨 (Leonids[ˈli.əˌnɪdz] \"lee-uhnids\")是與周期大約33年的坦普爾·塔特爾彗星有關的一個流星雨。獅子座流星雨的得名是因為這個流星雨輻射點的位置在獅子座。在2009年，這個流星雨的尖峰時間在11月17日（世界時），每小時的數量可能高達500顆，尚不足以成為流星暴（每小時超過1,000顆流星的大流星雨）。
Summary: 上一次狮子座流星雨发生在2009年11月17日。狮子座流星雨是与周期大约33年的坦普尔·塔特尔彗星有关的一个流星雨。狮子座流星雨的得名是因为这个流星雨辐射点的位置在狮子座。
Question [Chinese]: 上一次狮子座流星雨发生在什么时间?

Article: 清华大学（，縮寫：），简称清华，舊称清华学堂、游美肄业馆、清华学校、國立清華大學，是一所位于中华人民共和国北京市海淀区清华园的公立大学。始建于1911年，因北京西北郊清华园而得名。初为清政府利用美国退还的部分庚子赔款所建留美预备学校“遊美学务处”及附设“肄业馆”，於1925年始设大学部。抗日战争爆发后，清华与北大、南开南迁长沙，组建国立长沙临时大学。1938年再迁昆明，易名国立西南联合大学。1946年迁回清华园复校，拥有文、法、理、工、农等5个学院。1949年中华人民共和国成立后，国立清华大学归属中央人民政府教育部，更名“清华大学”；而原国立清华大学校长梅貽琦于1955年在台湾新竹复校，仍沿用原名。
Summary: 清华大学始建于1911年，因北京西北郊清华园而得名。初为清政府利用美国退还的部分庚子赔款所建留美预备学校“遊美学务处”及附设“肄业馆”。
Question [Chinese]: 清华大学什么时候成立的?

Article: **{Input Wikipedia Article in Chinese}**
Summary:

Figure 10: 3-shot SAP (*Summarize-then-Ask Prompting*) for MIRACL (Zhang et al., 2023b) is shown for Chinese (zh). There are three exemplars (3-shot) in our monolingual query generation task. The query-passage pairs are randomly selected from MIRACL training set. Finally, the summary for all above exemplars is automatically generated in Chinese (zh) using Google Bard (bard.google.com).

**5-shot Summarize-then-Ask Prompting for XTREME-UP**

Read the following article and write a factual summary. Your summary will act as a surrogate for asking a question based on the article. Finally, translate the question to Hindi.

Article: Long Lost Family is a BAFTA award winning British television series that has aired on ITV since 21 April 2011. The programme, which is presented by Davina McCall and Nicky Campbell, aims to reunite close relatives after years of separation. It is made by the production company Wall to Wall. "Long Lost Family" is based on the Dutch series "Spoorloos" (), airing on NPO 1 since February 1990 and it is made by KRO-NCRV. Presented by Davina McCall and Nicky Campbell, the series offers a last chance for people who are desperate to find long lost relatives.
Summary: Long Lost Family is a BAFTA award winning British television series aired since 2011. The series aim to reunite close relatives after years of seperation which is presented by Davina McCall and Nicky Campbell.
Question [Hindi]: ब्रिटिश टेलीविजन लॉन्ग लॉस्ट फैमिली ने कौन सा पुरस्कार जीता?

Article: Muscular activity accounts for much of the body's energy consumption. All muscle cells produce adenosine triphosphate (ATP) molecules which are used to power the movement of the myosin heads. Muscles have a short-term store of energy in the form of creatine phosphate which is generated from ATP and can regenerate ATP when needed with creatine kinase. Muscles also keep a storage form of glucose in the form of glycogen. Glycogen can be rapidly converted to glucose when energy is required for sustained, powerful contractions. Within the voluntary skeletal muscles, the glucose molecule can be metabolized anaerobically in a process.
Summary: All muscle cells produce adenosine triphosphate (ATP) molecules for movement of myosin heads. A short term store of energy is generated from ATP in the form of cratine phosphate and can regenerate ATP when needed with creatine kinase.
Question [Hindi]: मायोसिन हेड्स की गति के लिए मांसपेशियों की कोशिकाएं ऊर्जा को कैसे शक्ति देती हैं?

Article: The 1960s brought anime to television and in America. The first anime film to be broadcast was "Three Tales" in 1960. The following year saw the premiere of Japan's first animated television series, "Instant History", although it did not consist entirely of animation. Osamu Tezuka's "Tetsuwan Atom" ("Astro Boy") is often miscredited as the first anime television series, premiering on January 1, 1963. "Astro Boy" was highly influential to other anime in the 1960s, and was followed by a large number of anime about robots or space.
Summary: First anime movie broadcast on TV was 'Three Tales' in 1960. First anime TV series was 'Instant History' in 1961. 'Astro Boy' first aired in 1963 was a highly influential anime about robots or space.
Question [Hindi]: १९६० में टीवी पर प्रसारित होने वाली पहली एनीमे फिल्म कौन सी थी?

Article: Łęczna is a town in eastern Poland with 19,780 inhabitants (2014), situated in Lublin Voivodeship. It is the seat of Łęczna County and the smaller administrative district of Gmina Łęczna. The town is located in northeastern corner of historic province of Lesser Poland. Łęczna tops among the hills of the Lublin Upland, at the confluence of two rivers —the Wieprz, and the Świnka. On December 31, 2010, the population of the town was 20,706. Łęczna does not have a rail station, the town has been placed on a national Route 82 from Lublin to Włodawa. And shall be considered as a
Summary: Łęczna is a town in eastern Poland with 19,780 inhabitants. It is a hill located in the Lublin Upland, at the confluence of two rivers - Wieprz and Świnka. It is a road hub, and has no rail station.
Question [Hindi]: लेक्ज़ना पोलैंड में किन दो नदियों के संगम पर स्थित है?

Article: The μ-law algorithm (sometimes written "mu-law", often approximated as "u-law") is a companding algorithm, primarily used in 8-bit PCM digital telecommunication systems in North America and Japan. It is one of two versions of the G.711 standard from ITU-T, the other version being the similar A-law, used in regions where digital telecommunication signals are carried on E-1 circuits, e.g. Europe. Companding algorithms reduce the dynamic range of an audio signal. In analog systems, this can increase the signal-to-noise ratio (SNR) achieved during transmission; in the digital domain, it can reduce the quantization error (hence increasing signal to quantization noise ratio).
Summary: The μ-law algorithm is a companding algorithm, which is used to reduce the dynamic range of audio signals. In analog systems, this can increase the signal-to-noise ratio (SNR) achieved during transmission.
Question [Hindi]: कैसे μ-नियम एल्गोरिथम एनालॉग सिस्टम में संचरण को प्रभावित करता है?

Article: **{Input Wikipedia Article in English}**
Summary:

Figure 11: 5-shot SAP (*Summarize-then-Ask Prompting* with Machine Translation (MT) for XTREME-UP (Ruder et al., 2023) is shown for Hindi (hi). There are five exemplars (5-shot) in our cross-lingual query generation. The exemplars are re-used from XOR-Retrieve. A summary and a query for all above exemplars is manually written in English by the authors. Finally, the English written query is translated to Hindi (hi) for all above exemplars using Google Translate (translate.google.com).

# Annotation Guidelines for SWIM-IR

**Nandan Thakur**

**June 2nd 2023**

- The goal of this task is to evaluate the quality of LLM-generated (PaLM 2-S) generated questions.
- Every annotator will receive a set of annotations containing the wikipedia paragraph and the question in the ${target_language}.
- Annotators should read each annotation carefully and provide feedback on the following:
    - The ***fluency*** of the question.
    - The ***adequacy*** of the question.
    - The ***language*** of the question.
- Annotators should be respectful and professional in their feedback.
- Annotators should complete all annotations within the allotted duration.

Here below we define the following terms:

## Fluency

| Rating Level | Explanation |
|---|---|
| **2** (Flawless) | Perfect use of ${target_language} with no mistakes at all. |
| **1** (Good) | Few or minor spelling or grammar mistakes; the text is still mostly understandable and readable. |
| **0** (Poor) | Many or serious spelling, grammar, or other mistakes, which make the text difficult to understand or hard to read. |

## Adequacy

| Rating Level | Explanation |
|---|---|
| **2** (Relevant) | Highly related to the wiki passage. The question can be answered using the wiki passage. |

| | |
|---|---|
| **1** (Moderate) | The question is somewhat related to the wiki paragraph, the question cannot be answered using the passage. |
| **0** (Not Relevant) | The question is not at all related to the wiki passage. |

## Language

| **Rating Level** | **Explanation** |
|---|---|
| **2** (Flawless) | The whole question is perfectly in the ${target_language}. |
| **1** (Good) | Code-switching occurs with part of the question in the ${target_language}. |
| **0** (Poor) | The whole question is not at all in ${target_language}. |

Thank you for your participation in this task!