Few-Shot Multilingual Open-Domain QA from Five Examples

Fan Jiang and **Tom Drummond** and **Trevor Cohn***
School of Computing and Information Systems
The University of Melbourne, Victoria, Australia
fan.jiang1@student.unimelb.edu.au
{tom.drummond, trevor.cohn}@unimelb.edu.au

Abstract

Recent approaches to multilingual open-domain question answering (MLODQA) have achieved promising results given abundant language-specific training data. However, the considerable annotation cost limits the application of these methods for underrepresented languages. We introduce a *few-shot learning* approach to synthesize large-scale multilingual data from large language models (LLMs). Our method begins with large-scale self-supervised pre-training using WikiData, followed by training on high-quality synthetic multilingual data generated by prompting LLMs with few-shot supervision. The final model, FsModQA, significantly outperforms existing few-shot and supervised baselines in MLODQA and cross-lingual and monolingual retrieval. We further show our method can be extended for effective zero-shot adaptation to new languages through a *cross-lingual prompting* strategy with only English-supervised data, making it a general and applicable solution for MLODQA tasks without costly large-scale annotation.

1 Introduction

Open-domain QA has demonstrated impressive performance by employing the *retrieve-then-read* (Figure 1(a)) pipeline (Chen et al., 2017), which is built upon dense retrievers (Karpukhin et al., 2020) and efficient generative readers (Izacard and Grave, 2021). However, this success has been primarily limited to English, leaving the multilingual setting under-explored. This limitation is mainly due to the difficulty and costs of creating high-quality and balanced human-supervised training data for languages other than English. Moreover, multilingual open-domain QA introduces additional challenges with retrieving evidence from multilingual corpora, requiring the underlying retrieval system to be capable of both cross-lingual and monolingual retrieval (Asai et al., 2021b).

More recently, efforts have been made to create multilingual open-domain QA benchmarks from existing multilingual machine reading comprehension tasks (e.g., Xor-TyDi QA [Asai et al., 2021a]) and by translating English datasets (e.g., MKQA [Longpre et al., 2021]). These datasets have enabled various approaches to address multilingual open-domain QA problems, including iterative data augmentation (Asai et al., 2021b) and extensive additional pre-training on Wikipedia texts (Abulkhanov et al., 2023; Jiang et al., 2024). However, these methods still heavily depend on abundant high-quality language-specific data for fine-tuning, making them less effective solutions when language resources are limited. Therefore, a more generalizable approach to multilingual open-domain QA should aim to mitigate this reliance and be capable of facilitating language adaptation with minimally supervised samples.

In this paper, we present **FsModQA**, a method for ***Few-Shot*** **M**ultilingual **O**pen-**D**omain **QA** using minimally-sized supervised data (i.e., up to 5 per language).[1] Our approach consists of two core components: a self-supervised pre-training objective on multilingual corpora; and a synthetic data generation pipeline that prompts a large language model (LLM) using few-shot supervised examples. Concretely, we generate question-answer pairs from WikiData triples by leveraging LLMs' In-Context Learning (ICL) ability. To facilitate ICL prompts, we incorporate ChatBots to generate curated input-output pairs, which serve as examples for prompting LLMs to generate millions of questions from WikiData triples across various languages. After generating these question-answer pairs, we identify the supported

* Also at Google.

[1] We use the term *few-shot* throughout this paper to denote that our method relies on only a small number of human-annotated examples. Thus, we classify our method as a *few-shot learning* approach, consistent with Dai et al. (2023).

Transactions of the Association for Computational Linguistics, vol. 13, pp. 481–504, 2025. https://doi.org/10.1162/tacl_a_00750
Action Editor: Shay Cohen. Submission batch: 7/2024; Revision batch: 1/2025; Published 6/2025.

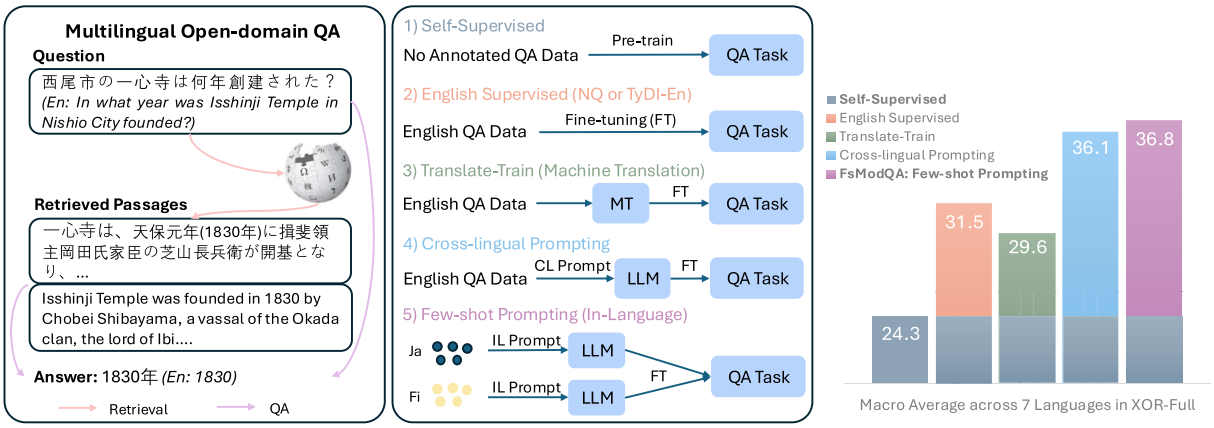

Figure 1: **Left**: Multilingual open-domain QA pipeline. **Middle**: Training strategies: 1) self-supervised pre-training; 2–4) baselines using English QA data: 2) used directly; 3) machine translated into target languages; 4) used to prompt LLMs to generate target language QA samples; 5) our method using few-shot in-language data to prompt an LLM. **Right**: Result comparison (Avg. F1) on the XOR-Full dataset.

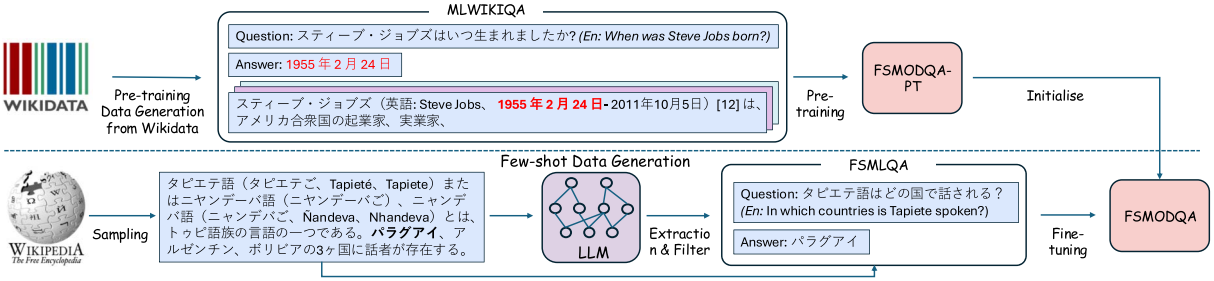

Figure 2: Full pipeline for data construction and model training: (1) generate large-scale data from Wikidata for self-supervised pre-training; (2) use few-shot prompting to generate synthetic Q&A pairs from Wikipedia passages of target languages, on which the pre-trained model is further fine-tuned.

Wikipedia passages through answer string matching. We further gather cross-lingual answers and evidence passages through Wikipedia language links to facilitate cross-lingual retrieval. Employing this generated data, we train a multilingual model with a joint objective for retrieval and QA, producing a promising pre-trained model (Figure 1(c)) for subsequent *few-shot learning*.

In *few-shot learning*, we employ LLMs for data generation from few-shot examples. For each target language, we feed the few-shot examples to an LLM and prompt it to generate question-answer pairs from a given document. The few-shot examples are assumed to encapsulate the QA style and distribution of the target dataset, enforcing the LLM to generate synthetic data with similar characteristics. With this abundant synthetic data, the pre-trained model can be further fine-tuned to achieve superior results (Figure 1(c)). As an unsupervised alternative, we explore a *zero-shot cross-lingual prompting* strategy that uses data from other languages as prompts for data generation, and we show this is almost on par with *few-shot prompting* (Figure 1(c)).

We evaluate FsModQA on various datasets, including cross-lingual and monolingual retrieval, and multilingual open-domain QA. We observe notable improvements over competitive few-shot baselines, with +5.1% gain on retrieval and +8.4% gain on multilingual open-domain QA. To further test FsModQA language adaptation ability, we conduct zero-shot adaptation experiments using our *cross-lingual prompting* strategy on 15 languages. This adaptation improves performance in both monolingual retrieval and multilingual QA significantly, achieving results that are superior or comparable to strong translation-based methods.[2]

2 FsModQA

Figure 2 presents the full pipeline for generating self-supervised pre-training and fine-tuning data.

[2]Code, data, and checkpoints are available here.

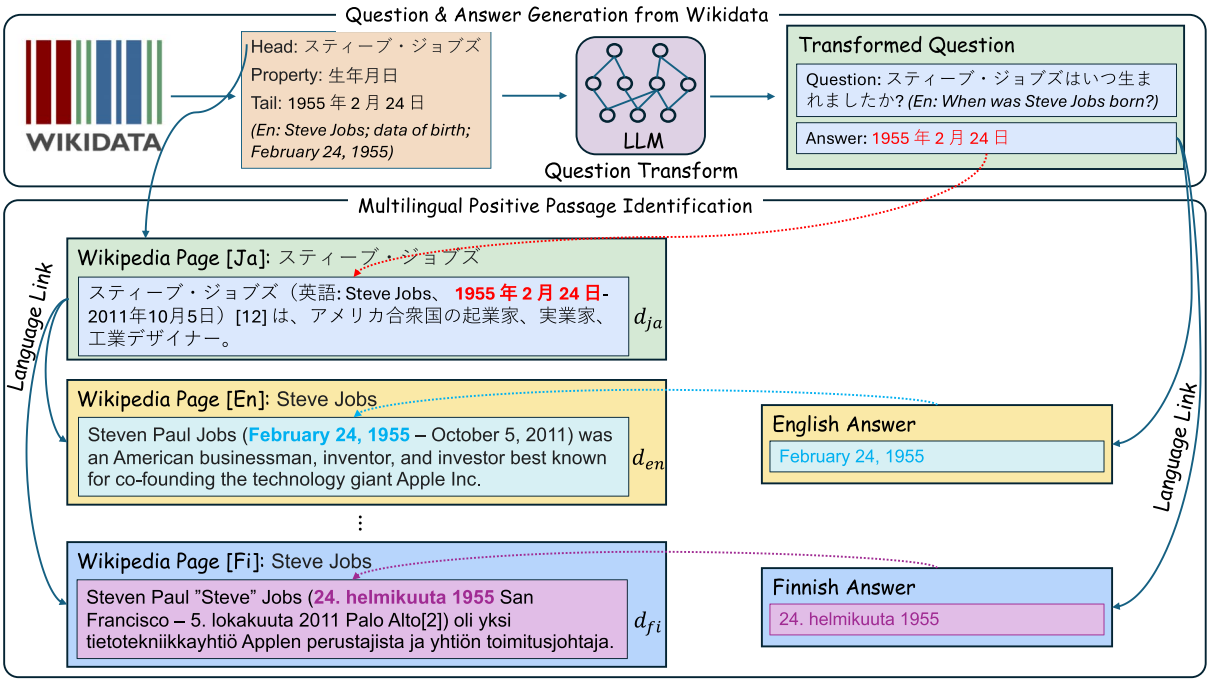

Figure 3: Pre-training data construction pipeline: (1) transform WikiData triples into QAs using LLMs for each target language L, and (2) identify in-language and cross-lingual positive passages from the head entity's Wikipedia page and through language links. *English translations* are added for readability.

2.1 Self-Supervised Data Construction

Sampling Factual Triplets. Our self-supervised training dataset is constructed based on Wikidata (Vrandečić and Krötzsch, 2014), a multilingual knowledge base consisting of fact triplets linked to millions of entities. We manually select 50 common properties (Appendix Table 10) based on English and consider all triples associated with these relations. We then gather fact triplets in the desired target languages through language links.

Generating Questions. Given a triplet $\mathcal{T} = (s, r, o)$, we aim to write a question q about the head entity s's property r with the gold answer a being the tail entity o. One can use relation-specific templates to efficiently transform each triple into natural questions (Sciavolino et al., 2021). However, this method lacks diversity, making triples with the same properties generate questions with similar surface forms. Instead, we adopt a generative approach by using a LLM to automatically generate questions with more diverse styles.

Specifically, we first sample five triples for each property and prompt ChatGPT (`gpt-3.5-turbo`) to generate three questions for each triple. This process yields a curated set of high-quality questions: $\mathbb{K} = \{s_i, r_i, o_i, q_i\}_{i=0}^{k}$.

We additionally generate questions with `Yes/No` answers from the same set of sampled triples. It is easy to generate `Yes` questions. For `No` questions, we need to create false fact triples from existing triples. Specifically, we randomly replace a triple's head or tail entity with the most similar Wikidata entity, and check the perturbed triple is not a valid fact according to Wikidata. We then generate questions using ChatGPT as before. Examples are included in Appendix Table 11.

Subsequently, these curated questions are used as ICL examples to prompt a smaller LLM to transform all sampled triples into natural questions. We use Gemma-7B (Gemma Team et al., 2024) as the LLM and include the prompts we used in Appendix Table 12.

Multilingual Positive Passage Identification. As shown in Figure 3, for a question q^{ja} and answer a^{ja} derived from a triple (s^{ja}, r, o^{ja}), we gather all passages from the Wikipedia page $\mathcal{W}^{ja}$ linked by s^{ja} and add passages containing a^{ja} as positive $\mathcal{D}_q^{ja}$. If no such passage exists, we use partial match and select the one with the highest lexical overlaps with a^{ja} as positive. We further include positive passages in other languages to facilitate cross-lingual retrieval. We first translate the triple into target languages (s^L, r, a^L) using

language links and identify cross-lingual positives by searching a^L in the Wikipedia page $\mathcal{W}^L$ linked by s^L as above. This derives monolingual and cross-lingual positive passages $\mathcal{D}_q = \mathcal{D}_q^{ja} \cup \mathcal{D}_q^L$. We generate 18.7M $(q, a, \mathcal{D}_q)$ triples across 8 languages in total, denoted as MLWIKIQA.[3]

2.2 Few-shot Synthetic Data Generation

Few-shot Setting. The main idea of FSMODQA is to amplify a limited number of annotated examples into a substantially larger volume of synthetic data by prompting LLMs. In this work, we consider XOR-TYDI QA (Asai et al., 2021a) as our target dataset. For each language in XOR-TYDI QA, we randomly sample five triples $\mathcal{K} = \{(q_i^L, a_i^L, d_i^L)\}_{i=1}^5$ from the training set as *few-shot* examples. Each triple contains the question, answer, and the ground truth passage. We ensure that three examples are span answers, while the remaining two are `Yes` and `No` answers to align with XOR-TYDI QA distribution.

Prompt-based Question & Answer Generation. We populate a hand-engineered template with our *few-shot* language-specific examples $\mathcal{K}$ and use them as the ICL examples to prompt LLM. Given a randomly sampled passage d^L from language L, we append d^L to the template, and the LLM is expected to generate a relevant question q^L and answer a^L in language L. We further constrain the answer a^L to be a span within d^L, a property of the original XOR-TYDI QA dataset.

Many questions classified as *unanswerable* in Clark et al. (2020) can be answered by referring to English Wikipedia (Asai et al., 2021a). These questions are included as *cross-lingual* questions in XOR-TYDI QA. To simulate this scenario, we generate synthetic cross-lingual data from English passages. We first use Google Translate to translate the *few-shot* examples to English: $\mathcal{K}' = \{(q_i^{En}, q_i^L, a_i^{En}, a_i^L, d_i^{En})\}_{i=1}^5$. Subsequently, we use these translated *few-shot* examples $\mathcal{K}'$ to fill another template and instruct the LLM to generate QA from a randomly sampled English passage d^{En}, first in English (q^{En}, a^{En}) and then in target language (q^L, a^L). Similarly, we restrict a^{En} to be a span within d^{En}. We include the prompts we used in Tables 13 and 14 in Appendix.

[3]We classify MLWIKIQA as a silver-standard dataset rather than a synthetic one, as it is derived from the structured information in WikiData and Wikipedia.

Data Filtering. We employ a method based on Natural Language Inference (NLI) to enhance the quality of our synthetic data. NLI techniques aim to classify whether a hypothesis text is entailed by, neutral, or contradictory to a given premise text (Bowman et al., 2015). They have been widely used for identifying factual errors in text summarization (Laban et al., 2022) and hallucinations in machine-generated texts (Honovich et al., 2022). In this study, we employ NLI methods for data filtering (Yoran et al., 2024). Given a synthetic example (q, a, d), we consider the source passage d as the premise and the concatenation of the generated question q and answer a as the hypothesis. We retain an example only when the premise entails the hypothesis.

In more detail, we apply a novel *local-to-global* filtering mechanism. In *local filtering*, we evaluate whether the originating passage d entails the synthetic QA (q, a) pairs. We take the output probability of the entailment label as the score and keep examples when the entailment score exceeds a threshold $\mathcal{T}_l$. In *global filtering*, we use a pre-trained model (i.e., the self-supervised model in Figure 1(b)) to perform retrieval for the question q and obtain a set of passages $\hat{\mathcal{D}}_q$. We compute an entailment score vector $\boldsymbol{x} \in \mathcal{R}^{|\hat{\mathcal{D}}_q|}$, with each entry being the entailment score between (q, a) and a retrieved passage $d \in \hat{\mathcal{D}}_q$. We then apply a maximum pooling operation $\max(\boldsymbol{x})$ to derive the final score. The intuition behind this is that a valid (q, a) should be supported by at least one of the retrieved passages, which aligns with open-domain settings. Similarly, we retain only those examples whose scores surpass a predefined threshold $\mathcal{T}_g$. In this way, we end up having 1.7M synthetic data in total across 7 languages, denoted as FSMLQA.

2.2.1 Zero-shot Cross-lingual Prompting

Our *few-shot* setting relies on a few annotated examples to generate synthetic QA pairs in target languages. However, this approach encounters significant challenges when the target language is extremely low-resourced, making it nearly impossible to obtain even a few examples. For this setting, we explore *zero-shot* prompting, which uses *cross-lingual* examples to prompt LLMs to generate synthetic QA pairs in target languages.

We consider two *zero-shot* prompting settings. In *English-Prompting* setting, we use English QA data to fill up a template and use it as the prompt to ask LLMs to generate QA pairs from passages

randomly sampled from the target language. In *Multilingual-Prompting* setting, we assume access to a handful of examples in a held-out language set. We randomly sample five multilingual examples from this held-out set to populate another template, and prompt LLMs to generate QA pairs in target languages. We include the prompts used in Tables 15 and 16 in Appendix.

2.2.2 Data Sampling

Our synthetic dataset, FsMLQA, exhibits a strongly skewed distribution towards shorter answer lengths (often single tokens), whereas the human-annotated answers in XOR-TYDI QA tend to be substantially longer. To address this mismatch, we resample the training data from FsMLQA according to answer length, using a geometric distribution, $l \sim \text{Geo}(p)$, to achieve a better balance between short and long answers.[4]

2.3 FsMODQA Model

Model Structure. As shown in Figure 4, we employ a single encoder-decoder model to perform both passage retrieval and QA tasks. The first half of the encoder functions as a dual-encoder with shared parameters, which separately encodes the question $\boldsymbol{q}$ and the passage corpus $\mathcal{D}$. Additionally, we append an instruction to the question to inform the language of the target answer: ''*Answer in* {lang}''. A LayerNorm operation, followed by average pooling, is applied to compress the inputs into single vectors: $E_{\boldsymbol{q}}$ and $\{E_{\boldsymbol{d}_i} | \boldsymbol{d}_i \in \mathcal{D}\}$, which are used for matching via dot products. The top-k most relevant passages to the question are selected: $\mathcal{D}_{\boldsymbol{q}} = \arg\text{topk}_{\boldsymbol{d}_i \in \mathcal{D}}(E_{\boldsymbol{q}} \cdot E_{\boldsymbol{d}_i})$. The embeddings of the question and each top-k passage in $\mathcal{D}_{\boldsymbol{q}}$ are concatenated and fed into the remaining cross-encoder layers. Finally, the cross-encoder embeddings are flattened and incorporated into the decoder through cross-attention to generate the answer $\boldsymbol{a}$, following the Fusion-in-Decoder approach (Izacard and Grave, 2021).

Model Training. FsMODQA is first pre-trained on MLWIKIQA and later fine-tuned on FsMLQA. In self-supervised pre-training, we use a simple contrastive loss and answer generation loss to train FsMODQA. The dual-encoder is updated by

[4]Empirically, we set $p = 0.4$ ($\mu = 2.5$) for all languages except for Japanese, where we set $p = 0.1$ ($\mu = 10$) to favor longer answers. When computing the distribution, we truncate the answer length to 30.

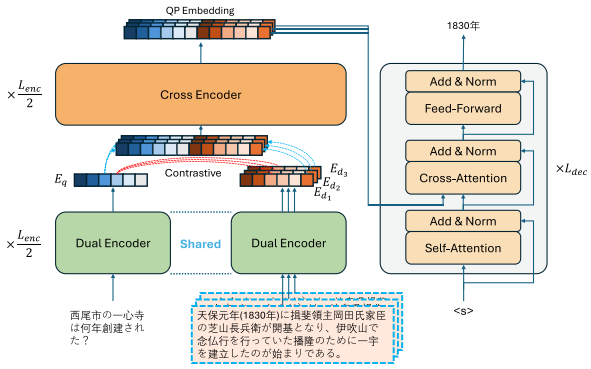

Figure 4: The unified model for passage retrieval and question answering.

contrasting the paired question passage against the targets of other questions in one training batch (i.e., in-batch negative). Formally, for i-th training example, the loss function $\mathcal{L}^i_{\text{ssl}}$ is:

$$-\log \frac{e^{(E_{\boldsymbol{q}_i} \cdot E_{\boldsymbol{d}_i})}}{\sum_{j=1}^{N} e^{(E_{\boldsymbol{q}_i} \cdot E_{\boldsymbol{d}_j})}} - \log \prod_{t=1}^{T} P(\boldsymbol{a}^i_t | \boldsymbol{a}^i_{<t}, \boldsymbol{q}_i, \boldsymbol{d}_i)$$

The pre-trained FsMODQA is subsequently fine-tuned on our synthetic data through an end-to-end training mechanism. The dual-encoder is trained using signals derived from the answer generation task, with the cross-attention score from the decoder serving as the target for assessing question-passage relevance. For i-th training example, the loss function is formally defined as:

$$\mathcal{L}^i_{\text{ret}} = \mathbb{KL}(P_{\text{ret}}(\cdot|\boldsymbol{q}_i, \mathcal{D}_{\boldsymbol{q}_i}) || P_{\text{ca}}(\cdot|\boldsymbol{q}_i, \mathcal{D}_{\boldsymbol{q}_i})),$$
$$P_{\text{ret}}(\cdot|\boldsymbol{q}_i, \mathcal{D}_{\boldsymbol{q}_i}) = \text{softmax}(E_{\boldsymbol{q}_i} \cdot E_{\boldsymbol{d}_1}, \ldots, E_{\boldsymbol{q}_i} \cdot E_{\boldsymbol{d}_{|\mathcal{D}_{\boldsymbol{q}_i}|}}),$$
$$P_{\text{ca}}(\cdot|\boldsymbol{q}_i, \mathcal{D}_{\boldsymbol{q}_i}) = \sum_{h=0}^{H} \sum_{t=0}^{|\boldsymbol{d}_j|} \frac{\text{SG}(\text{CA}(0,h,t))}{H} | \, \boldsymbol{d}_j \in \mathcal{D}_{\boldsymbol{q}_i},$$

where $\mathcal{D}_{\boldsymbol{q}_i}$ is the passages returned by the dual-encoder itself and P_{ca} is the target distribution gathered from the decoder's cross-attention scores. SG signifies stop-gradient, which prevents the decoder from being affected by the retriever loss, and CA denotes the cross-attention score at the last decoder layer. The term 0 refers to the first output token, H is the number of cross-attention heads, and $|\boldsymbol{d}_j|$ stands for the length of passage $\boldsymbol{d}_j$.

The entire model is optimized to generate the target answer $\boldsymbol{a}_i$ given $\boldsymbol{q}_i$ and relevant passages $\mathcal{D}_{\boldsymbol{q}_i}$. The final loss is: $\mathcal{L}^i_{\text{e2e}} = \mathcal{L}^i_{\text{ret}} + \mathcal{L}^i_{\text{ans}}$, where $\mathcal{L}^i_{\text{ans}} = \log \prod_{t=1}^{T} P(\boldsymbol{a}^i_t | \boldsymbol{a}^i_{<t}, \boldsymbol{q}_i, \mathcal{D}_{\boldsymbol{q}_i})$.

3 Experiments

3.1 Datasets and Metrics

We evaluate on the XOR-TyDi QA dataset (Asai et al., 2021a), with XOR-Retrieve for cross-lingual retrieval and XOR-Full for multilingual open-retrieval QA. We conduct zero-shot evaluations on two benchmarks, MIRACL (Zhang et al., 2023) for monolingual retrieval and MKQA (Longpre et al., 2021) for multilingual open-domain QA. For XOR-Retrieve, we use the February 2019 English Wikipedia dump as the retrieval corpus and the same dumps from 13 languages for XOR-Full and MKQA (Asai et al., 2021a). For MIRACL, we use the monolingual Wikipedia preprocessed by Zhang et al. (2023). Following prior work, we evaluate models at Recall@5kt (top 5000 tokens) on XOR-Retrieve; F1, exact match (EM) and BLEU on XOR-Full; nDCG@10 on MIRACL; and F1 on MKQA.

3.2 Baselines

We evaluate three ranges of representative baselines based on the type of supervised data used: (i) Zero-shot baselines (''-En'') fine-tuned on supervised English-only data (i.e., Natural Questions (Kwiatkowski et al., 2019)). (ii) Supervised baselines that fine-tuned on human-annotated multilingual data (i.e., XOR-TyDi QA). (iii) Few-shot models that improve zero-shot baselines with only a few supervised multilingual instances.

Retriever Baselines. For XOR-Retrieve, we include: (1) Zero-shot retrievers: translate-test methods: DPR+MT (Asai et al., 2021a) and ReATT+MT (Jiang et al., 2022); models pre-trained on multilingual Wikipedia: CLASS-En (Jiang et al., 2024) and LAPCA (Abulkhanov et al., 2023). (2) Supervised retrievers: multilingual dense retrievers: mDPR (Asai et al., 2021a), CORA (Asai et al., 2021b), Sentri (Sorokin et al., 2022), QuiCK (Ren et al., 2022); token-level dense retrievers: DrDecr (Li et al., 2022) pre-trains ColBERT on WikiMatrix (Schwenk et al., 2021). (3) Few-shot retrievers: SWIM-X (Thakur et al., 2024) generates massive synthetic data from LLMs through a summarisation-then-ask technique. CLASS (5-shot) fine-tunes CLASS-En on our 5-shot examples. For MIRACL (Zhang et al., 2023), we include two supervised retrievers: fine-tuned mContriever (Izacard et al., 2022) and Hybrid that combines the results of BM25, mDPR, and mColbert (Khattab and Zaharia, 2020).

Reader Baselines. (1) Zero-shot baselines: translate-test methods MT+DPR, ReAtt+MT, and GMT+GS generate answers from English retrieved passages with question and answer translations. (2) Supervised baselines: BM25 does in-language retrieval with an extractive multilingual QA model; MT+Mono first applies BM25 and then MT+DPR if no answer was generated. Fusion-in-decoder methods (i.e., CORA, CLASS, Sentri, LAPCA) use retrieval-augmented generation, generating target language answers from multilingual retrieved passages. (3) Few-shot readers: Gemma (5-shot) (Gemma Team et al., 2024) and LLaMa3 (5-shot) (Touvron et al., 2023) prompt LLMs with few-shot examples and retrieved passages using the template in Appendix Table 17; CLASS (5-shot) fine-tunes CLASS-En on few-shot examples. We use the same 5-shot examples for all methods.

3.3 Implementation Details

With the proposed self-supervised data construction method, we generate 18,735,159 triplets for pre-training across 8 languages, with statistics in Appendix Table 19. We initialize our model from the mT5-large checkpoint (Xue et al., 2021) and pre-train it using the loss function $\mathcal{L}_{\text{ssl}}$ for 100K steps with a batch size of 800 on 16 A100 GPUs for 64 hours. We set the learning rate to 5×10^{-5} with 10% steps of warm-up, and linear decay to 0.

With our few-shot data generation method, we obtain 1,746,156 question-answer pairs across 7 languages included in XOR-TyDi QA after data filtering with $\mathcal{T}_l = 0.5$ and $\mathcal{T}_g = 0.8$, with detailed statistics shown in Table 19 in Appendix. For fine-tuning, we first train the pre-trained model using NQ data for 8K steps and then on FsMLQA for 6K–14K steps depending upon the size of the sampled training dataset, with the loss function $\mathcal{L}_{\text{e2e}}$. We set the batch size to 128 and the learning rate to 5×10^{-5}. We apply an asynchronous passage update mechanism, where we periodically refresh the retrieved passages for each training query using the most recent checkpoint every 1K steps.

3.4 Retrieval Results

XOR-Retrieve. Table 1 shows that FsMODQA, fine-tuned on 100K synthetic data, surpasses the

Method	Backbone	# Total Params	Pre-training Data	Fine-tuning Data	R@5kt							
					Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg.
Zero-shot Retrievers												
DPR+MT†	mBERT	220M	–	NQ	52.4	62.8	61.8	48.1	58.6	37.8	32.4	50.6
ReAtt+MT*	T5-L	583M	–	NQ	67.3	71.0	29.3	61.8	67.0	61.2	66.4	60.6
CLASS-En*	mT5-L	410M	Wikipedia	NQ	66.7	78.6	66.6	60.2	63.2	58.2	78.2	67.4
Supervised Retrievers												
CORA	mBERT	557M	–	NQ + XOR	42.7	52.0	49.0	32.8	43.5	39.2	41.6	43.0
mDPR†	mBERT	557M	–	NQ + XOR	48.9	60.2	59.2	34.9	49.8	43.0	55.5	50.2
Sentri	XLM-R	560M	–	NQ + TQA + XOR	56.8	62.2	65.5	53.2	55.5	52.3	80.3	60.8
QuiCK	mBERT	557M	–	NQ + XOR	63.8	78.0	65.3	63.5	69.8	67.1	74.8	68.9
DrDecr	XLM-R	278M	WikiMatrix	NQ + XOR	70.2	85.9	69.4	65.1	68.8	68.8	83.2	73.1
LAPCA	XLM-R	560M	Wikipedia	NQ + XPAQ + XOR	70.2	83.8	79.6	69.7	73.6	75.5	83.1	76.5
CLASS	mT5-L	410M	Wikipedia	NQ	70.6	84.9	71.0	66.0	72.6	70.0	81.9	73.9
Few-shot Retrievers												
SWIM-X (7M)	mT5-B	580M	mC4	SWIM-IR	57.9	75.0	65.6	59.3	58.9	64.6	74.4	65.1
CLASS (5-shot)	mT5-L	410M	Wikipedia	NQ + XOR (5-shot)	**67.0**	78.6	65.6	59.0	63.6	59.0	**79.5**	67.5
FsModQA (100K)	mT5-L	410M	MLWikiQA	NQ + FsMLQA	66.3	79.3	**67.8**	**66.4**	65.6	73.8	75.2	**70.6**
FsModQA (1.7M)	mT5-L	410M	MLWikiQA	NQ + FsMLQA	63.4	**80.6**	67.5	66.0	**66.7**	**74.3**	75.6	**70.6**

Table 1: Results on XOR-Retrieve dev sets. Best performance is in bold. † and ∗ denotes results reported by Asai et al. (2021a) and Jiang et al. (2024), respectively. Others are copied from original papers.

	Seen Languages								Unseen Languages										
	ar	bn	en	fi	ja	ko	ru	te	es	fa	fr	hi	id	sw	th	zh	de	yo	Avg.
Supervised Retrievers																			
Hybrid	67.3	65.4	54.9	67.2	57.6	60.9	53.2	60.2	64.1	59.4	52.3	61.6	44.3	44.6	59.9	52.6	56.5	37.4	56.6
mContriever	66.4	68.4	44.2	65.2	56.8	58.8	51.2	79.0	42.8	48.9	46.2	45.0	45.8	67.7	70.7	49.4	42.3	48.4	55.4
Few-shot Retrievers																			
SWIM-X (180K)	60.2	57.1	34.7	40.6	40.8	43.3	49.7	55.9	33.4	36.3	**64.3**	33.0	39.5	40.0	56.3	**63.3**	**5**0.2	36.5	46.4
FsModQA (100K)	**64.4**	**63.6**	**45.4**	**64.7**	**55.1**	**49.6**	**50.0**	**76.2**	**40.5**	**43.7**	36.5	**43.2**	**42.6**	**50.2**	**60.4**	43.2	36.7	**60.2**	**51.5**

Table 2: Monolingual retrieval results on MIRACL dev sets. Best performance is in bold. Hybrid scores are taken from Zhang et al. (2023). mContriever and SWIM-X are copied from Thakur et al. (2024).

few-shot SWIM-X (7M) by 5.5% at Recall@5kt, despite the latter using substantially more synthetic data generated by a significantly larger proprietary LLM (PaLM2). This indicates our method's great efficiency in training and data generation. Further scaling up the training data to full size does not improve retrieval accuracy. In addition, we find that fine-tuning CLASS, a sophisticated pre-training method, on the same set of 5-shot examples, lags FsModQA by 3.1 points. This shows our method of amplifying data through LLM prompting is superior to direct fine-tuning.

MIRACL. Table 2 shows that FsModQA surpasses the few-shot retriever SWIM-X by 5.1%, although SWIM-X generates synthetic data on each MIRACL language through 3-shot prompting, whereas FsModQA is exclusively trained on synthetic data generated from 5-shot examples of Xor-TyDi QA and thus, evaluated on a *zero-shot* manner. We further divide languages into seen and unseen groups based on FsModQA's training data. It outperforms SWIM-X on all seen languages and 7 out of 10 unseen languages, except on zh, fr, and de. We suspect SWIM-X benefits significantly from large-scale synthetic data generation on these high-resource languages.

3.5 Multilingual QA Results

XOR-Full. In Table 3, we show FsModQA achieves the best results in few-shot settings, outperforming CLASS-En (directly fine-tuning on 5-shot examples) by 8.4% and directly few-shot promoting LLMs for QA by 18%. Compared to supervised readers, FsModQA surpasses CORA and other pipeline methods while achieving results comparable to the rest. It is also noteworthy that in two low-resource languages, FsModQA outperforms comparable supervised baselines in Bengali and achieves a closer match in Telugu, indicating the effectiveness of our method in handling low-resource languages.

Method	Backbone	# Total Params	Pre-training Data	Fine-tuning Data	F1							Macro Average		
					Ar	Bn	Fi	Ja	Ko	Ru	Te	F1	EM	BLEU
Zero-shot Readers														
MT+DPR†	mBERT	–	–	NQ	7.2	4.3	17.0	7.9	7.1	13.6	0.5	8.2	3.8	6.8
ReAtt+MT*	T5-L	1.19B	–	NQ	15.0	10.5	1.8	13.1	14.9	15.4	8.2	11.3	5.5	9.5
GMT+GS†	–	–	–	NQ	18.0	29.1	13.8	5.7	15.2	14.9	15.6	16.0	9.9	14.9
Supervised Readers														
BM25†	–	–	–	XOR	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–
MT+Mono†	mBERT	–	–	NQ + XOR	15.8	9.6	20.5	12.2	11.4	16.0	0.5	17.3	7.5	10.7
CORA	mBERT+mT5-B	1.14B	–	NQ + XOR	42.9	26.9	41.4	36.8	30.4	33.8	30.9	34.7	25.8	23.3
CLASS	mT5-L	1.23B	Wikipedia	NQ + XOR	49.1	32.0	46.7	44.1	38.4	39.9	41.1	41.6	32.5	28.2
Sentri	XLM-R+mT5-B	1.14B	–	NQ + TQA + XOR	52.5	31.2	45.5	44.9	43.1	41.2	30.7	41.3	34.9	30.7
LAPCA	XLM-R+mT5-B	1.14B	Wikipedia	NQ + XPAQ + XOR	53.4	50.2	49.3	44.7	49.5	49.3	38.9	47.8	38.7	35.5
Few-shot Readers														
Gemma (5-shot)	Gemma	7B	–	–	13.4	19.0	21.7	20.2	20.5	23.0	23.4	20.2	12.2	15.3
LLaMA3 (5-shot)	LLaMA3	8B	–	–	22.7	13.2	22.9	17.8	19.0	19.2	28.9	20.5	12.8	15.6
CLASS (5-shot)	mT5-L	1.23B	Wikipedia	NQ + XOR (5-shot)	32.3	28.1	29.9	25.7	29.5	27.7	24.7	29.8	20.5	21.2
FsModQA	mT5-L	1.23B	MLWIKIQA	NQ + FsMLQA	**41.3**	**35.4**	**39.6**	**41.5**	**35.0**	**38.2**	**36.3**	**38.2**	**27.9**	**24.4**

Table 3: Multilingual QA results on the XOR-Full dev set. Best performance is in bold. † and ∗ denotes results taken from Asai et al. (2021b) and Jiang et al. (2024). Others are copied from original papers.

Method	Da	De	Es	Fr	He	Hu	It	Km	Ms	Nl	No	Pl	Pt	Sv	Th	Tr	Vi	cn	hk	tw	Avg
Supervised Readers																					
CORA	30.4	30.2	32.0	30.8	15.8	18.4	29.0	5.8	27.8	32.1	29.2	25.6	28.4	30.9	8.5	22.2	**20.9**	5.2	6.7	5.4	21.8
CLASS	28.3	32.3	33.3	31.2	10.3	23.1	30.6	7.1	24.7	30.2	28.4	**25.6**	29.3	28.9	14.1	**24.8**	19.0	8.0	7.8	6.7	22.2
Few-shot Readers																					
FsModQA	**34.8**	**33.3**	**38.5**	**34.8**	**19.5**	**28.4**	**31.9**	**7.5**	**36.7**	**34.1**	**35.5**	18.4	**33.4**	**37.2**	**15.1**	**24.8**	9.9	**9.1**	**8.6**	**7.9**	**25.0**

Table 4: Zero-shot multilingual QA results (F1) on MKQA. Best performance is in bold. ''cn'': ''Zh-cn'' (Chinese, simplified). ''hk'': ''Zh-hk'' (Chinese, Hong Kong). ''tw'': ''Zh-tw'' (Chinese, traditional).

	XOR-Full				XOR-Retrieve
	In-LG Avg. F1	Cross-LG Avg. F1	All Avg. F1	Retrieval R^M@100	CL-Retrieval R@5kt
FsModQA	46.8	**31.2**	**36.9**	**75.0**	**70.6**
- CL Queries	**49.3**	30.0	36.8	72.0	68.4

Table 5: The effects of generating cross-lingual queries from English passages, at 100K data scale.

	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg.
FsModQA	**40.6**	34.3	**38.4**	**40.7**	32.9	**37.7**	33.9	**36.9**
- Data Filtering	39.0	31.7	37.4	39.2	32.3	35.5	**35.3**	35.8
- Geo Sampling	37.9	**35.9**	36.7	38.5	**34.1**	35.0	33.5	36.0
- MLWIKIQA	11.2	7.2	10.2	17.5	7.9	8.5	4.4	9.6

Table 6: Ablations by removing one component of our method, at 100K data scale.

MKQA. In Table 4, FsModQA achieves the best zero-shot results on MKQA in almost all languages, with an improvement of +2.8% compared to supervised CORA and CLASS. This suggests that training on our synthetic data can well generalize to other new languages, indicating that generating synthetic data for each target language may not be necessary for language adaptation.

3.6 Ablation

We perform ablation studies to justify each of our designs, with results shown in Tables 5 and 6.

Cross-lingual Data Improves Cross-lingual Ability. Excluding cross-lingual synthetic training data enhances performance in answering questions that require only the retrieval of in-language passages. However, the result on questions relying on cross-lingual passage retrieval declines, reducing the overall results. This is further evidenced by retrieval results R^M@100, where the accuracy of finding evidence in any language (e.g., English and in-language) drops, with additional support from the cross-lingual passage retrieval results.

Data Filtering Improves Data Quality. By using the raw synthetic data from LLMs without any quality control, the performance suffers in every examined language except Telugu. We suspect that the NLI model is deficient in this language.

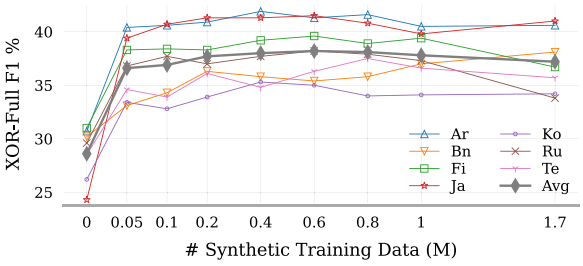

Figure 5: Performance when trained with different sizes of our synthetic data.

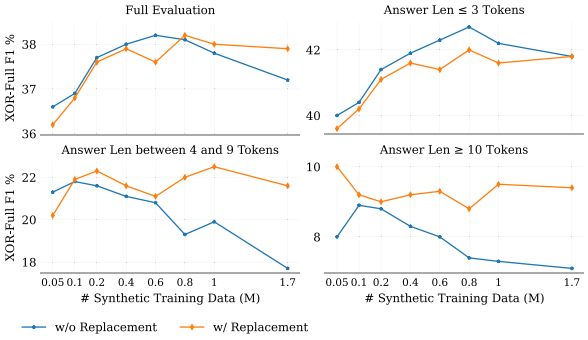

Figure 6: Performance comparison when sampling data with or without replacement by using our geometric sampling strategy.

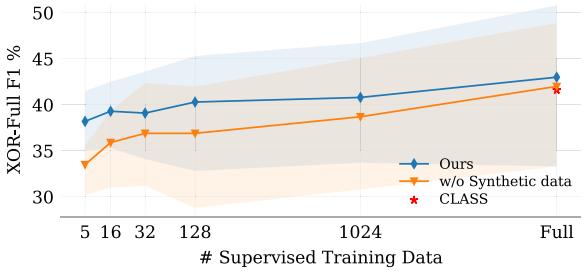

Figure 7: Results when trained with varying sizes of supervised data. The average together with the best and worst languages are reported.

Geometry Sampling Improves Long-answer Generation. Sampling data according to geometry distribution over answer length leads to a 0.9% gain on average. In languages that contain a significant number of long answers (i.e., `ar`, `fi`, `ja`, `ru`), geometry sampling shows gains of up to 2.7%. Conversely, in `bn`, and `ko`, where short answers dominate, random sampling is usually better.

Pre-training is Crucial. We observe extremely poor results in all languages without pre-training on our MLWIKIQA, primarily due to the model's low retrieval accuracy in identifying relevant passages. We believe pre-training enables the model to achieve good initial retrieval accuracy, which is essential in the subsequent fine-tuning process.

3.7 Training Data Scaling

Performance Improves with More Synthetic Data. To investigate the effect of our data scale on models, we train FSMODQA on subsets ranging from 0.05M to the entire 1.7M QA pairs, Results on each language and the average performance are shown in Figure 5. As the data size increases, FSMODQA shows enhanced average performance up to the 0.6M data scale and gradually decreases afterward. We observe that as data size increases, the proportion of examples with short answers increases (78.4% → 95.3%), and the result on long-answer examples drops from 18.0% to 15.1%, indicating overfitting to short answers.

Our geometric sampling method (§2.2.2) attempts to balance the answers by length, however its use of *sampling without replacement* means the few long answer instances are quickly exhausted, such that larger sampled datasets become skewed toward shorter answers. To mitigate this issue, we employ *sampling with replacement*. This method upsamples longer-answer examples such that the length distribution follows the precomputed geometric distribution.[5] As a result, it effectively increases the number of training epochs for data points with longer answers. As shown in Figure 6, *sampling with replacement* significantly improves performance on longer answers (≥4 tokens) while maintaining comparable performance on shorter answers relative to the current method.

Few-shot Prompting Is Superior to Direct Fine-tuning and Benefits from More Supervised Data. In Figure 7, we show that directly fine-tuning on the 5-shot examples is beneficial (28.6% → 33.5%) but remains inferior to our *few-shot* method. When increasing the size of supervised data, both methods achieve consistent improvements although the performance gap narrows. With full-sized training data, FSMODQA surpasses CLASS (43.0% *v.s.* 41.6%), achieving new state-of-the-art results. See Appendix Table 21 for results in each language.

[5] We do not cap the number of repeats.

3.8 Zero-shot Prompting Strategies

We compare our *few-shot* prompting strategy with two *zero-shot cross-lingual* prompting methods in §2.2.1. In *English-Prompting*, we consider NQ training data and TyDi QA English training data as prompting sources, respectively. In *Multilingual-Prompting*, we use 5-shot examples from all languages in Xor-TyDi QA (i.e., those used in our *few-shot* setting) for prompting. When generating synthetic data for each target language, we exclude its 5-shot examples from the prompting source. We compare the success rate of generating valid examples using different prompting strategies in Appendix Table 20, with *few-shot* prompting achieving the highest rate and *English-Prompting* with NQ yielding the lowest rate.

Zero-shot prompting is comparable to few-shot prompting. Table 7 shows that all three zero-shot prompting variants achieve consistent improvements over FsModQA-En with up to 8.1% gains, highlighting the versatility of our method in zero-shot language adaptation. Prompting with English datasets created with the same guidelines achieves better results (TyDi-En *v.s.* NQ-En), and using multilingual examples for prompting (i.e., Xor-TyDi-*) is comparable to FsModQA. Specifically, the diversity and QA styles in prompts are more important for `fi` and `te`, while for other languages, employing in-language prompts usually leads to the best performance.

English-prompting is the best way of using English data and is complementary to existing methods. We compare three different ways of using TyDi QA English data for zero-shot learning, direct English fine-tuning, fine-tuning on machine-translated data from English, and *English Prompting*. Table 8 shows the benefits of all three methods, with our *English-Prompting* approach yielding the best results in all languages. Additionally, combining data from all three methods results in improvements over any of them when used independently, and matches the performance of our few-shot setting.

4 Zero-shot Language Adaptation

In §2.2.1, we propose a *zero-shot prompting* strategy that uses few-shot examples from other languages to generate synthetic data for a distinct target language. The effectiveness of this approach is demonstrated in §3.8. In this section, we evaluate the impact of this strategy in adapting FsModQA to a diverse range of previously unseen languages, using only English labeled data.

	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg.
FsModQA-En	30.7	30.2	31.0	24.3	26.2	29.6	28.5	28.6
FsModQA	41.7	**34.7**	38.7	**39.4**	**34.7**	**35**.0	33.5	**36.8**
NQ-En	38.8	33.9	40.1	33.0	33.0	34.9	34.3	35.4
TyDi-En	39.1	34.4	**41.2**	35.6	31.9	34.6	**36.0**	36.1
Xor-TyDi-*	**42.5**	33.9	40.3	37.9	33.7	34.6	34.3	36.7

Table 7: XOR-Full performance comparison when using zero-shot prompting strategies for synthetic data generation, at 100K scale. FsModQA-En indicates the model pre-trained on MLWikiQA and fine-tuned on the English NQ dataset.

	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg.
FsModQA-En	30.7	30.2	31.0	24.3	26.2	29.6	28.5	28.6
+ Fine-tuning	36.8	30.7	35.5	29.1	28.6	30.4	29.4	31.5
+ Translate-train	31.5	31.2	29.9	26.5	28.8	27.7	31.5	29.6
+ English-prompt	39.1	34.4	41.2	35.6	31.9	34.6	36.0	36.1
+ All	**41.9**	**36.2**	**43.0**	**37.3**	**33.7**	**37.0**	**37.4**	**38.1**

Table 8: Result comparison on XOR-Full for different means of using TyDi-En data.

4.1 Experimental Setup

Languages We select ten languages unseen by FsModQA from the MIRACL dataset for monolingual retrieval adaptation. We choose ten unseen languages from the MKQA dataset with high, medium, and low resources for multilingual open-domain QA adaptation.

Data Generation We consider the English NQ training data as the source for prompts. For each target language, we randomly sample five-shot examples from the NQ dataset to prompt the generation of Q&A pairs from selected Wikipedia passages, following the procedure described in §2.2. This approach yields 128,000 training instances for each target language. Additionally, we compare this method to the translate-train baseline (MT), which uses Google Translate to translate the NQ training data into the target languages.

Model Training For both methods, we fine-tune FsModQA for 3K steps following the same procedure used in FsMLQA (§3.3). The final checkpoint

	High				Medium			Low			
MIRACL	De	Es	Fr	Zh	Fa	Hi	Id	Sw	Th	Yo	Avg.
FsModQA	36.7	40.5	36.5	43.2	43.7	42.6	43.2	50.2	60.4	60.2	45.7
+ MT	**41.3**	**41.8**	37.1	41.7	40.7	42.4	44.1	50.7	60.5	23.9	42.4
+ Adapt	38.8	41.6	**38.6**	**47.0**	**47.7**	**45.9**	**44.2**	**62.3**	**66.6**	**78.3**	**51.1**
	High				Medium				Low		
MKQA	De	Es	Fr	Zh	He	Pl	Tr	Vi	Km	Th	Avg.
FsModQA	33.3	38.5	34.8	8.5	19.5	18.4	24.9	9.9	7.5	15.1	21.0
+ MT	**42.6**	41.6	41.2	**12.8**	**32.1**	29.5	39.9	39.5	**13.8**	22.1	31.5
+ Adapt	42.1	**42.1**	**43.0**	12.0	27.4	**39.7**	**40.4**	**40.4**	13.3	**22.7**	**32.3**

Table 9: Zero-shot adaptation to unseen languages in monolingual retrieval (nDCG@10) and multilingual open-domain QA (F1).

obtained at the last training step is used for evaluation. Note that separate models are created per language in this experiment.

4.2 Results

Monolingual Retrieval Adaptation. As shown in the upper part of Table 9, the zero-shot adaptation significantly improves FsModQA's monolingual retrieval results by an average of 5.4% across ten unseen languages. These improvements are particularly pronounced in low-resource languages (i.e., `th`, `yo`, `sw`), whereas the MT baseline results in notable declines both in these languages (e.g., −36.3% in `yo`) and overall (−3.3%). Note that MIRACL was created by native speakers from texts in the target languages, which aligns with our data generation process. This explains the consistent gains achieved by our method and shows its superiority to translation-based approaches.

Multilingual Open-domain QA Adaptation. As shown in the bottom of Table 9, the adaptation effectively enhances multilingual open-domain QA performance across seven languages, achieving an average improvement of 11.3%. MT-based approaches yield results comparable to our adaptation, which is expected since MKQA was translated from NQ and the machined-translated data share the same topic distributions (i.e., American-centric). In contrast, our method generates data from Wikipedia texts written in target languages to simulate how native speakers ask questions, which is more common for real-world scenarios.

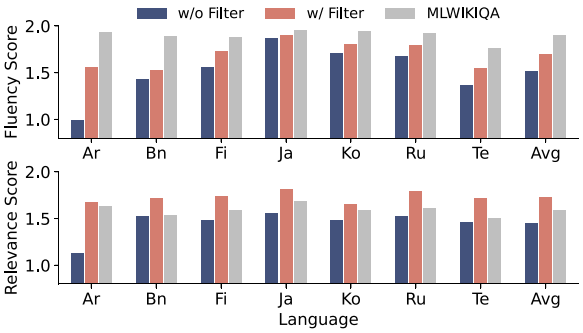

Figure 8: Quality validation results on the synthetic FsMlQA (with and without filter), comparing against the silver-standard pre-training data MlWikiQA. We employ *Model-as-Judge* to evaluate the quality of generated data on a three-level rating scale (0–2) based on two factors: fluency and relevance.

5 Data Analysis

5.1 Quality Validation

To assess the overall quality of our synthetic data, we randomly sample 1,000 examples from the silver pre-training data (MlWikiQA) and few-shot synthetic data (FsMlQA). These samples are evaluated using the `GPT-4o mini` to assess quality based on: 1) Fluency (0–2): assessing whether the query is understandable, readable, and free of spelling or grammatical mistakes; 2) Relevance (0–2): evaluating the alignment between the generated query-answer pair and the passage used for data generation. The prompts employed for quality assessment are included in Appendix Table 18.

Figure 8 illustrates that both types of our generated queries exhibit fluency and strong grounding in the corresponding positive passages. The silver-standard MlWikiQA, derived using heuristics from WikiData (§2.1), consistently achieves higher scores across both metrics in all languages compared to the unfiltered synthetic FsMlQA (w/o Filter columns). However, the quality of FsMlQA improves significantly after applying our tailored filtering mechanism (w/ Filter columns), almost matching the quality and fluency scores for MlWikiQA. This finding underscores the critical role of the filtering procedure in producing a synthetic dataset of quality comparable to the silver-standard dataset.

5.2 Query Distribution Comparison

To examine the distributional differences between our synthetic FsMlQA and the gold-standard data in Xor-TyDi QA, we randomly sample up

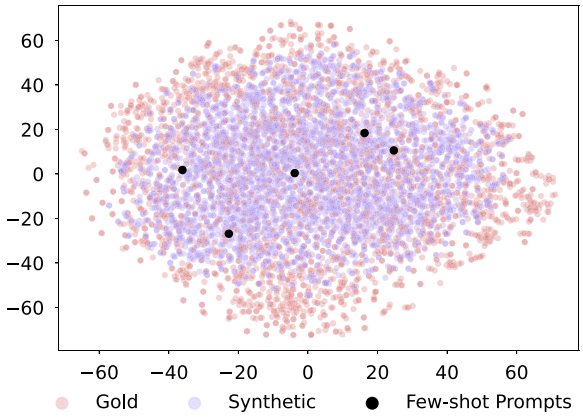

Figure 9: Distribution comparison between FsMlQA and Xor-TyDi QA in Japanese. We show that the synthetic data is diverse and significantly overlaps with the gold standard.

to 20,000 examples from both datasets and visualize their distributions using t-SNE (van der Maaten and Hinton, 2008), which projects the queries onto a two-dimensional space. Figure 9 highlights several key findings: 1) The synthetic queries exhibit sufficient diversity, as they are scattered across the plot, indicating that our approach is capable of generating queries of various types using only five labeled examples. 2) The synthetic data shows significant overlap with the gold-standard data, demonstrating that it retains the core characteristics of the gold distribution. 3) The gold-standard data exhibits greater diversity than the synthetic data, suggesting that there is still room for improvement in enhancing diversity and variation during the data generation process, which we leave for future work. Similar findings are observed in the other languages (see Appendix Figure 10).

5.3 Safety

We employ `Llama-Guard-2`[6] as the content safety classifier to assess the presence of unsafe content within our synthetic dataset. Our analysis reveals that 98.9% of the 1,746,156 queries in FsMlQA are classified as safe.

6 Related Work

Pre-training for Open-domain QA. Open-domain QA requires retrieving relevant passages and extracting answers from them. This necessity has driven various methods that jointly train retrievers and readers. REALM (Guu et al., 2020), RAG (Lewis et al., 2020), EMDR2 (Sachan et al., 2021), YONO (Lee et al., 2022), ReAtt (Jiang et al., 2022), and Atlas (Izacard et al., 2024) first pre-train retrievers or initialize from pre-trained (Izacard et al., 2022) and fine-tuned retrievers. Subsequently, both components are fine-tuned jointly: the reader is trained using an answer generation loss, and the retriever is trained to promote passages that increase the likelihood of generating correct answers. Recently, this joint training mechanism has been adapted for multilingual open-domain QA (Jiang et al., 2024), where retrievers are initially trained by learning from English teachers using multilingual parallel data, followed by a joint training stage with query-answer pairs generated by LLMs. Our approach follows this joint training paradigm for model pre-training but differs significantly. We use WikiData as a source to generate more informative natural questions and answers. Additionally, our pre-training method is more efficient by eliminating knowledge distillation from English models.

LLMs for Few-shot Data Generation. Prompting LLMs to generate synthetic data has been widely adopted to improve the performance of retrieval and QA tasks. UPR (Sachan et al., 2022) and InPars (Bonifacio et al., 2022) use zero-shot or few-shot prompting for passage reranking. Promptagator (Dai et al., 2023) and SWIM-X (Thakur et al., 2024) prompt LLMs with few-shot examples to generate massive synthetic queries, either in English or in multiple languages, for retriever fine-tuning. Gecko (Lee et al., 2024) prompts LLMs to generate synthetic instructions and queries from Web documents and create high-quality labels for retriever fine-tuning. Beyond retrieval, LLMs are employed to generate QA data, where QAmeleon (Agrawal et al., 2023) prompts a 540B LLM to generate multilingual QA pairs from only five examples. Nevertheless, these methods primarily focus on retrieval tasks and the more narrowly defined machine reading comprehension tasks. In our work, we rigorously investigate how LLMs can improve the more challenging multilingual open-domain QA tasks under few-shot settings. In addition, we explore zero-shot prompting, demonstrating that

[6] `https://huggingface.co/meta-llama/Meta-Llama-Guard-2-8B`.

cross-lingual prompting using English data or limited multilingual data from held-out languages can yield results comparable to few-shot prompting, and we show this technique can also be leveraged for effective zero-shot language adaptation.

7 Conclusion and Limitation

In this work, we propose FsModQA, a *few-shot learning* approach for multilingual open-domain retrieval tasks. We present a novel self-supervised pre-training framework that exploits WikiData to effectively initialize both multilingual retrieval and QA capabilities. This process is followed by few-shot synthetic multilingual QA generation from LLMs using only five human-annotated examples. We demonstrate that the resulting model achieves competitive multilingual retrieval and QA performance through fine-tuning on the high-quality synthetic data. We further show that this few-shot approach generalizes to zero-shot settings that only require English-supervised data. This mechanism serves as an effective approach for language adaptation, enabling the adapted model to achieve both boosted retrieval and end-to-end QA performance across fifteen previously unseen languages.

This work uses LLMs for synthetic data generation, which may propagate undesirable biases to generated data. We believe such biases will not be amplified as we sample prompts from Xor-TyDi QA, a dataset annotated with strict guidelines. Our preliminary safety analysis also reveals that only less than 1% data contains potentially harmful queries, as identified by `Llama-Guard-2`.

Acknowledgments

We thank the action editor Shay Cohen and anonymous reviewers for their helpful feedback and suggestions. The first author is supported by the Graduate Research Scholarships funded by the University of Melbourne. This work was funded by the Australian Research Council, Discovery grant DP230102775.

References

Dmitry Abulkhanov, Nikita Sorokin, Sergey Nikolenko, and Valentin Malykh. 2023. Lapca: Language-agnostic pretraining with cross-lingual alignment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2098–2102, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3539618.3592006

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. QAmeleon: Multilingual QA with Only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771. https://doi.org/10.1162/tacl_a_00625

Akari Asai, Jungo Kasai, Jonathan 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*, pages 547–564, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.46

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2387–2392, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3477495.3531863

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia

to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1171

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

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.dialdoc-1.19

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*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.74

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(1).

Fan Jiang, Tom Drummond, and Trevor Cohn. 2024. Pre-training cross-lingual open domain question answering with large-scale synthetic supervision. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13906–13933, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.770

Zhengbao Jiang, Luyu Gao, Zhiruo Wang, Jun Araki, Haibo Ding, Jamie Callan, and Graham Neubig. 2022. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2336–2349, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.149

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. https://doi.org/10.18653/v1/2020.emnlp-main.550

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

New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3397271.3401075

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

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177. https://doi.org/10.1162/tacl_a_00453

Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher Manning, and Kyoung-Gu Woo. 2022. You only need one model for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3047–3060, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.198

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint*, 2403.20327v1.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

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

Houxing Ren, Linjun Shou, Ning Wu, Ming Gong, and Daxin Jiang. 2022. Empowering dual-encoder with query generator for cross-lingual dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3107–3121, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.203

Devendra 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*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.249

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems*.

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. https://doi.org/10.18653/v1/2021.eacl-main.115

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.496

Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya, and Valentin Malykh. 2022. Ask me anything in your native language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 395–406, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.30

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint*, 2403.08295v4.

Nandan Thakur, Jianmo Ni, Gustavo Hernandez Abrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In *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, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.426

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint*, 2302.13971v1.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. https://doi.org/10.1145/2629489

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*, pages 483–498, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131. https://doi.org/10.1162/tacl_a_00595

Property ID	Description	Property ID	Description
P264	record label	P175	performer
P176	manufacturer	P112	founded by
P127	owned by	P840	narrative location
P495	country of origin	P20	place of death
P407	language of work or name	P582	end time
P69	educated at	P159	headquarters location
P740	location of formation	P17	country
P136	genre	P800	notable work
P36	capital	P570	date of death
P190	twinned administrative body	P4552	mountain range
P915	filming location	P3086	speed limit
P84	architect	P2046	area
P569	date of birth	P86	composer
P515	phase of matter	P2048	height
P40	child	P580	start time
P828	has cause	P50	author
P2067	mass	108	employer
P170	creator	P2049	width
P364	original language of film or TV show	P277	programmed in
P276	location	P413	position played on team / speciality
P131	located in the administrative territorial entity	P26	spouse
P106	occupation	P607	conflict
P942	theme music	P571	inception
P6	head of government	P19	place of birth
P1830	owner of	P61	discoverer or inventor

Table 10: List of English properties used for generating MLWIKIQA. Note that we do not generate data for a property if it does not exist in the Wikidata of target languages.

System:
You are a powerful question generator that generates natural and fluent questions from Wikidata triples.
User:
Rewrite this triplet "(テリー・ジャックス, 出生地, ウィニペグ)" into three questions in natural Japanese whose answer is "ウィニペグ". The questions should be diverse and the verbs used should be different. The generated questions should be separated by "\n".
System:
テリー・ジャックスの出生地はどこですか？
テリー・ジャックスはどこで生まれましたか?
テリー・ジャックスが生まれた都市は何という名前ですか？

System:
You are a powerful question generator that generates natural and fluent questions from Wikidata triples.
User:
Rewrite this triplet "(レアンダー・クラブ, 結成場所, ロンドン)" into three questions in natural Japanese whose answer is "yes". The questions should be diverse and the verbs used should be different. All elements in the triplet should be included in the question. The generated questions should be separated by "\n".
System:
レアンダー・クラブはロンドンで結成されましたか？
ロンドンはレアンダー・クラブの結成場所ですか？
レアンダー・クラブの創立がロンドンで行われたのですか？

Table 11: Examples of using ChatGPT to generate questions from triples. We use the same prompt as `Yes` questions to generate `No` ones by sampling perturbed triples. Highlighted texts indicate system outputs.

Triple: (伝染性単核球症, 原因, エプスタイン・バール・ウイルス)
Question: どのウイルスが伝染性単核球症を引き起こすことが知られていますか？
Answer: エプスタイン・バール・ウイルス

Triple: (東日本大震災による電力危機, 原因, 福島第一原子力発電所事故)
Question: 東日本大震災後の電力危機を引き起こした出来事は何ですか？
Answer: 福島第一原子力発電所事故

Triple: (脳死, 原因, 脳損傷)
Question: 脳死を引き起こすものとして、主に何が挙げられますか？
Answer: 脳損傷

Triple: (壊血病, 原因, ビタミンC欠乏症)
Question: 何が壊血病につながるのか
Answer: ビタミンC欠乏症

Table 12: An example of prompting Gemma-7B to generate questions with ICL examples from ChatGPT. Highlighted texts indicate system outputs.

Document: アダム・スミスは、私たちが一般的に資本主義と呼ぶものの最初の理論家と見なされています。彼の1776年の著作『国富論』は、ある安定した商業システムと評価システムの中で、個人が生産を専門化することでより多くの収益を得るというインセンティブに応えるだろうと理論化しました。これらの個人は特定の国家の介入なしに自然に「その産業を生産物が最も価値あるものとなるように指示することができる」でしょう。
Question: 資本主義の提唱者は誰？ => Answer: アダム・スミス

Document: 人気のあるフロントは、1936年5月3日の総選挙で、608議席中386議席を獲得しました。初めて、社会主義者がラディカル社会主義者よりも多くの議席を獲得し、社会主義者の指導者レオン・ブルムがフランス初の社会主義首相およびその職を保持する初のユダヤ人となりました。初めての人気のあるフロント内閣は、20人の社会主義者、13人のラディカル社会主義者、2人の社会主義共和党員で構成されており（共産主義者の閣僚はいませんでした）、初めて3人の女性が含まれていました（当時、フランスでは女性が投票することができませんでした）。
Question: フランスではいつから女性権利大臣が存在する？ => Answer: 1936年5月3日

Document: ハンガリー王国は中央ヨーロッパに存在した君主国で、中世から20世紀にかけて存在しました（1000年から1946年まで、1918年から1920年を除く）。ハンガリー公国は約1000年にエステルゴムで最初の国王スティーブン1世が戴冠し、キリスト教王国として現れました。彼の家族（アールパード王朝）は300年にわたって王国を指導しました。
Question: ハンガリー王国は何年間存在した？ => Answer: 1000年から1946年まで、1918年から1920年を除く

Document: アントマルキと英国人は別々に死体解剖報告書を書き、それぞれがナポレオンが父親を殺した病気である内出血によって死んだと結論づけました。後に、ナポレオンの髪のサンプルから高濃度のヒ素が見つかったことに基づく別の理論では、ナポレオンがヒ素中毒で死んだとされています。ただし、後の研究でもナポレオンの幼少期や息子、ジョゼフィーヌの髪のサンプルからも高濃度のヒ素が見つかりました。19世紀には医薬品やヘアクリームなどの製品でヒ素が広く使われていました。2021年に国際チームの消化器病理学者による研究では、ナポレオンは胃がんで亡くなったと結論づけられました。
Question: ナポレオンが死んだのはマラリアのせい？ => Answer: no

Document: フィリピンの1987年憲法は次のように述べています：「国と教会の分離は不可侵であるべきです。」（第II条第6節）、および、「宗教の設立を尊重する法律は制定されず、またその自由な行使を禁止する法律も制定されません。宗教の職業と礼拝の自由な行使と享受は、差別や優遇なしに永遠に許可されます。公民権や政治的権利の行使に宗教的試験は求められません。
Question: フィリピンは政教分離を原則としている？ => Answer: yes

Table 13: Complete prompt for few-shot question answer generation from passages in target language.

Document: Adam Smith is considered the first theorist of what we commonly refer to as capitalism. His 1776 work, An Inquiry into the Nature and Causes of the Wealth of Nations, theorized that within a given stable system of commerce and evaluation, individuals would respond to the incentive of earning more by specializing their production. These individuals would naturally, without specific state intervention, "direct ... that industry in such a manner as its produce may be of the greatest value."
Question: English: Who are the proponents of capitalism? => Japanese: 資本主義の提唱者は誰
Answer: English: Adam Smith => Japanese: アダム・スミス

Document: The Popular Front won the general election of 3 May 1936, with 386 seats out of 608. For the first time, the Socialists won more seats than the Radical-Socialists, and the Socialist leader Léon Blum became the first Socialist Prime Minister of France as well as the first Jew to hold that office. The first Popular Front cabinet consisted of 20 Socialists, 13 Radical-Socialists and two Socialist Republicans (there were no Communist Ministers) and, for the first time, included three women (women were not able to vote in France at that time).
Question: English: Since when does France have a Minister of Women's Rights? => Japanese: フランスではいつから女性権利大臣が存在する？
Answer: English: 3 May 1936 => Japanese: 1936 年 5 月 3 日

Document: The Kingdom of Hungary was a monarchy in Central Europe that existed from the Middle Ages into the twentieth century (1000–1946 with the exception of 1918–1920). The Principality of Hungary emerged as a Christian kingdom upon the coronation of the first king Stephen I at Esztergom in about the year 1000; his family (the Árpád dynasty) led the monarchy for 300 years.
Question: English: How many years did the Kingdom of Hungary exist? => Japanese: ハンガリー王国は何年間存在した
Answer: English: 1000 – 1946 with the exception of 1918–1920 => Japanese: 1000年から1946年まで、1918年から1920年を除く

Document: Antommarchi and the British wrote separate autopsy reports, each concluding that Napoleon had died of internal bleeding caused by stomach cancer, the disease that had killed his father. A later theory, based on high concentrations of arsenic found in samples of Napoleon's hair, held that Napoleon had died of arsenic poisoning. However, subsequent studies also found high concentrations of arsenic in hair samples from Napoleon's childhood and from his son and Joséphine. Arsenic was widely used in medicines and products such as hair creams in the 19th century. A 2021 study by an international team of gastrointestinal pathologists concluded that Napoleon died of stomach cancer.
Question: English: Did Napoleon die because of malaria? => Japanese: ナポレオンが死んだのはマラリアのせい？ Answer: English: no => Japanese: no

Document: The 1987 Constitution of the Philippines declares: The separation of Church and State shall be inviolable. (Article II, Section 6), and, No law shall be made respecting an establishment of religion, or prohibiting the free exercise thereof. The free exercise and enjoyment of religious profession and worship, without discrimination or preference, shall forever be allowed. No religious test shall be required for the exercise of civil or political rights.
Question: English: Does the Philippines follow the principle of separation of church and state?=> Japanese: フィリピンは政教分離を原則としている？
Answer: English: yes => Japanese: yes

Table 14: Complete prompt for few-shot cross-lingual question answer generation from English passages.

English Document: Lindzen was born on February 8, 1940 in Webster, Massachusetts.[1] His father, a shoemaker, had fled Hitler's Germany with his mother. He moved to the Bronx soon after his birth and grew up in a Jewish household in a predominantly Catholic neighborhood there.[3][5] Lindzen attended the Bronx High School of Science (winning Regents' and National Merit Scholarships), Rensselaer Polytechnic Institute,and Harvard University.[6] From Harvard, he received an A.B. in physics in 1960, followed by an S.M. in applied mathematics in 1961 and a PhD in applied mathematics in 1964. His doctoral thesis, Radiative and photochemical processes in strato- and mesospheric dynamics,[7] concerned the interactions of ozone photochemistry, radiative transfer, and dynamics in the middle atmosphere.
English Question: Where was Richard Siegmund Lindzen born? => English Answer: Webster, Massachusetts

English Document: On 30 September 2011, Justice Johnson Lam of the Court of First Instance of the High Court (CFI) ruled in Vallejos' case that existing legislation restricting FDHs from qualifying for permanent residence contravened the Hong Kong Basic Law. Lam also found that Vallejos and Domingo, but not the three other applicants, had fulfilled the condition of taking Hong Kong as their only permanent home and being ordinarily resident in Hong Kong for seven years. The Court of Appeal of the High Court overturned the CFI's decision on Vallejos' case on 28 March 2012. Vallejos and Domingo then jointly appealed to the Court of Final Appeal (CFA), which heard their case on 26–28 February 2013; the CFA rejected their appeal on 25 March 2013.
English Question: Who was the judge in the case of Vallejos and Domingo v. Commissioner of Registration? => English Answer: Justice Johnson Lam

English Document: Human dissections were carried out by the Greek physicians Herophilus of Chalcedon and Erasistratus of Chios in the early part of the third century BC.[7] During this period, the first exploration into full human anatomy was performed rather than a base knowledge gained from 'problem-solution' delving.[8] While there was a deep taboo in Greek culture concerning human dissection, there was at the time a strong push by the Ptolemaic government to build Alexandria into a hub of scientific study.[8] For a time, Roman law forbade dissection and autopsy of the human body,[9] so physicians had to use other cadavers. Galen, for example, dissected the Barbary macaque and other primates, assuming their anatomy was basically the same as that of humans.[10][11][12]
English Question: Who first performed human dissection? => English Answer: Herophilus of Chalcedon and Erasistratus of Chios

English Document: On 16 March 1915, Watson gained his Royal Aero Club Certificate No. 1,117 (equivalent of a pilot's licence) with the London and Provincial School at the London Aerodrome, Hendon, having sought a commission with the Royal Naval Air Service with the outbreak of war in 1914.[6] Sadly, on 30 June 1915 he lost his life when the Caudron G.3 aeroplane he was flying disintegrated in flight and crashed in Dunlye field, a few miles from the Cross-in-Hand Hotel, Sussex. Watson is buried in Dundee's Western Cemetery.[2]
English Question: Where is Preston Albert Watson buried? => English Answer: Dundee's Western Cemetery

English Document: A paywall is a method of restricting access to content via a paid subscription.[1][2] Beginning in the mid-2010s, newspapers started implementing paywalls on their websites as a way to increase revenue after years of decline in paid print readership and advertising revenue.[3] In academics, research papers are often subject to a paywall and are available via academic libraries that subscribe.[4][5]
English Question: What is a paywall? => English Answer: a method of restricting access to content via a paid subscription

Japanese Document: モンロー郡はは1815年6月29日に、ヨーロッパ系アメリカ人によって設立された。アラバマが州に昇格する以前のこと>だった。当初の開拓者はイギリス人の子孫であり、バージニア州、ジョージア州および両カロライナ州から来ていた。アッパー・クリーク族ウィンド・クランの著名な酋長レッド・イーグル（ウィリアム・ウェザーフォードとも呼ばれた）が、クリーク戦争（1813年-1814年）の後でこの地に入り、プランテーションを造り上げた。レッド・イーグルはクリーク>族とヨーロッパ人の血を引いており、資産である奴隷を農園主や馬の飼育者に育てた。クリーク族員の大半は、1830年代にアラバマからインディアン準州（現在のオクラホマ州）に移住させられた。その後に入ってきたヨーロッパ系アメリカ人は、奴隷労働者を連れてくるか、土地を取得した後に奴隷を購入した。
Japanese Question: モンロー郡はいつ設立されましたか? => Japanese Answer: 1815年6月29日

Table 15: An example for zero-shot English Prompting. Highlighted texts indicate system outputs.

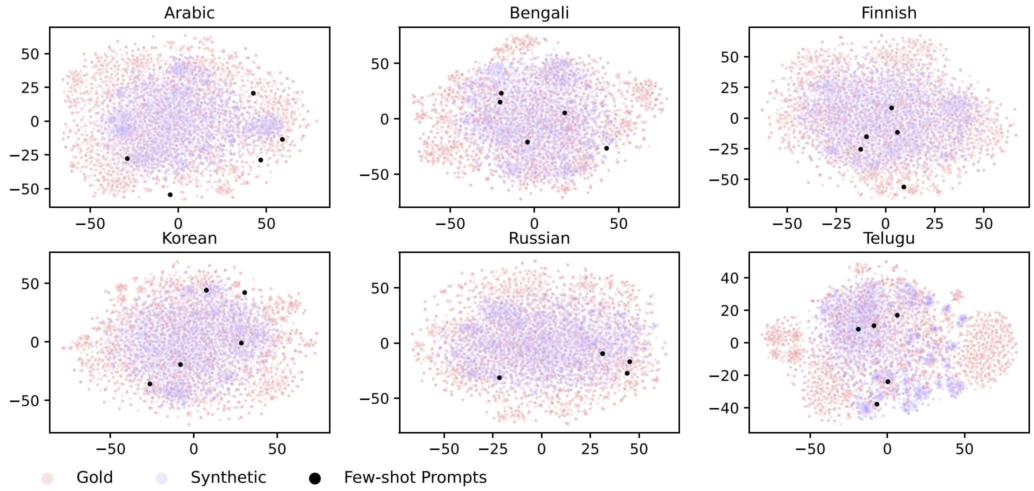

Figure 10: Distribution comparison between FsMlQA and Xor-TyDi QA in the rest languages. We demonstrate that the diverse synthetic data can be expanded from only five-shot examples and retains the core characteristics of the gold distribution.

Finnish Document: Tel Avivissa asuu 467 875 ihmistä, jotka jakautuvat 52 000 dunamin (52 km²; 20 neliömailia) suuruiselle alueelle, mikä tuottaa väestötiheyden 7 606 ihmistä neliökilometrillä (19 699 ihmistä neliömaililla). Israelin keskusviraston (CBS) mukaan vuoteen 2009 mennessä Tel Avivin väestö kasvaa vuosittain 0,5 prosentilla. Kaikkien taustojen juutalaiset muodostavat 91,8 prosenttia väestöstä, muslimit ja arabikristityt 4,2 prosenttia ja loput kuuluvat muihin ryhmiin (mukaan lukien eri kristilliset ja aasialaiset yhteisöt). Koska Tel Aviv on monikulttuurinen kaupunki, siellä puhutaan monia kieliä heprean lisäksi. Joissakin arvioissa noin 50 000 rekisteröimätöntä afrikkalaista ja aasialaista ulkomaalaistyöntekijää asuu kaupungissa. Verrattuna länsimaisiin kaupunkeihin, rikollisuus Tel Avivissa on suhteellisen vähäistä.
Finnish Question: Onko Tel Aviv monikulttuurinen maa? => Finnish Answer: no

Russian Document: Во время войны в Вооруженных Силах Соединенных Штатов служило более 16 миллионов американцев, из которых 405 399 погибли в бою, а 671 278 были ранены. Также было 130 201 американских военнопленных, из которых 116 129 вернулись домой после войны. Ключевыми гражданскими советниками президента Рузвельта были министр войны Генри Л. Стимсон, который мобилизовал промышленность и центры призыва для обеспечения армии, командованной генералом Джорджем Маршаллом, и Военно-воздушных сил под командованием генерала Хапа Арнольда. Военно-морской флот, возглавляемый министром военно-морского флота Фрэнком Ноксом и адмиралом Эрнестом Кингом, оказался более автономным. Общие приоритеты устанавливал Рузвельт.
Russian Question: Сколько американских солдат участвовало во Второй Мировой войне? => Russian Answer: 16 миллионов

Bengali Document: দক্ষিণের বিরুদ্ধে সংগ্রাম চালানো যারা কানসাস-নেব্রাস্কা আইনের প্রতি বিরোধী ছিলেন, সে সম্মিলিত হযেছিল সন ১৮৫৪ তে। সে আইন একটি আইন ছিল যা কানসাস ও নেব্রাস্কা এলাকার পশ্চিমাঞ্চলে মালিকানা ক্রমবর্ধন সম্ভব করত। সে ক্লাসিক্যাল লিবারেলিজম এবং অর্থনৈতিক সংস্কারে সমর্থন করতে, কিন্তু মুক্ত অঞ্চলে গুলামির প্রশারণের বিরুদ্ধে ছিল। পার্টির অধিকাংশই প্রাথমিকভাবে দক্ষিণে উপস্থিতি ছিল, কিন্তু উত্তরে সফল ছিল। ১৮৫৮ সালের পর্যন্ত, এটি অধিকাংশ পূর্বগামী ও প্রাক্তন ফ্রি সইলারদের সহযোগিতা নিযে উত্তরের প্রায সমস্ত রাজ্যে বড বড বৃহত্তরত্ব গঠন করেছিল। সাদা দক্ষিণের মানুষরা গুলামির বিপন্নতায় চিন্তিত হযে উঠেছিলেন। ১৮৬০ সালে প্রথম রিপাব্লিকান রাষ্ট্রপতি আব্রাহাম লিংকনের নির্বাচনের সাথে, মহান দক্ষিণী রাষ্ট্রগুলি ইউনাইটেড স্টেটস থেকে বিচ্ছিন্ন হযে গেল।
Bengali Question: মার্কিন যুক্তরাষ্ট্রের রিপাবলিকান পার্টির প্রথম প্রেসিডেন্ট কে ছিলেন ? => Bengali Answer: আব্রাহাম লিঙ্কন

Arabic Document: ةسماخلا ةيقارعلا ةقرفلا عاطق يف ةيلاتقلا تابكرملاو ةيقارعلا ئجلاملا ترمد ةنماثلا ناسرفلا نم ىلولأا ةبيتكلاو ،عوردلا كلذكو نيطابلا يداو برغ يقارعلا عافدلا يف رغثلا ربع ةاشملل ىلولأا ةقرفلا نم ةيناثلا ةقرفلا تمدقت ،١٩٩١ رياربف ٢٤ يف .ةاشملل نيرشعلاو ريهطتو يقارعلا عافدلا رسكب عوردلا نم ٣-٣٧ ةمهملا ةعومجم تماق .ةيناودعلا ةمواقملا نم رغثلا عقوملل يقرشلا يلامشلا عاطقلا ريهطتب تماق ةعبرأ تارمم عيسوتو ةوجفلا تحت نارين ودعلا ةرشابملا. ًاضيأ يف ٢٤ رياربف، تماق ةقرفلا ىلولأا نم ةقرفلا ىلولأا ةاشملل نواعتلاب عم ةقرفلا ىلولأا نم ناسرفلا ريمدتب زجاوحلا ةيقارعلا ودو اهتايرو ةعباتلا ةقرفلل ةيقارعلا ةسداسلا نيرشعلاو.
Arabic Question: وهام مسا لوأ ةقطنم ةيقارع تضرعت وزغلل يكيرملأا؟ => Arabic Answer: يداو نطابلا

Korean Document: 스미스-풋넘 풍력 터빈은 세계 최초의 1메가와트 규모의 풍력 터빈이었습니다. 1941년에 버몬트주 캐슬턴의 그랜드파스 노브에 연결되어 현지 전력 공급 시스템에 연결되었습니다. 이 터빈은 파머 코슬렛 풋넘이 디자인하고 S. 모건 스미스 회사에서 제조했습니다. 이 1.25메가와트 터빈은 1100시간 동안 작동한 후 전쟁 중 재료 부족으로 약점이 알려진 곳에서 날개가 파손되었습니다. 이후 1979년까지는 최대 규모의 풍력 터빈으로 남았습니다.
Korean Question: 세상에서 가장 큰 풍력 에너지 발전소는 무엇인가? => Korean Answer: 스미스-퍼트남 풍력 터빈

Japanese Document: 細菌性髄膜炎の原因として>多い肺炎球菌と髄膜炎菌は鼻咽腔上皮細胞に付着しコロニーを形成する。そこから血管内に侵入し脳室内脈絡叢に到達する。脈絡叢上皮細胞に直接感染し脳脊髄液中に入ることができる。脳脊髄液中では免疫防御機構が機能しないため細菌は急速に増殖する。細菌性髄膜炎の発症機序において重要なのは浸潤した細菌が誘発する炎症反応である。細菌性髄膜炎の神経症状や合併症の多くは、細菌による組織の直接的な破壊よりもむしろ、浸潤した細菌に対する免疫応答によって引き起こされている。結果として、抗生物質療法により脳脊髄液が無菌化された後になっても神経の損傷は進行しうる
Japanese Question: 細菌性髄膜炎の最も一般的な原因は何か? => Japanese Answer: 肺炎球菌と髄膜炎菌は鼻咽腔上皮細胞に付着しコロニーを形成する。

Table 16: An example for zero-shot Multilingual Prompting. Highlighted texts indicate system outputs.

Question: {Example question #1}
Answer: {Example answer #1}
...
Question: {Example question #5}
Answer: {Example answer #5}

Passage #1 Title: {Passage #1 Title}
Passage #1 Text: {Passage #1 Text}
...
Passage #N Title: {Passage #N Title}
Passage #N Text: {Passage #N Text}

Task description: predict the {Test Question Language} answer to the following question. The answer should be a minimal span extracted from the document. You should only output the answer.

Question: {Test question}
Answer:

Table 17: Prompt template for few-shot multilingual QA with LLMs.

Relevance Assessment
You are given a Q&A pair and a paragraph. Your goal is to Rate the relevance of the Q&A pair to the paragraph on a scale from 0 to 2.
0: Very low relevance, the Q&A pair and paragraph are almost unrelated.
1: Moderate relevance, the Q&A pair and paragraph share some overlap.
2: High relevance, the Q&A pair are strongly grounded by the paragraph.
Output Format:
Relevance (0–2)
Only provide the final result in the above structured format without any additional explanations.
Paragraph: {Paragraph}
Q: {Synthetic Query}
A: {Synthetic Answer}

Fluency Assessment
You are given a question. Your goal is to Rate the fluency of the question on a scale from 0 to 2.
0: Poor fluency, the question is unclear, contains significant grammatical errors, or is incomprehensible.
1: Moderate fluency, the question has minor grammatical errors or awkward phrasing but is still understandable.
2: High fluency, the question is clear, well-structured, and grammatically correct.
Output Format:
Fluency (0–2):
Only provide the final result in the above structured format without any additional explanations.
Question: {Synthetic Query}

Table 18: Prompt template for quality validation of synthetic data using `Model-as-Judge`.

	MLWIKIQA			FSMLQA		
	# Q-A Paris	Question Length	Answer Length	# Q-A Paris	Question Length	Answer Length
Arabic	1,803,765	$7.00_{\pm 2.13}$	$1.65_{\pm 0.84}$	80,575	$8.20_{\pm 2.86}$	$1.57_{\pm 1.30}$
Bengali	407,496	$6.13_{\pm 1.80}$	$1.65_{\pm 0.85}$	127,562	$8.97_{\pm 2.99}$	$1.63_{\pm 1.44}$
English	7,963,985	$7.95_{\pm 2.54}$	$1.78_{\pm 1.01}$	–	–	–
Finnish	2,135,790	$6.02_{\pm 1.75}$	$1.32_{\pm 0.64}$	270,627	$5.83_{\pm 2.09}$	$1.38_{\pm 0.90}$
Japanese	2,735,635	$14.74_{\pm 3.57}$	$3.57_{\pm 1.73}$	143,265	$10.19_{\pm 2.18}$	$3.96_{\pm 4.69}$
Korean	1,018,348	$5.46_{\pm 1.78}$	$1.55_{\pm 0.80}$	192,002	$5.72_{\pm 2.29}$	$1.42_{\pm 0.92}$
Russian	2,561,925	$6.94_{\pm 2.17}$	$1.70_{\pm 1.10}$	792,914	$7.34_{\pm 2.64}$	$1.44_{\pm 1.03}$
Telugu	108,215	$5.60_{\pm 1.84}$	$1.50_{\pm 0.74}$	139,211	$6.48_{\pm 2.48}$	$1.49_{\pm 1.17}$

Table 19: Dataset statistics of our pre-training data MLWIKIQA and few-shot synthetic data FSMLQA in each language.

Prompting Strategy	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg.
FSMODQA	5.9%	13.9%	16.9%	7.7%	21.4%	15.4%	13.2%	13.4%
NQ-En	3.4%	8.8%	8.2%	1.2%	8.3%	5.9%	4.2%	5.3%
TYDI-En	5.0%	13.6%	12.8%	1.9%	17.0%	6.3%	5.9%	7.0%
XOR-TYDI-*	10.2%	14.6%	14.2%	2.5%	22.7%	9.1%	10.7%	9.8%

Table 20: Success Rate of synthetic data generation across seven languages with different prompting strategies. Success Rate = valid examples after data filtering / total examples (i.e., # Documents).

Method	F1							Macro Average		
	Ar	**Bn**	**Fi**	**Ja**	**Ko**	**Ru**	**Te**	**F1**	**EM**	**BLEU**
5-shot										
FsModQA-En	35.6	32.7	35.5	35.1	30.2	33.6	31.8	33.5	23.8	23.0
FsModQA	**41.3**	**35.4**	**39.6**	**41.5**	**35.0**	**38.2**	**36.3**	**38.2**	**27.9**	**24.4**
16-shot										
FsModQA-En	38.3	31.0	39.4	38.3	35.2	34.9	34.6	35.9	26.1	24.1
FsModQA	**42.0**	**35.6**	**41.4**	**41.7**	**35.3**	**39.2**	**40.0**	**39.3**	**29.3**	**26.6**
32-shot										
FsModQA-En	42.4	31.2	40.8	38.1	33.0	37.9	34.9	36.9	26.3	25.5
FsModQA	**43.6**	**35.6**	**42.2**	**42.5**	**34.1**	**38.6**	**37.0**	**39.1**	**28.8**	**26.6**
128-shot										
FsModQA-En	42.0	28.8	41.7	40.3	**34.6**	34.7	36.0	36.9	27.0	25.2
FsModQA	**45.3**	**32.8**	**44.3**	**43.8**	34.0	**39.9**	**42.1**	**40.3**	**30.5**	**27.4**
1024-shot										
FsModQA-En	45.0	30.8	45.1	39.2	34.1	39.1	37.5	38.7	29.3	26.5
FsModQA	**47.5**	**33.7**	**46.7**	**41.4**	**35.9**	**40.2**	**40.1**	**40.8**	**31.3**	**27.9**
full										
FsModQA-En	48.9	**33.3**	47.7	42.9	**39.6**	40.0	41.7	42.0	32.7	28.5
FsModQA	**50.8**	**33.3**	**47.8**	**45.0**	38.9	**42.0**	**43.1**	**43.0**	**33.4**	**29.6**

Table 21: Detailed results in each language when trained with varying sizes of supervised data.