Pre-training Cross-lingual Open Domain Question Answering with Large-scale Synthetic Supervision

Anonymous ACL submission

Abstract

Cross-lingual open domain question answering (CLQA) is a complex problem, comprising cross-lingual retrieval from a multilingual knowledge base, followed by answer generation in the query language. Both steps are usually tackled by separate models, requiring substantial annotated datasets, and typically auxiliary resources, like machine translation systems to bridge between languages. In this paper, we show that CLQA can be addressed using a single encoder-decoder model. To effectively train this model, we propose a self-supervised method based on exploiting the cross-lingual link structure within Wikipedia. We demonstrate how linked Wikipedia pages can be used to synthesise supervisory signals for cross-lingual retrieval, through a form of cloze query, and generate more natural questions to supervise answer generation. Together, we show our approach, CLASS, outperforms comparable methods on both supervised and zero-shot language adaptation settings, including those using machine translation.

1 Introduction

Open Domain Question Answering (QA) is the task of generating an answer for a given question based on the evidence gathered from a large collection of documents. A widely adopted pipeline "*retrieve-then-read*" is employed for this task (Chen et al., 2017; Karpukhin et al., 2020), which begins by retrieving a small set of passages using a dense retrieval model and subsequently processes retrieved passages to generate the answer with a dedicated reader. Unlike English open-domain QA, where both questions and knowledge sources share the same language, multilingual open-domain QA presents new challenges, as it involves retrieving evidence from multilingual corpora, considering that many languages lack comprehensive support documents or the questions require knowledge from diverse cultures (Asai et al., 2021b).

Several attempts have been made to enhance the performance of multilingual open-domain QA (Asai et al., 2021b; Abulkhanov et al., 2023). These approaches typically require passage labels for retriever training through supervised contrastive learning. This requirement complicates cross-lingual retrieval training significantly due to the challenge of constructing a large-scale dataset containing query-passage labels. This challenge emerges from the unavailability of prior knowledge regarding which language contains the relevant evidence. Furthermore, these efforts often involve separate training of the retriever and reader, leading to error propagation within the resulting pipeline.

Evidence in the context of English open-domain QA reveals that integrating retriever and reader training typically leads to improved performance on both components. This achievement is often realised by training both components (Guu et al., 2020; Lewis et al., 2020) or a unified model that performs both tasks (Lee et al., 2022; Jiang et al., 2022) through fully end-to-end training. Nonetheless, such a joint training paradigm has not been extensively explored in multilingual open-domain QA, and how to adapt it to suit the complexities of multilingual settings remains an open question.

In this paper, we introduce the first *unified model* capable of performing both cross-lingual retrieval and multilingual open-domain QA tasks. To achieve this, we propose **CLASS** (**C**ross-**L**ingual **QA** Pre-training with **S**ynthetic **S**upervision), a self-supervised method to pre-train the model with multilingual texts at scale. CLASS comprises two core components: **cross-lingual retrieval pre-training** that equips the model with robust cross-lingual retrieval ability, and **multilingual QA pre-training** that further enhances retrieval and QA abilities jointly. Concretely, as depicted in Figure 1, the pre-training data is created by mining parallel queries from parallel Wikipedia pages, using salient entities within English sentences as answers. To facil-

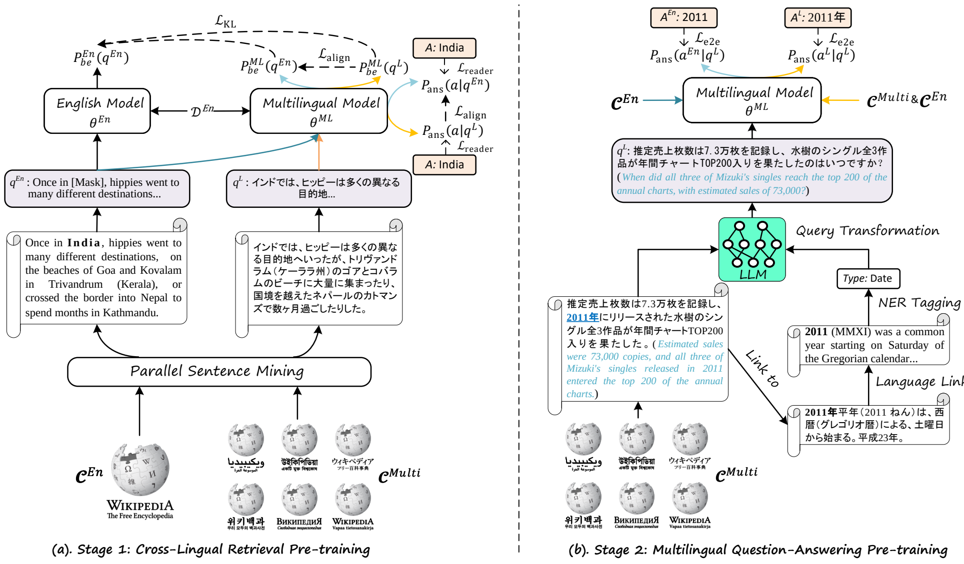


Figure 1: The overview of our two-stage unsupervised pre-training method for cross-lingual open domain question answering. English translations from Google Translate are added in (b) for readability.

itate cross-lingual retrievals, a knowledge distillation process is introduced, requiring the model to match the distributions of a well-trained English teacher when given queries in both languages. The follow-up is a self-supervised learning task for end-to-end pre-training by propagating training signals derived from the end QA task. This process entails generating pre-training data using anchor texts indicated by hyperlinks and a *question transformation* technique to resemble the formats of natural questions. Notably, our approach does not necessitate additional tools such as machine translation and offers a more convenient application to low-resource languages, requiring only comparable documents (i.e., Wikipedia language links).

This large-scale pre-training framework empowers the model to demonstrate promising unsupervised performance, and it can even outperform many competitive supervised counterparts. By fine-tuning it with supervised English and multilingual QA data, we can attain further improvements, ultimately establishing new state-of-the-art performance in both cross-lingual retrieval and multilingual open-domain QA tasks. In summary, our contributions are:[1]

1. Empirical results on the XOR-TYDI QA benchmark demonstrate that CLASS outperforms a wide range of prominent unsupervised, zero-shot, and supervised models on both tasks, while solely relying on QA pairs throughout the whole training processes.
2. On the MKQA dataset, CLASS exhibits remarkable generalisation capabilities across linguistically diverse languages without using human-annotated data.
3. To the best of our knowledge, we are the pioneers in systematically exploring the advantages of pre-training for multilingual retrieval and open-domain QA tasks. This demonstrates the feasibility of achieving multilingual open-domain QA within a unified model.

[1] Code and data will be released upon acceptance.

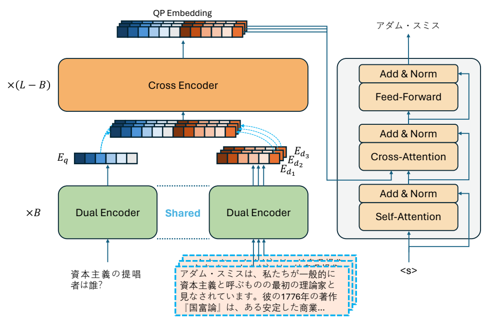


Figure 2: The unified model for passage retrieval and question answering.

2 Preliminaries

2.1 Task Definition

Given a query q^L in language L, **Cross-lingual Passage Retrieval** requires retrieving a collection of passages $\mathcal{D}^{En}$ from English Wikipedia $\boldsymbol{C}^{\mathbf{En}}$ that potentially provide evidence to answer q^L. In contrast, **Multilingual Open-Domain Question Answering** aims at answering q^L in language L by referring to a multilingual Wikipedia $\boldsymbol{C}^{\mathbf{Multi}}$. In this setting, the prior knowledge of which language contains the evidence is unavailable, and the relevant passages can be retrieved from any language.

2.2 Model Architecture

Figure 2 shows the overall structure of our model. In this model, the bottom layers of the encoder function as the *retriever*, encoding queries and passages independently for efficient retrieval. The remaining encoder layers and the entire decoder are designated as the *reader* for question answering.

Retriever. The retriever is a bi-encoder that uses the first B encoder layers with H heads to encode query q and passages d from a corpus $\mathcal{D}$. We use the query Q and key vectors K in $B+1$-th layer as their embeddings, respectively (Jiang et al., 2022):

$$E_{\mathbf{d}} = \{K_{\mathbf{d}}^{B+1,h} \in \mathbb{R}^{|\mathbf{d}|\times e}\}_{h=1}^{H},$$
$$E_{\mathbf{q}} = \{Q_{\mathbf{q}}^{B+1,h} \in \mathbb{R}^{|\mathbf{q}|\times e}\}_{h=1}^{H},$$

where $|\mathbf{q}|$ and $|\mathbf{d}|$ are sequences lengths and e is the dimension of each head.

The self-attention matrix $\mathrm{SA}_{\mathbf{q},\mathbf{d}}^{B+1,h}$ from a specific head ($h=6$ (Jiang et al., 2022)) is considered the source of retrieval scores. A sum of max computations (Khattab and Zaharia, 2020) is performed to reduce it to yield the retrieval score:

$$s_{\text{mv}}(q,d) = \textstyle\sum_{i\in|\mathbf{q}|} \max_{j\in|\mathbf{d}|} \mathrm{SA}_{i,j}^{B+1,h},$$
$$\mathrm{SA}_{\mathbf{q},\mathbf{d}}^{B+1,h} = Q_{\mathbf{q}}^{B+1,h} \times K_{\mathbf{d}}^{B+1,h^{\top}} \in \mathbb{R}^{|\mathbf{q}|\times|\mathbf{d}|}.$$

We denote this as **Multi-Vector Retrieval** and consider it as our *default setting*. We also explore **Dense Retrieval**, which takes the average pooling of layer B's output with LayerNorm as query $Q_{\mathbf{q}}$ and passage $K_{\mathbf{d}}$ representations, and the relevance is measured by their dot product:

$$s_{\text{dense}}(q,d) = \mathrm{LN}(Q_{\mathbf{q}}) \cdot \mathrm{LN}(K_{\mathbf{d}}).$$

The top-k most relevant passages are then retrieved by $\mathcal{D}_q = \arg\mathrm{topk}_{d_i\in\mathcal{D}} P_{\text{be}}(\cdot|q,D) = \arg\mathrm{topk}\left[s(q,d_0),\ldots,s(q,d_{|d\in\mathcal{D}|})\right]$.

Reader. The encoded query and each top-k passage in $\mathcal{D}_q$ are concatenated and fed into the remaining *cross-encoder* layers. Finally, the joint encodings $\{E_{\mathbf{q},\mathbf{d_i}}\}_{i=0}^{|\mathcal{D}_q|}$ are integrated into the decoder through cross-attention to generate the answer a efficiently (Izacard and Grave, 2021b): $P_{\text{ans}}(a|q,\mathcal{D}_q) = \log\prod_{t=1}^{T} P(a_t|a_{<t},q,\mathcal{D}_q)$.

3 Method

We propose an unsupervised two-stage pre-training method for cross-lingual open-retrieval question answering, as depicted in Figure 1. Our approach starts with **cross-lingual retrieval pre-training**, where the *unified multilingual model* develops excellent cross-lingual dense retrieval capabilities. This proficiency is acquired through learning from a well-trained English model, employing cloze-style parallel queries and retrieved English passages as inputs. The subsequent stage involves **pre-training for multilingual question-answering (QA)**, where the *unified model* is further pre-trained on multilingual question-answer pairs that are automatically generated. This process entails selecting potential answers from anchor texts and applying our novel *question transformation* techniques to convert cloze questions into natural questions by prompting a large language model.

3.1 Cross-Lingual Retrieval Pre-training

Pre-training Data. We consider cloze questions, which are statements with the answer masked, as pseudo queries. The answers are salient spans selected from named entities. We extract all named entities for an English sentence using a NER system, generating queries for each. Formally, let s^{En} be a sentence sampled from an English Wikipedia page $\mathcal{W}^{En}$, along with its associated named entities $\{a_i\}_{i=1}^{n}$. This allows us to derive cloze queries $\{q_i^{En}\}_{i=1}^{n}$ by masking each entity a_i. Then, for each q_i^{En}, the objective is to identify its translation q_i^{L} in language L by searching from sentences $\{q_j^{L}\}_{j=0}^{n}$ within a Wikipedia page $\mathcal{W}^{L}$, which is connected to $\mathcal{W}^{En}$ via language links in Wikipedia.

We use a margin-based mining method (Artetxe and Schwenk, 2019) to identify parallel sentences based on their similarity in the embedding space:

$$\mathrm{M}(q_i,q_j) = \frac{\cos(q_i,q_j)}{\sum_{z\in N_{q_i}}\frac{\cos(q_i,z)}{2k}+\sum_{z\in N_{q_j}}\frac{\cos(q_j,z)}{2k}},$$

where N_{q_i} and N_{q_j} are the top-k neighbours of sentence q_i and q_j in the other language, respectively. $cos(q_i,q_j)$ denotes the cosine similarity between the embeddings of q_i and q_j extracted using `mSimCSE` (Wang et al., 2022). We apply this scoring function to q_i^{En} and each $q_j^{L} \in \{q_j^{L}\}_{j=0}^{n}$. Pairs whose scores surpass a threshold T are selected as parallel queries, denoted as $\{q_i^{En}, q_j^{L}, a_i\}$.[2]

Training. A well-trained English model θ^{En} is employed to teach a multilingual model θ^{ML} using parallel queries. Specifically, given a training example $\{q^{En}, q^{L}, a\}$, we employ θ^{En} to retrieve a set

[2]We identify a_i and mask it in q_j^{L} through string match if L is written in Latin script and leave q_j^{L} unchanged otherwise.

of relevant passages $\mathcal{D}_{q^{En}}$ from English Wikipedia $\boldsymbol{C}^{\mathbf{En}}$ for q^{En}. The multilingual model is then compelled to align its retrieval distributions with those of θ^{En} over $\mathcal{D}_{q^{En}}$ through KL divergence loss:

$$\begin{aligned}\mathcal{L}_{\text{KL}} = \mathbb{KL}(P^{ML}_{\text{be}}(\cdot|q^{L}, \mathcal{D}_{q^{En}})||P^{En}_{\text{be}}(\cdot|q^{En}, \mathcal{D}_{q^{En}})) \\ +\mathbb{KL}(P^{ML}_{\text{be}}(\cdot|q^{En}, \mathcal{D}_{q^{En}})||P^{En}_{\text{be}}(\cdot|q^{En}, \mathcal{D}_{q^{En}})).\end{aligned}$$

Additionally, θ^{ML} is trained to predict the answer a with either q^{En} or q^{L} as the question:

$$\mathcal{L}_{\text{reader}} = -P_{\text{ans}}(a|q^{En}, \mathcal{D}_{q^{En}}) - P_{\text{ans}}(a|q^{L}, \mathcal{D}_{q^{En}}).$$

Moreover, to ensure that the multilingual model generates consistent predictions across languages, we introduce an alignment regularisation term:

$$\begin{aligned}\mathcal{L}_{\text{align}} = \mathbb{KL}(P^{ML}_{\text{be}}(\cdot|q^{L}, \mathcal{D}_{q^{En}})||P^{ML}_{\text{be}}(\cdot|q^{En}, \mathcal{D}_{q^{En}})) \\ +\mathbb{KL}(P_{\text{ans}}(a|q^{L}, \mathcal{D}_{q^{En}})||P_{\text{ans}}(a|q^{En}, \mathcal{D}_{q^{En}})).\end{aligned}$$

Overall, θ^{ML} is trained with the weighted combined loss: $\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{reader}} + \alpha \cdot (\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{align}})$.

3.2 Multilingual QA Pre-training

The cloze questions used in §3.1 are substantially different from the formats of natural questions asked by real users, which inherently impedes the development of advanced QA skills. Moreover, the incapacity to precisely locate and mask the answer a within q^{L} for perfectly aligned queries makes the QA task notably simpler, as a implicitly appears in q^{L} (e.g., "インド" in q^{L} is the Japanese answer in Figure 1 (a)). Meanwhile, since q^{En} and q^{L} could be roughly aligned, the querying of a by q^{L} is not assured, thereby introducing noise into the pre-trained data (e.g., "In *1945*, his father sent him to Collège des Frères" and "父はサブリーをヤッファのカトリック系フランス語学校に送った。" are aligned but the Japanese query does not mention the answer *1945*). Thus, we design another pre-training technique to address the limitations above.

3.2.1 Pre-training Data

The construction of pre-training data in this stage involves two sequential steps. Initial data are first acquired from a multilingual Wikipedia source in the format of cloze questions, followed by a format transformation into natural questions.

Initial Data. In contrast to English texts, where robust NER systems facilitate the detection of named entities with high precision for answer generation, such systems in other languages exhibit inherent deficiencies. Instead, we employ anchor texts with hyperlinks as answer candidates. Specifically, for a given sentence s^{L} in language L, we consider the anchor texts $\{a^{L}_{i}\}^{n}_{i=0}$ within it as potential answers and construct cloze questions $\{s^{L}_{i}\}^{n}_{i=0}$ accordingly.

For each a^{L}_{i}, we fetch the Wikipedia page $\mathcal{W}^{L}$ to which it links and access the corresponding English Wikipedia page $\mathcal{W}^{En}$ via language link. Subsequently, the title a^{En}_{i} of $\mathcal{W}^{En}$ is assumed to be the pseudo translation of a^{L}_{i} (Figure 1 (b)). Moreover, NER tagging is performed on the first paragraph of $\mathcal{W}^{En}$ to identify the type t_i of the title entity a^{En}_{i}, which is then assigned to a^{L}_{i}. Finally, a training example is derived as $(s^{L}_{i}, a^{L}_{i}, a^{En}_{i}, t_i)$.

Query Transformation. We employ large language models (LLMs) for query transformation via In-Context Learning (ICL) (Brown et al., 2020).

We first prompt ChatGPT (`gpt-3.5-turbo`) to generate a few examples as meta-examples (Fan et al., 2023) for ICL. Specifically, we randomly sample instances from the initial dataset and generate transformed questions based on the structure of the prompt shown in Prompt 3.1.

Prompt 3.1: Meta-Example Generation

Rewrite this sentence $\{s^{L}_{i}\}$ into a natural question whose question word is {`wh_word`} and answer is $\{a^{L}_{i}\}$. Please respond in the format: "The transformed question is: $\{q^{L}_{i}\}$"

where `wh_word` is chosen according to the entity type t_i through heuristics (Lewis et al., 2019). This step yields a curated set of ICL examples: $\mathbb{K} = \{c^{L}_{i}, \texttt{wh\_word}, a^{L}_{i}, q^{L}_{i}\}^{k}_{i=0}$. An example is shown in Figure 11 in the Appendix.

Subsequently, the curated ChatGPT examples are used as the source to few-shot prompt a smaller LLM, LLaMA-2-7B (Touvron et al., 2023), to generate many more instances efficiently. We include the prompting examples in Appendix E.

3.2.2 Joint Training

The retriever learns indirectly from the answer generation task, taking the cross-attention score from the decoder as the target for query-passage relevance measurement (Izacard and Grave, 2021a):

$$\mathcal{L}_{\text{KL}} = \mathbb{KL}(P_{\text{be}}(\cdot|q^{L}, \mathcal{D}_{q^{L}})||P_{\text{ca}}(\cdot|q^{L}, \mathcal{D}_{q^{L}})),$$

$$P_{\text{ca}}(d_i|q^{L}, \mathcal{D}_{q^{L}}) = \sum_{h=0}^{H}\sum_{t=0}^{|d_i|}\frac{\text{SG(CA}(0,h,t))}{H} \mid d_i \in \mathcal{D}_{q^{L}},$$

where $\mathcal{D}_{q^{L}}$ is the set of passages returned by the retriever itself and P_{ca} is the target distribution gathered from the decoder's cross-attention scores. SG signifies stop-gradient, which blocks the gradient to ensure the decoder is not 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 $|d_i|$ is the length of passage d_i.

The reader optimises the negative log-likelihood of generating a^L given q^L and relevant passages $\mathcal{D}_{q^L}$ as input: $\mathcal{L}_{\text{reader}} = -P_{\text{ans}}(a^L|q^L, \mathcal{D}_{q^L})$. The final loss combines reader and retriever loss: $\mathcal{L}_{\text{e2e}} = \alpha \cdot \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{reader}}$.

Asynchronous Passage Update. During training, we need to use the retriever to gather a set of passages $\mathcal{D}_{q^L}$ from $\boldsymbol{C}^{\textbf{Multi}}$ for each (q^L, a^L).[3] However, since the retriever's parameters are updated constantly, employing the latest model for retrieval becomes computationally expensive due to the need for recomputing all passage embeddings. To ensure efficient training, we periodically update the retrieved passages for each training query using the most recent model every 1000 steps.

4 Experiments

Datasets, Baselines and Metrics. We evaluate our model on the XOR-TYDI QA dataset (Asai et al., 2021a), with XOR-Retrieve for cross-lingual retrieval, and XOR-Full for multilingual open-domain QA. We employ MKQA (Longpre et al., 2021) for zero-shot evaluation on unseen languages. We use the February 2019 English Wikipedia dump as $\boldsymbol{C}^{\textbf{En}}$ and use the Wikipedia dumps of the same date, consisting of 13 diverse languages from all 7 languages of XOR-TYDI QA and a subset of MKQA languages as $\boldsymbol{C}^{\textbf{Multi}}$ (Asai et al., 2021a).

We compare retrieval performance with translate-test methods DPR+MT (Asai et al., 2021a), multilingual dense passage retrievers mDPR, CORA, Sentri, QuiCK, LAPCA, SWIM-X (Asai et al., 2021a,b; Sorokin et al., 2022; Ren et al., 2022; Abulkhanov et al., 2023; Thakur et al., 2023), and multi-vector retriever DrDecr (Li et al., 2022). We report top-n retrieval accuracy, the fraction of queries for which the top-n retrieved tokens contain the answer. We compare QA results with multilingual models that use BM25 for monolingual retrieval, translate-test models MT+DPR, GMT+GS, MT+Mono and ReAtt+MT (Asai et al., 2021a; Jiang et al., 2022), and multilingual fusion-in-decoder models CORA, Sentri and LAPCA using F1, exact match (EM) and BLEU scores.

[3] We replace a^L with a_i^{En} and $\boldsymbol{C}^{\textbf{Multi}}$ with $\boldsymbol{C}^{\textbf{En}}$ when focusing on cross-lingual retrieval from English corpus.

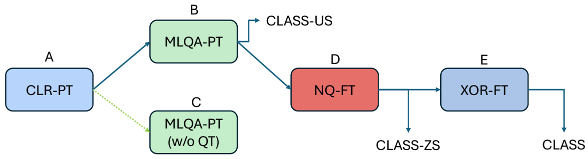


Figure 3: Our training pipeline. CLR: cross-lingual retrieval, MLQA: multilingual question answering, QT: query transformation, PT: pre-training, FT: fine-tuning.

4.1 Experimental Settings

Pre-training Corpus. In cross-lingual retrieval pre-training, we gather the parallel pages across various languages for each $\mathcal{W}^{En} \in \boldsymbol{C}^{\textbf{En}}$. We consider 15 distinct languages, with 7 from XOR-TYDI QA and 8 being high-resource or closely related to the 7 evaluated languages. Parallel sentences are mined from each pair of parallel pages. A state-of-the-art NER tagger is applied to each English sentence, and we retain pairs that contain named entities.

In multilingual QA pre-training, data generation is limited to 7 languages on XOR-TYDI QA. We employ LLaMA-2-7B to generate one transformed question per training example with 3 randomly sampled meta-examples in the same language as the prompt. We generate multiple questions for each example in low-resource languages. More details are in Appendices A.1.1 and A.1.2.

Training Sequence. Figure 3 shows the complete pre-training and fine-tuning sequence. *i) Cross-lingual Retrieval Pre-training* (CLR-PT): We pre-train `mt5-large` (Xue et al., 2021) as in §3.1 to get `CLASS-US-Stage1`, with English teacher being ReAtt (Jiang et al., 2022) trained on NQ (Kwiatkowski et al., 2019). *ii) Multilingual QA Pre-training* (MLQA-PT): `CLASS-US-Stage1` is further pre-trained as in §3.2 to obtain the unsupervised `CLASS-US`. *iii) Fine-tuning*: We first fine-tune `CLASS-US` on NQ to obtain the zero-shot `CLASS-ZS`, which is then trained on supervised data from XOR-TYDI QA to derive `CLASS`. We use the same training objective $\mathcal{L}_{\text{e2e}}$ as in MLQA-PT.

4.2 Main Results

XOR-Retrieve. Table 1 shows the results on the dev set of XOR-Retrieve. `CLASS`, which exclusively employs question-answer pairs for training, demonstrates a substantial performance advantage over all baselines that rely on passage labels for contrastive learning. This advantage is particularly pronounced

Method	# Total Params	Pre-train Data	Fine-tuning Data	R@2kt								R@5kt							
				Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg	Ar	Bn	Fi	Ja	Ko	Ru	Te	Avg
Unsupervised Retrievers																			
LAPCA§	560M	Wikipedia	—	51.1	50.2	48.6	35.1	57.3	32.2	64.4	48.4	61.0	58.4	52.6	40.5	66.7	40.8	70.1	55.7
SWIM-X	580M	mC4	SWIM-IR	50.8	65.1	56.1	48.1	54.0	55.7	66.4	56.6	57.9	75.0	65.6	59.3	58.9	64.6	74.4	65.1
CLASS-US	410M	Wikipedia	—	**66.0**	**75.7**	**63.4**	**57.7**	**63.5**	**68.8**	**70.6**	**66.5**	**71.2**	**81.6**	**69.4**	**66.8**	**70.5**	**75.1**	**77.3**	**73.1**
w/ Dense	410M	Wikipedia	—	54.4	67.4	58.6	47.7	51.6	59.9	65.6	57.9	64.8	73.0	64.7	57.3	58.6	67.9	70.6	65.3
Zero-shot Retrievers																			
DPR+MT†	220M	—	NQ	43.4	53.9	55.1	40.2	50.5	30.8	20.2	42.0	52.4	62.8	61.8	48.1	58.6	37.8	32.4	50.6
LAPCA§	560M	Wikipedia	NQ+XPAQ	46.2	50.3	56.6	41.4	48.7	52.3	54.6	50.0	53.0	60.5	66.2	49.7	56.1	60.7	63.8	58.6
ReAtt+MT	583M	—	NQ	63.1	67.7	20.7	55.9	60.3	55.3	58.4	54.5	67.3	71.0	29.3	61.8	67.0	61.2	66.4	60.6
CLASS-ZS	410M	Wikipedia	NQ	**65.1**	**79.3**	**67.8**	**60.6**	**61.1**	**69.2**	**74.4**	**68.2**	**72.5**	**83.2**	**73.9**	**70.5**	**69.1**	**75.1**	**81.9**	**75.2**
w/ Dense	410M	Wikipedia	NQ	59.2	70.1	59.9	51.5	57.2	51.5	72.3	60.2	66.7	78.6	66.6	60.2	63.2	58.2	78.2	67.4
(Semi-) Supervised Retrievers																			
CORA	557M	—	NQ+XOR	32.0	42.8	39.5	24.9	33.3	31.2	30.7	33.5	42.7	52.0	49.0	32.8	43.5	39.2	41.6	43.0
mDPR†	557M	—	NQ+XOR	38.8	48.4	52.5	26.6	44.2	33.3	39.9	40.5	48.9	60.2	59.2	34.9	49.8	43.0	55.5	50.2
Sentri§	560M	—	NQ+TQA+XOR	47.6	48.1	53.1	46.6	49.6	44.3	67.9	51.0	56.8	62.2	65.5	53.2	55.5	52.3	80.3	60.8
QuiCK	557M	—	NQ+XOR	52.8	70.1	62.2	54.8	62.8	57.8	70.6	61.3	63.8	78.0	65.3	63.5	69.8	67.1	74.8	68.9
DrDecr*	278M	WikiMatrix	NQ+XOR	-	-	-	-	-	-	-	66.0	70.2	85.9	69.4	65.1	68.8	68.8	83.2	73.1
LAPCA§	560M	Wikipedia	NQ+XPAQ+XOR	61.1	76.9	**72.6**	60.9	69.1	69.1	75.6	69.3	70.2	83.8	**79.6**	69.7	73.6	75.5	83.1	76.5
CLASS	410M	Wikipedia	NQ+XOR	**67.3**	**80.9**	67.2	**64.7**	**71.6**	**69.6**	**79.8**	**71.6**	**74.8**	**84.5**	72.3	**73.9**	**79.3**	**77.2**	**85.3**	**78.2**
w/ Dense	410M	Wikipedia	NQ+XOR	66.7	79.6	64.3	58.1	66.0	64.1	77.7	68.1	70.6	84.9	71.0	66.0	72.6	70.0	81.9	73.9

Table 1: Results on the dev set of XOR-Retrieve. The best and second-best results are marked in **bold** and underlined. † denotes results reported by Asai et al. (2021a). ∗ indicates human-translated supervised parallel queries released by XOR-Retrieve are used for training. § represents methods that employ MT systems for training data augmentation.

Method	# Total Params	Pre-training Data	Fine-tuning Data	F1							Macro Average		
				Ar	Bn	Fi	Ja	Ko	Ru	Te	F1	EM	BLEU
BM25†	—	—	XOR	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–
MT+DPR†	—	—	NQ	7.2	4.3	17.0	7.9	7.1	13.6	0.5	8.2	3.8	6.8
ReAtt+MT	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
MT+Mono†	—	—	NQ+XOR	15.8	9.6	20.5	12.2	11.4	16.0	0.5	17.3	7.5	10.7
CORA†	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	1.23B	Wikipedia	NQ+XOR	**49.5**	**32.0**	**49.6**	**44.7**	37.5	**41.4**	**42.0**	**42.4**	**32.7**	**29.2**
w/ Dense	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
Incomparable Models (for Reference)													
Sentri§	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§	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**

Table 2: QA results on the XOR-Full dev set. The best and second-best results are marked in **bold** and underlined. † denotes results from Asai et al. (2021b). § indicates methods that use synthetic and translated English datasets.

under unsupervised and zero-shot settings, where both variants, CLASS-US and CLASS-ZS, achieve improvements of more than 10% over state-of-the-art methods ($p < 0.001$).[4] The **Dense Retrieval** variant (i.e., w/ Dense) consistently outperforms other competitive baselines and is comparable to LAPCA with only 73% of the parameters. This highlights that our approach is versatile and can be applied to enhance various kinds of retrievers.

XOR-Full. Table 2 reports the results of CLASS on XOR-Full. Both CLASS and the variant employing dense retrieval achieve superior performance when compared to a series of baseline models and the prior state-of-the-art CORA model in all tested languages, showcasing an average improvement of up to 7.8% ($p < 0.001$). Compared to methods that rely on machine translation to generate a substantially larger pool of multilingual training data from English datasets, CLASS is comparable to Sentri but falls behind LAPCA.[5] The most pronounced performance gaps are in Bengali and Korean, with the fewest two training samples available within XOR-Full. We believe it is the translated QA pairs used by Sentri and LAPCA that alleviate such discrepancies, and further improvements are expected when integrating such augmented data.

MKQA. We assess the zero-shot performance of CLASS in various unseen languages included in MKQA. Figure 4 shows that in cross-lingual retrieval tasks, all variants of our method exhibit

[4]Paired Student's t-test (Dror et al., 2018).

[5]A direct comparison with Sentri and LAPCA is not feasible since the Wikipedia pages they employed as knowledge sources are different from ours and Asai et al. (2021b).

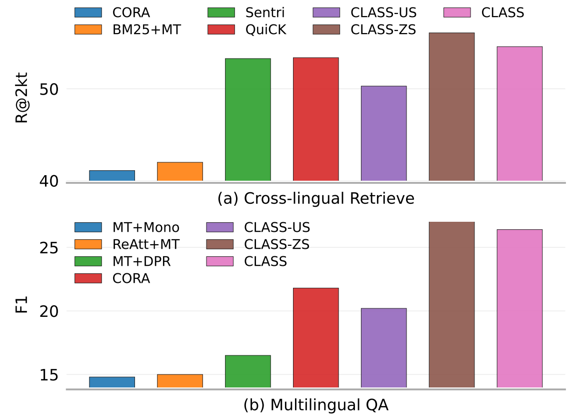


Figure 4: Zero-shot cross-lingual retrieval and multilingual QA results on unseen languages of MKQA.

promising results. Notably, CLASS-US surpasses the supervised model CORA significantly, and further fine-tuning on English data leads to substantial improvements. Interestingly, CLASS underperforms CLASS-ZS, despite being further fine-tuned on multilingual data. We attribute this phenomenon to three factors: the limited number of queries in XOR-Retrieve leads to overfitting to these specific languages; the query topics differ, as MKQA was translated from NQ while XOR-Retrieve questions were created by native speakers in target languages; the answer type differs (free spans *v.s.* WikiData aligned entities). In the multilingual QA task, we observe similar patterns where CLASS-ZS achieves the best zero-shot performance across unseen languages while supervised fine-tuning on XOR-Full hurts the generalisability. Detailed results in each language are in Appendix B Tables 4 and 5.

4.3 Analysis

We include quantitative and qualitative error analysis in Appendix C and additional numeric results in Appendix D (Figures 8, 9, 10)).

Cross-lingual Retrieval Pre-training Ablations. We conduct ablation studies to understand the impact of different components in cross-lingual retrieval pre-training, with results shown in Figure 5.

Effects of Learning from Parallel Queries. Removing queries either in English (-q^{En}) or in target languages (-q^{L}) leads to performance degradation. Meanwhile, the cross-lingual alignment regularisation (-$\mathcal{L}^{align}$) benefits the model by ensuring consistent predictions across languages.

Comparison with Different Parallel Query Sources. When comparing the approaches of gathering parallel queries, our method outperforms code-switching (w/ CS), which creates pseudo-

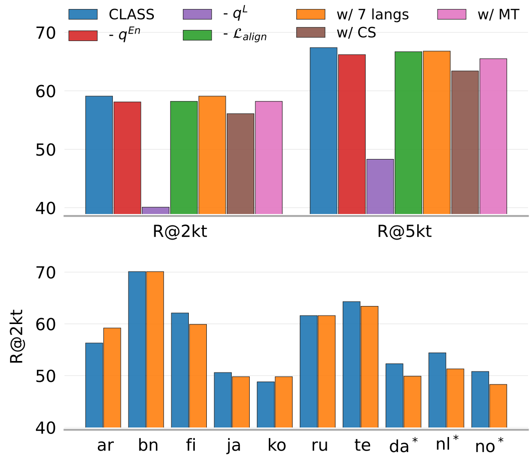


Figure 5: Ablations on cross-lingual retrieval pre-training, with results on the XOR-Retrieve dev set reported. * indicates unseen languages from MKQA.

Method	XOR-Retrieve		XOR-Full				
	R@2kt	R@5kt	F1	EM	BLEU	R^L@N	R^M@N
Unsupervised							
CLASS-US (AB)	**66.5**	**73.1**	**18.4**	**12.0**	**14.6**	60.0	69.1
- MLQA-PT (A)	59.1	67.4	5.7	3.9	4.0	55.7	**74.7**
- Query TF (AC)	66.1	**73.1**	7.2	4.8	4.9	**60.1**	65.4
Zero-shot							
CLASS-ZS (ABD)	**68.2**	**75.2**	**23.9**	**15.8**	**19.4**	**59.2**	69.1
- MLQA-PT (AD)	62.9	71.1	13.7	8.1	8.3	57.0	**76.2**
- Pre-train (D)	27.6	36.3	15.4	9.6	11.0	52.5	58.6
Supervised							
CLASS (ABDE)	**71.6**	**78.2**	42.4	32.7	29.2	62.8	**78.4**
- MLQA-PT (ADE)	69.6	75.7	**42.5**	**33.1**	29.1	**63.1**	77.8
- Pre-train (DE)	62.8	69.3	41.9	32.6	28.7	62.4	71.7

Table 3: Effects of two-stage pre-training. Results on the dev sets are reported. Symbols within brackets are described in Figure 3. R^L@N and R^M@N means the percentage of the questions whose top-N (N=100) passages contain an answer string in the target or any language.

translations through lexicon replacement based on bilingual dictionaries, and machine translations (w/ MT). This inferiority is primarily attributed to the limited coverage of bilingual dictionaries and poor translation quality in low-resource languages.

Sensitivity to Pre-training Language. Removing the extra 8 high-resource languages (w/ 7 langs) does not impact average performance but *affects specific low-resource languages* in XOR-TYDI QA. In particular, adding languages related to Telugu and Japanese (e.g., Tamil & Chinese) yields improvements. *Moreover, including a wider range of languages improves generalisation to unseen low-resource languages with limited parallel Wikipedia links* (e.g., adding German data enhances understanding of the West Germanic languages: Danish, Dutch, and Norwegian).

Effects of Two-stage Pre-training. We evaluate the efficacy of our two-stage proposed pre-training framework. Table 3 showcases the performance on both XOR-Retrieve and XOR-Full under unsupervised, zero-shot, and supervised settings. Integrating multilingual QA pre-training dramatically boosts performance in both unsupervised and zero-shot scenarios. Merely employing cloze-style questions instead of transformed natural questions has minimal impacts on retrieval but yields sub-optimal QA results, highlighting the importance of synthetic natural questions in QA tasks. When discarding the entire pre-training process, we observe a notable drop in both datasets. In supervised settings, the advantages of pre-training diminish with labelled data. This is especially evident in XOR-Full, where the differences between CLASS and the other two variants in QA and in-language retrieval (R^L@N) results diminish. While pre-training significantly improves cross-lingual evidence retrieval (R^M@N 71.7% -> 78.4%), CLASS does not benefit from this, suggesting its heavy reliance on in-language evidence and inability to reason over cross-lingual evidence when generating answers. See Appendix C for more detailed error analysis.

5 Related Work

Multilingual Dense Retrieval. Dense retrievers adopt pre-trained language models and follow a dual-encoder architecture (Karpukhin et al., 2020) to encode queries and passages into dense vectors and calculate the similarity scores. Effective techniques were proposed to advance English dense retrievals, including hard negative mining (Xiong et al., 2021), multi-vector representations (Khattab and Zaharia, 2020), and distilling from cross-encoder rerankers (Ren et al., 2021). With the advent of multilingual pre-trained models, these techniques were adapted to improve cross-lingual dense retrievals (Asai et al., 2021b; Ren et al., 2022). However, all these methods rely on passage labels for contrastive learning, which is challenging to obtain in cross-lingual settings. In contrast, our method explores a semi-supervised method and shows that a competitive cross-lingual retriever can be achieved using only query-answer pairs.

Multilingual Retrieval Pre-training. Large-scale unsupervised retrieval pre-training has significantly enhanced dense retrievers (Gao and Callan, 2021; Izacard et al., 2022) in processing English texts. Pre-training has also been explored in cross-lingual and multilingual dense retrieval, with a particular emphasis on augmenting the cross-lingual alignment capabilities of models. LAPCA (Abulkhanov et al., 2023) is trained through extensive cross-lingual contrastive learning, employing texts from parallel Wikipedia pages and parallel texts generated by machine translation systems. DrDecr (Li et al., 2022) learns from English models but operates on a smaller scale and relies on supervised parallel queries. In this work, we delve into the potential of large-scale unsupervised pre-training for cross-lingual dense retrieval and show that the resulting model exhibits high efficacy, outperforming many supervised ones.

Pre-training for Retrieval-Augmented Multilingual QA. In the context of English, jointly training a retriever and reader on supervised query-answer pairs (Sachan et al., 2021; Lewis et al., 2020) or large-scale unsupervised data derived from masked salient span masking (Guu et al., 2020; Lee et al., 2022) have been shown to enhance the performance of both retrieval and question answering tasks. However, the application of such a joint training paradigm, whether in supervised training or unsupervised pre-training, has not been explored in cross-lingual and multilingual settings. Our study represents the first investigation into this issue and proposes a curated pre-training framework within a unified model to address both retrieval and question-answering tasks. We introduce a two-stage pre-training procedure to initially equip a multilingual model with robust cross-lingual retrieval abilities by learning from English experts and then gradually evolving it through exposure to large-scale multilingual QA pairs. This approach yields remarkable unsupervised results and significant performance improvements across unseen languages without annotated training data.

6 Conclusion

In this paper, we explore the potential of a unified model for both cross-lingual retrieval and multilingual QA tasks. By incorporating our proposed pre-training paradigm, CLASS, the model's performance can be significantly improved, achieving both boosted retrieval and QA performance, while exhibiting impressive zero-shot transfer abilities to numerous unseen languages. Detailed ablations and thorough analyses are conducted to assess the efficacy of each component within our approach. Our future work aims at scaling CLASS to a broader range of languages to further enhance the model's cross-lingual transfer performance.

Limitations

The proposed pre-training framework incurs additional training costs when compared to standard supervised training, such as various pre-training data generation pipelines. The entire training pipeline requires approximately two weeks to complete with a maximum of 32 A100 GPUs. This could be less practical for researchers who do not have access to sufficient GPU resources. Nonetheless, common techniques such as *gradient accumulation* can be applied to adapt our approach for training in a more academic setting, although more training time is required to achieve comparable results.

Both stages in our pre-training paradigm depend on the availability of parallel Wikipedia pages. This can pose a challenge when dealing with languages that have limited resources even in terms of monolingual texts. Our approach may fail when no language links exist between English and a specific low-resource language. One may resort to employing a multi-hop approach to discover parallel Wikipedia pages, by first searching for the language linked to the low-resource language within Wikipedia and then repeating this process iteratively until reaching the corresponding English page. Another option could be relying on the generalisation of the multilingual model by training it in closely-related languages. Our analysis has revealed that incorporating a high-resource language in the pre-training phase consistently results in improvements for other languages within the same language family (Figure 5), which makes this issue less of a concern. Nevertheless, it remains imperative to explore methods for reducing the reliance on parallel Wikipedia texts, as this is essential to scale our method to more diverse and unique languages, which is worth exploring as a future work.

This work does not examine the benefits of pre-training in a broader range of languages and the scaling effects of both model size and data size for multilingual QA tasks, which is an interesting research topic that should be addressed rigorously in the future.

As this work uses large language models for *query transformation*, it is possible that undesirable biases (e.g., gender and cultural) inherent in these language models may be propagated to downstream systems. Furthermore, the extensive corpus of Wikipedia texts, drawn from a multitude of languages, could potentially introduce a diverse array of biases related to races and cultures to the pre-trained model. Assessing the magnitude of bias within the pre-training data and its subsequent impact on the model is an inherently intricate problem, which remains an open question for future research. Theoretically, our model can incorporate information extracted from any external corpus to generate answers to asked questions. This capability carries the potential for significant information leakage or the exposure of potentially toxic content from the corpus, which underscores the need for exercising caution when applying our method in sensitive domains.

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, page 2098–2102, New York, NY, USA. Association for Computing Machinery.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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. 2021a. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021b. 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.

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.

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 39–48, New York, NY, USA. Association for Computing Machinery.

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.

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.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. 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. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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.

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*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

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.

Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. 2023. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. 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 *International Conference on Learning Representations*.

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.

Overview of Appendix

Our supplementary includes the following sections:

- Section A: Experimental Settings, including implementation details, datasets, and compared baselines.
- Section B: Full zero-shot evaluation results on MKQA.
- Section C: Error analysis on multilingual open-domain question answering with quantitative and qualitative results.
- Section D: Additional numeric analysis.
- Section E: Prompts and examples for query transformation in each target language.

A Experimental Settings

A.1 Implementation Details

A.1.1 Parallel Queries Mining

Our implementation encompasses 15 distinct languages, namely **Arabic**, **Bengali**, German, Spanish, **Finnish**, French, Italian, **Japanese**, **Korean**, **Russian**, **Telugu**, Tamil, Malayalam, Kannada, Chinese. Parallel queries are collected from parallel Wikipedia pages for each `en-x`. Using unsupervised contrastive learning, we adopt the approach in Wang et al. (2022) to first pre-train a multilingual model `XLM-R`[6] on English Wikipedia texts by taking the dropout as a form of data augmentation. The resulting model is proficient in generating universal cross-lingual sentence embeddings without the need for parallel data, demonstrating robust zero-shot cross-lingual transfer capabilities. Subsequently, we deploy the pre-trained model for extracting multilingual sentence embeddings and mining parallel queries for each `en-x` language pair. Empirically, we set the margin-score threshold to 1.5 for most languages; however, for Japanese and Chinese, we observe improved performance with a larger threshold of 1.65. This process yields 5.4 million examples for the training, with the number of parallel queries for each language pair `en-x` shown in Figure 6.

We employ a balanced sampling strategy to avoid the training bias towards high-resource languages. For N number of languages $\{D_i\}_{i=1}^N$ with probabilities, $\{p_i\}_{i=1}^N$, we define the following multinomial distribution to sample from:

$$p_i = \frac{f_i^\alpha}{\sum_{j=1}^N f_j^\alpha}, \text{where } f_i = \frac{n_i}{\sum_{j=1}^N n_j},$$

where α is the sampling factor, which is set to 0.5 by following CONNEAU and Lample (2019) and n_i is the total number of parallel queries in the i-th language. During training, we use this to determine n'_i, the number of parallel queries in each language; and top-n'_i queries are used for training according to the margin-based scores. For every pair of mined query, we employ a state-of-the-art Named Entity Tagger from Stanza (Qi et al., 2020)[7] to find salient entities within the English query and take all identified entities as answer candidates to construct cloze-style queries.

A.1.2 Query Transformation

We use ChatGPT to generate 32 meta-examples. We then employ LLaMA-2-7B[8] for query transformation by randomly sampling 3 meta-examples to construct prompts for each test instance, with the format as shown in Prompts E.1, E.2, E.3, E.4, E.5, E.6, and E.7. We use Bloomz-7B[9] for Telugu as we find LLaMA-2-7B does not work well in this language. The Question word `wh_word` is chosen based on the entity type of the answer according to the heuristic rules in Table 10. Ultimately, 146K examples are generated per language, resulting in a total of 1M training instances.

A.1.3 Training Details

We use `mt5-large`[10] to initialise the model. In stage-1, we train the model for 64k steps on 32 A100 GPUs, which takes about one week to complete. The passages for all training queries are retrieved by the English teacher at once before training. In stage-2, we further train the model for 16k steps on 16 A100 GPUs with roughly 4 days. We periodically update the retrieved passages for each training instance every 1k steps using the most recent model. For fine-tuning, we first train the model on NQ with 8k steps and fine-tune the model on XOR-Retrieve for 6k steps and 12k steps on XOR-Full, which takes about 19 hours and 156 hours to complete, respectively. Likewise, we also do passage refreshing periodically every 1k steps.

For all training stages, we use the same batch size of 64 queries with each paired with 100 retrieved passages and learning rate 5×10^{-5}. We set α to 8 in all training loss functions. We set the maximum query and passage lengths to 50 and 200

[6] `https://huggingface.co/xlm-roberta-large`

[7] `https://github.com/stanfordnlp/stanza`

[8] `https://huggingface.co/meta-llama/Llama-2-7b`

[9] `https://huggingface.co/bigscience/bloomz-7b1`

[10] `https://huggingface.co/google/mt5-large`

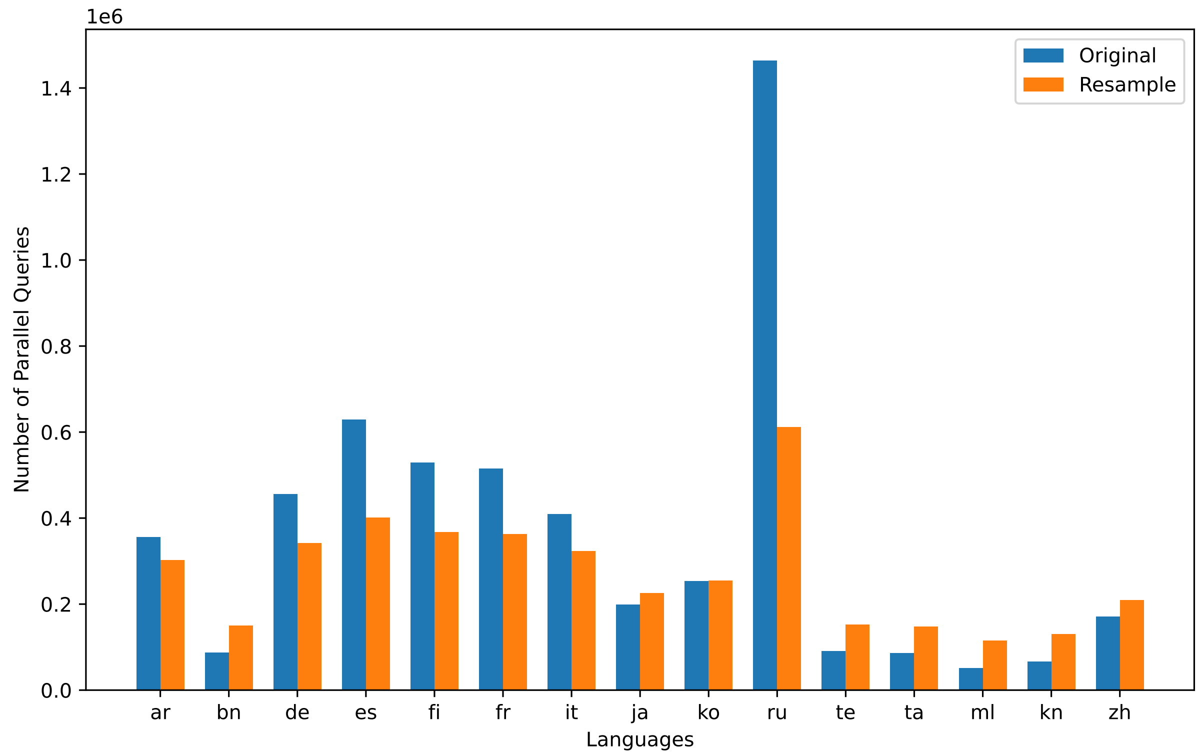


Figure 6: The number of mined parallel queries for each language pair en−x.

for both training and evaluation.

For the **Dense Retrieval** variant, we follow the same training and hyperparameter settings. The only difference is that this configuration is significantly more efficient, with training time reduced by half for multilingual QA pre-training and fine-tuning.

A.2 Datasets

We used the following datasets for model evaluation in our experiments:

- **XOR-Retrieve** (Asai et al., 2021a). It is under the MIT License. It contains 15250 QA pairs for training and takes the 20190201 English Wikipedia dump which contains 18M passages as the retrieval database.
- **XOR-Full** (Asai et al., 2021a). It is under the MIT License, containing 61360 training examples and a set of 43M passages as the retrieval corpus, collected from 20190201 Wikipedia dumps across 13 languages, namely English, Arabic, Finnish, Japanese, Korean, Russian, Bengali, Telugu, Indonesian, Thai, Hebrew, Swedish, and Spanish.
- **Natural Questions** (Kwiatkowski et al., 2019). It is under the Apache License and contains 79168 QA pairs.
- **MKQA** (Longpre et al., 2021). It is under the Apache License. This dataset covers 26 linguistically diverse languages, namely Arabic, Danish, German, English, Spanish, Finnish, French, Hebrew, Hungarian, Italian, Japanese, Korean, Khmer, Malay, Dutch, Norwegian, Polish, Portuguese, Russian, Swedish, Thai, Turkish, Vietnamese, Chinese (Simplified), Chinese (Hong Kong), and Chinese (Traditional). For the cross-lingual retrieval task, each language contains 6620 questions and the retrieval database consists of 18M English Wikipedia passages. For the multilingual QA task, each language contains 6758 questions and it uses the same retrieval database as XOR-Full.

A.3 Baselines

A.3.1 Cross-lingual Passage Retrieval

We compare our proposed model with a range of strong baselines:

- **mDPR.** This is the multilingual version of Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) encoder, which undergoes initial training on English NQ queries followed by fine-tuning on XOR-Retrieve.
- **DPR+MT** (Asai et al., 2021a). This is a translate-test baseline that involves the translation of queries into English during test time, followed by monolingual passage retrieval using the English DPR encoder.

- **CORA** (Asai et al., 2021b). This method trains a multilingual DPR encoder iteratively, with positive and negative passages identified by a multilingual QA model.
- **Sentri** (Sorokin et al., 2022). An iterative self-training method that uses the latest retriever to identify positive and negative passages through answer string matching for updating the training dataset. Machine translation is used for data augmentation.
- **QuiCK** (Ren et al., 2022). A knowledge distillation method that trains a multilingual bi-encoder retriever, learning from a query generator as the teacher. The query generator is also used for generating synthetic multilingual queries to enhance knowledge distillation.
- **DrDecr** (Li et al., 2022). A multilingual ColBERT model that learns from an English ColBERT on parallel queries, sourced from both parallel corpora and human-translated gold queries released by XOR-Retrieve.
- **LAPCA** (Abulkhanov et al., 2023). A pre-training method that takes the first paragraphs of parallel Wikipedia pages as the parallel corpus for cross-lingual pre-training, with augmented data through machine translation.
- **SWIM-X** (Thakur et al., 2023). A method that uses large language models to generate synthetic queries from unlabelled corpus with textual summary generation as an intermediate step. A multilingual dense retrieval model is fine-tuned exclusively on synthetic data.

A.3.2 Multilingual Open Domain Question Answering

- **MT+DPR** (Asai et al., 2021a). This represents the translate-test baseline, in which queries are translated into English and the answers are identified within English passages retrieved by the DPR+MT retriever. The English answer is then translated back to the target language if necessary.
- **ReAtt+MT** (Jiang et al., 2022). This is the English teacher employed in the cross-lingual retrieval pre-training. We use a state-of-the-art machine translation model[11] to translate the queries into English at test time. It always retrieves passages from English Wikipedia and generates answers in English. The generated answer is translated back to the target language.
- **GMT+GS** (Asai et al., 2021a). This pipeline follows the same procedure as **MT+DPR** except that we employ Google Search for passage retrieval and Google Machine Translation services for query and answer translation.
- **Monolingual baseline (BM25)** (Asai et al., 2021a). Instead of using a multilingual DPR or an English DPR model with query translation, this baseline always retrieves the passage from the target language and extracts the answer using a multilingual reader.
- **MT+Mono** (Asai et al., 2021a). This is a combination of the **BM25** and **MT+DPR** baselines, which first does monolingual QA for the target language using the BM25 method and resorts to the **MT+DPR** baseline if no answer is found.
- **Fusion-in-Decoder.** This encompasses a family of multilingual retrieval-augmented generation models, which take the passages returned by a multilingual retriever as inputs to generate the answer in the target language. **CORA** (Asai et al., 2021b), **Sentri** (Ren et al., 2022) and **LAPCA** (Abulkhanov et al., 2023) are included in this family by using the passages returned by their respective retrievers.

[11] https://huggingface.co/facebook/m2m100_418M

B Detailed Zero-shot Evaluation

Cross-lingual Retrieval. Table 4 presents the detailed result comparisons in each of the 20 unseen languages covered by MKQA. Notably, `CLASS-ZS` outperforms other baselines significantly on average and achieves the best results in nearly all languages except for Vietnamese. Comparing the three variants of our method, fine-tuning on supervised English data significantly enhances cross-lingual transfer abilities to every unseen language (i.e., `CLASS-US` vs `CLASS-ZS`). However, fine-tuning `CLASS-ZS` on a limited number of supervised multilingual data with a restricted language set does not lead to improved generalization performance, as indicated by the result comparison in every language between `CLASS-ZS` and `CLASS`. Furthermore, a decrease in performance is also observed in both supervised and zero-shot settings when either multilingual QA pre-training or the entire pre-training procedures are omitted, highlighting the effectiveness of our pre-training approach in enhancing cross-lingual ability.

Multilingual QA. Table 5 presents the detailed multilingual QA results for each of the 20 unseen

Method	Da	De	Es	Fr	He	Hu	It	Km	Ms	Nl	No	Pl	Pt	Sv	Th	Tr	Vi	cn	hk	tw	Avg
Unsupervised																					
CLASS-US	50.5	53.4	53.8	53.9	44.1	49.1	52.6	39.8	55.3	53.3	49.5	52.6	50.4	52.5	54.9	50.9	48.0	48.0	46.3	46.4	50.3
Zero-shot																					
BM25+MT	44.1	43.3	44.9	42.5	36.9	39.3	40.1	31.3	42.5	46.5	43.3	46.5	45.7	49.7	46.5	42.5	43.5	37.5	37.5	36.1	42.0
CLASS-ZS	**59.3**	**58.9**	**59.4**	**59.2**	**50.1**	**54.0**	**58.7**	**46.2**	**59.6**	**60.4**	**58.5**	**57.5**	**58.0**	**59.4**	**58.0**	**55.1**	**54.1**	**52.1**	**51.5**	**51.4**	**56.1**
- MLQA-PT	58.0	57.6	57.7	58.0	47.3	51.8	57.2	44.4	58.0	59.3	57.1	56.1	56.2	57.7	56.4	53.6	52.3	50.6	49.8	49.1	54.4
- Pre-train	50.9	50.5	49.9	50.0	32.5	41.9	49.6	32.9	49.9	52.3	50.2	46.6	49.3	51.5	44.2	44.7	41.3	37.8	37.7	37.1	45.0
Supervised																					
CORA	44.5	44.6	45.3	44.8	27.3	39.1	44.2	22.2	44.3	47.3	48.3	44.8	40.8	43.6	45.0	34.8	33.9	33.5	41.5	41.0	41.1
Sentri	57.6	56.5	55.9	55.1	47.9	51.8	54.3	**43.9**	56.0	56.3	56.5	55.8	54.8	56.9	55.3	53.0	**54.4**	50.2	**50.7**	49.4	53.3
QuiCK	**58.3**	56.4	55.2	55.5	44.7	52.4	52.3	42.0	56.9	57.5	**57.0**	54.9	54.7	**58.0**	55.7	53.9	54.9	50.4	49.3	48.9	53.4
CLASS	57.4	**57.5**	**58.0**	**57.8**	**48.5**	**52.5**	**57.1**	43.4	**58.2**	58.4	56.7	**56.0**	**56.4**	57.6	**57.2**	**54.2**	52.5	51.3	49.9	**50.2**	**54.6**
- MLQA-PT	56.9	57.3	57.2	57.0	47.3	51.8	56.2	42.9	57.6	**58.7**	56.0	55.3	55.5	56.8	56.1	53.3	51.5	**51.4**	49.9	49.4	53.9
- Pre-train	56.5	55.3	55.9	55.1	44.8	50.8	55.0	41.3	56.4	57.4	55.8	53.3	54.8	56.5	53.7	51.9	49.6	47.3	46.4	45.8	52.2

Table 4: Zero-shot cross-lingual retrieval results (R@2kt) on the MKQA dataset. "cn": "Zh-cn" (Chinese, simplified). "hk": "Zh-hk" (Chinese, Hong Kong). "tw": "Zh-tw" (Chinese, traditional).

Method	Da	De	Es	Fr	He	Hu	It	Km	Ms	Nl	No	Pl	Pt	Sv	Th	Tr	Vi	cn	hk	tw	Avg
Unsupervised																					
CLASS-US	24.9	27.4	29.1	27.1	12.9	21.7	25.2	9.3	26.3	27.0	25.0	23.7	22.4	26.0	13.2	22.8	17.5	7.3	8.9	6.3	20.2
Zero-shot																					
ReAtt+MT	22.4	23.9	21.6	23.5	**24.2**	6.3	13.7	3.2	12.7	22.1	21.5	11.2	18.6	17.3	7.2	6.3	24.0	**10.8**	4.7	4.0	15.0
MT+DPR	26.2	25.9	28.4	21.9	8.9	15.7	25.1	1.2	12.6	28.3	18.3	24.6	24.7	19.7	6.9	18.2	15.1	3.3	3.8	3.8	16.5
CLASS-ZS	**37.6**	**38.5**	**40.2**	**37.6**	17.0	**29.1**	**36.2**	**16.2**	**36.9**	**38.6**	**37.4**	**34.4**	**33.6**	**38.6**	**18.9**	**30.9**	**29.6**	8.7	**13.8**	**8.5**	**29.1**
Supervised																					
MT+Mono	19.3	21.6	21.3	21.9	8.9	16.5	20.9	1.2	12.6	21.5	17.4	24.6	19.9	20.0	8.3	16.6	15.1	4.9	3.8	5.1	14.8
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	**33.4**	**35.4**	**37.5**	**35.7**	12.3	**27.7**	**35.3**	**10.2**	**34.6**	**36.1**	**34.3**	**31.9**	**32.8**	**33.3**	**17.6**	**29.3**	**25.1**	**8.6**	**10.2**	**7.4**	**26.4**

Table 5: Zero-shot multilingual question answering results (F1) on the MKQA dataset. "cn": "Zh-cn" (Chinese, simplified). "hk": "Zh-hk" (Chinese, Hong Kong). "tw": "Zh-tw" (Chinese, traditional).

Model	F1	EM	BLEU
CLASS	**30.4**	21.0	**20.6**
CLASS w/o MLQA-PT	30.2	**21.4**	20.3
CLASS w/o Pre-train	29.7	20.9	19.8

Table 6: Multilingual QA results on queries requiring cross-lingual evidence retrieval.

languages covered by MKQA. We observe similar patterns where CLASS-US surpasses a range of machine-translation-based methods and CLASS-ZS outperforms the supervised CORA by a significant margin. Further fine-tuning CLASS-ZS on a limited number of supervised multilingual data with a restricted language set hampers its generalizability, with a decline in performance across all examined languages.

C Error Analysis

We add additional error analysis regarding the issue identified in multilingual QA (i.e., XOR-Full).

C.1 Quantitative Analysis

Our focus is on analysing the behaviour of our model when handling cross-lingual queries in XOR-Full. These queries require answers based on English evidence (Asai et al., 2021a). Initially, we analyse the retrieval accuracy of our model by assessing whether the top-n retrieved tokens contain the answer string in English or the target language.

As shown in Table 7, our pre-training method shows significant improvements in finding correct English evidence for those queries requiring cross-lingual evidence retrieval (e.g., 50.4% -> 70.8%) while maintaining competitive performance (Table 7(c)) in finding in-language (i.e., the question language) evidence if there exists. Nevertheless, we have observed that these advancements do not translate into enhancements in the subsequent QA task, wherein the model is supposed to produce an answer in the same language as the question with English supporting documents. Table 6 shows that our complete CLASS model fails to achieve additional benefits in QA tasks despite its outstanding performance in retrieving cross-lingual evidence.

To gain deeper insights into the behaviour of our model, we specifically analyse its QA performance whenever the top-n retrieved evidence contains the gold answer in either English or the target language. As indicated in Table 8, our model demonstrates reasonable performance only when the correct answer string is presented in the target language. However, it often fails to generate the correct an-

Model	R@2kt	R@5kt	R@10kt
CLASS	**61.0**	**70.6**	**75.6**
CLASS w/o MLQA-PT	59.0	68.8	74.6
CLASS w/o Pre-train	50.6	59.3	65.4

(a) English or target language answer is in top-n retrieved tokens.

Model	R@2kt	R@5kt	R@10kt
CLASS	**50.4**	**63.1**	**70.8**
CLASS w/o MLQA-PT	46.3	59.2	67.8
CLASS w/o Pre-train	32.7	42.1	50.4

(b) Only English answer is in top-n retrieved tokens.

Model	R@2kt	R@5kt	R@10kt
CLASS	41.8	47.3	50.8
CLASS w/o MLQA-PT	**42.7**	**47.9**	**51.2**
CLASS w/o Pre-train	41.0	46.9	50.4

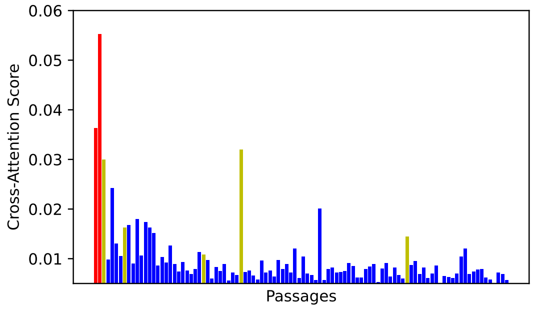
(c) Only Target language answer is in top-n retrieved tokens.

Table 7: Retrieval accuracy of queries requiring answers based on English evidence.

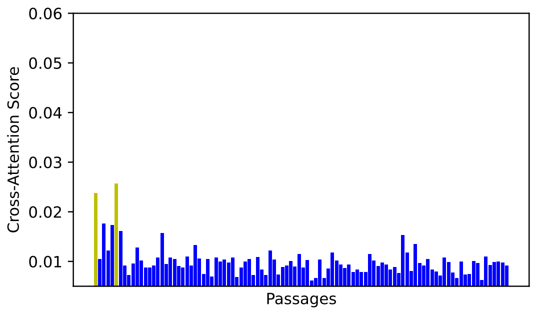
	Contain English Ans	No English Ans
Contain Target Ans	F1: 41.0/EM: 30.6/BLEU: 33.2	F1: 37.7/EM: 28.3/BLEU: 32.5
No Target Ans	F1: 13.5/EM: 2.1/BLEU: 12.9	F1: 10.1/EM: 1.0/BLEU: 6.8

Table 8: Multilingual QA results on queries requiring cross-lingual evidence retrieval, grouped by whether the gold-standard answer string in English or the target language appears within the top-n retrieved tokens.

swer when the gold standard answer is provided solely in English, despite our model being able to include the correct English answer in its top-10k retrieved tokens 71% of the time. This indicates a deficiency in our model's ability to identify correct clues for QA among cross-lingual evidence. An example is shown in Figure 7. In cases where the top 100 retrieved passages contain answer strings in the target language, our model tends to assign significantly higher scores to passages containing these target language answer strings. By contrast, when only English answer strings are present, the distribution of cross-attention scores across all retrieved passages becomes more uniform, leading to a general narrowing of the gap between positively relevant passages and irrelevant ones.



(a) Answer strings in target language or English are in top-100 retrieved passage



(b) Only answer strings in English are in top-100 retrieved passages

Figure 7: Cross-Attention score to each of top-100 retrieved passages. Passages that contain the answer string in target languages or English are denoted with red and yellow bars, respectively.

C.2 Case Study

As shown in Table 9, our model successfully retrieves the appropriate supporting document as its top-1 retrieval. However, it encounters challenges in generating Telugu answers, whereas it performs accurately in English. This highlights our model's inability to translate English evidence into answers in the target language, necessitating further efforts to enhance the model's capabilities in cross-lingual evidence reasoning and answer generation.

D More Analysis

Performance Evolution during Pre-training. Figure 8 illustrates the trajectory of the performance on the XOR-Retrieve cross-lingual retrieval task. As shown in the Figure, the use of code-switching consistently yields inferior results compared to CLASS and the variant using machine translation. After training on around 45 billion tokens, CLASS consistently outperforms MT, matching the performance of CS and MT with only 30% and 50% computation costs. This demonstrates greater training efficiency. The performance continues to improve over the next 50% of the training tokens, implying that the scalability of pre-training data remains beneficial as training progresses.

Few-Shot Cross-lingual Retrieval. We consider a few-shot learning task with varying numbers of labelled training examples. Figure 9 shows that CLASS is consistently better than the other two variants, although the performance gap diminishes as more labelled data becomes available. Notably, as

- Query: ఆక్సిజన్ చిత్ర కథానాయకుడు ఎవరు? ("en": Who is the protagonist of the movie 'Oxygen'?)

- Gold Ans: [గోపీచంద్, అను ఇమ్మాన్యుయేల్] ("en": [Gopichand, Anu Emmanuel])

- TOP-1 Retrieved Passage: Oxygen is a 2017 Indian Telugu-language action film produced by S. Aishwarya on Sri Sai Raam Creations banner, presented by A. M. Rathnam and directed by A. M. Jyothi Krishna. **Starring Gopichand, Raashi Khanna, Anu Emmanuel in the lead roles** while Jagapati Babu in crucial supporting role and music composed by Yuvan Shankar Raja

- Telugu Prediction: బ్రహ్మానందం ("en": Brahmananda)
- English Prediction: Gopichand

Table 9: An example of our model in finding correct evidence while failing to generate the right answer in the target language.

High Level Answer Category	Named Entity Types	Most appropriate wh_word
PERSON/NORP/ORG	PERSON, NORP, ORG	Who
PLACE	GPE, LOC, FAC	Where
THING	PRODUCT, EVENT, WORKOFART, LAW, LANGUAGE	What
TEMPORAL	TIME, DATE	When
NUMERIC	PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL	How much/How many

Table 10: The heuristics rules for choosing the most appropriate question word based on named entity types (taken from Lewis et al. (2019)).

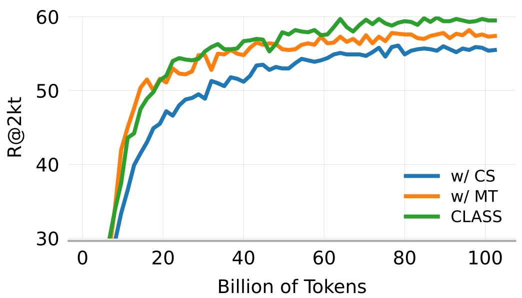


Figure 8: Performance evolution in stage-1 pre-training.

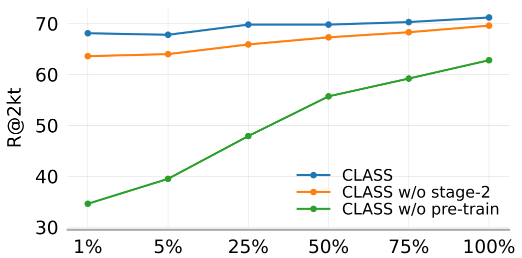


Figure 9: Scaling training data on cross-lingual retrieval.

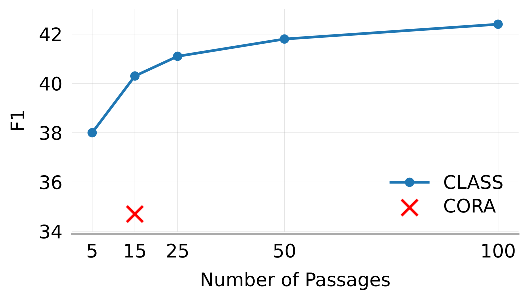


Figure 10: Effects of employing different numbers of retrieved passages for QA during inference time.

illustrated in Figure 9, the introduction of stage-2 pre-training results in a 75% reduction in the required amount of labelled data. Furthermore, employing pre-training of both stages eliminates the need for any labelled data, in contrast to the approach that solely relies on supervised data for training (i.e., CLASS w/o pre-train).

Effects of Number of Retrieved Passages. Figure 10 reports the performance concerning the number of retrieved passages for QA during inference. We observe the performance improves consistently as the number of retrieved passages increases. CLASS significantly outperforms CORA when using only top-5 retrieved passages, showcasing superior inference efficiency.

E Query Transformation Examples

Figure 11 showcases examples illustrating the generation of meta-examples through prompting ChatGPT. Prompts E.1, E.2, E.3, E.4, E.5, E.6, and E.7 provide detailed illustrations of prompting a much smaller large language model, LLaMA-2-7B, to perform query transformation using In-Context Learning, which incorporates meta-examples in the target language L from $\mathbb{K}$ into the prompt to guide the model's behaviour. The choice of the *question word* is determined based on the detected entity type of the answer and the heuristic rules outlined in Table 10.

Finnish Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence "Strapping Young Lad (lyh. SYL) oli Devin Townsendin vuonna 1994 perustama kanadalainen metalliyhtye." into a natural question whose question word is "Milloin" and answer is "1994". Please respond in the format: "The transformed question is: Milloin Devin Townsend perusti kanadalaisen metalliyhtyeen Strapping Young Lad (lyh. SYL)? "

Russian Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence " В 215 году Цао Цао атаковал Чжан Лу и разгромил его в битве в проходе Янпингуань. " into a natural question whose question word is " Кто " and answer is " Чжан Лу ". Please respond in the format: "The transformed question is: Кто был атакован Цао Цао и разгромлен в битве в проходе Янпингуань в 215 году? "

Japanese Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence "熊野那智神社（くまのなちじんじゃ）は、宮城県名取市にある神社である。" into a natural question whose question word is "どこ" and answer is "宮城県". Please respond in the format: "The transformed question is: 熊野那智神社はどこにある神社ですか？ "

Korean Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence "19세기 후반에 아일랜드에는 독립과 토지개혁을 요구하는 운동이 크게 확산되었다." into a natural question whose question word is "어디" and answer is "아일랜드". Please respond in the format: "The transformed question is: 19세기 후반에 독립과 토지개혁을 요구하는 운동이 크게 확산된 나라는 어디입니까? "

Arabic Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence " ولكن في مجال التعليم العالي تفتقر المسلمين، ونظرا لأفراد أسرهم وأقاربهم أخذها عن وظائف في دول الخليج وجنوب شرق آسيا (أساسا وسنغافورة، وألمانيا وماليزيا وبروناي في جنوب آسيا) ذفسها في سن مبكرة. " into a natural question whose question word is " أين " and answer is " جنوب شرق آسيا ". Please respond in the format: "The transformed question is:

أين يأخذ أفراد أسر المسلمين وأقاربهم وظائف غالباً، مما يؤدي إلى فقدان الاهتمام بالتعليم العالي؟ "

Bengali Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence "ভারতে হাজার হাজার মানুষ অনাহারে মারা যায়, কিন্তু ধর্মপ্রচারকরা তাদের প্রতি উদাসীন।" into a natural question whose question word is "কোথায়" and answer is "ভারত". Please respond in the format: "The transformed question is: কোথায় হাজার হাজার মানুষ অনাহারে মারা যায় এবং ধর্মপ্রচারকরা তাদের প্রতি উদাসীন থাকে? "

Telugu Prompt

You are an AI model that rewrites sentences into questions, using a given question word and answer.
Rewrite this sentence " వీటిలో ప్రసిద్ధి చెందిన శ్రీ వాసవి కన్యకా పరమేశ్వరీ దేవి ఆలయం ఆంధ్ర ప్రదేశ్ రాష్ట్రంలో పశ్చిమ గోదావరి జిల్లాలో పెనుగొండ అనే పట్టణంలో ఉంది. " into a natural question whose question word is " ఎవరు " and answer is " పెనుగొండ ". Please respond in the format: "The transformed question is: ఆంధ్ర ప్రదేశ్ రాష్ట్రం పశ్చిమ గోదావరి జిల్లాలోని ప్రసిద్ధి చెందిన శ్రీ వాసవి కన్యకా పరమేశ్వరీ దేవి ఆలయం ఉన్న పట్టణం ఎవరు? "

Figure 11: Meta-examples obtained by prompting ChatGPT are shown for each language coverd by XOR-TYDI QA. Lightblue texts indicate the transformed questions.

Prompt E.1: Finnish Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: Toisaalta hän oli taiteiden suosija ja hänen valtakaudellaan Preussi sai haltuunsa suuren osan Puola-Liettuasta Puolan jaoissa vuosina 1793 ja 1795.
Question word: Missä
Answer: Preussi
Transformed Question: Missä maassa taiteiden suosija hallitsi ja missä valtakunnassa saatiin haltuunsa suuri osa Puola-Liettuasta Puolan jaoissa vuosina 1793 ja 1795?

Sentence: Hän pelasi urallaan myös Ruotsissa ja Slovakiassa.
Question word: Missä
Answer: Slovakia
Transformed Question: Missä maassa hän pelasi urallaan Ruotsin lisäksi?

Sentence: Barokin jälkeen concerto grossoja ovat säveltäneet muun muassa Heitor Villa-Lobos, Bohuslav Martinů, Alfred Schnittke ja Philip Glass.
Question word: Kuka
Answer: Bohuslav Martinů
Transformed Question: Kuka säveltäjistä Heitor Villa-Lobosin, Alfred Schnittken ja Philip Glassin ohella on säveltänyt concerto grossoja barokin jälkeen?

Sentence: Hänen ajatteluunsa vaikuttivat muun muassa buddhalaiset ja taolaiset ideat, joihin hän tutustui Aasian matkoillaan, Mahatma Gandhin väkivallattomuusliike, sekä hänen katolinen uskontonsa.
Question word: Kuka
Answer: Mahatma Gandhi
Transformed Question: Kuka vaikutti hänen ajatteluunsa, mahtimaailmaan ja katoliseen uskontonsa?

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: On the other hand, he/she was a fan of the arts and during his/her reign, Prussia took over a large part of Poland-Lithuania in the partitions of Poland in 1793 and 1795.
Question word: Where
Answer: Prussia
Transformed Question: In which country did the lover of the arts rule and in which kingdom was a large part of Poland-Lithuania taken over during the partitions of Poland in 1793 and 1795?

Sentence: He/She also played in Sweden and Slovakia during her career.
Question word: Where
Answer: Slovakia
Transformed Question: In which country did he/she play in his/her career besides Sweden?

Sentence: After the Baroque, concerto grossos have been composed by, among others, Heitor Villa-Lobos, Bohuslav Martinů, Alfred Schnittke and Philip Glass.
Question word: Kuka
Answer: Bohuslav Martinů
Transformed Question: Besides Heitor Villa-Lobos, Alfred Schnittke and Philip Glass, which of the composers has composed concerto grossos after the Baroque?

Sentence: His/Her thinking was influenced, among other things, by Buddhist and Taoist ideas, which he/she got to know during his/her travels in Asia, Mahatma Gandhi's non-violence movement, and his/her Catholic religion.
Question word: Who
Answer: Mahatma Gandhi
Transformed Question: Who influenced his/her thinking, the world of power and his/her Catholic religion?

Prompt E.2: Russian Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: Корабли проекта выполняли контроль за учениями ВМС стран НАТО в Норвежском и Средиземном морях, следили за корабельными и авианосными группами флотов США и Великобритании.
Question word: Кто
Answer: НАТО
Transformed Question: Кто выполнял контроль за учениями ВМС в Норвежском и Средиземном морях и следил за корабельными и авианосными группами флотов США и Великобритании?

Sentence: 1 апреля 1768 года Доверню назначают пенсию Королевской академии музыки в размере 1000 ливров как автору музыки.
Question word: Кто
Answer: Королевской академии музыки
Transformed Question: Кто 1 апреля 1768 года назначил пенсию в размере 1000 ливров Доверню как автору музыки?

Sentence: Софи́я Шарло́тта Авгу́ста (22 февраля 1847, Мюнхен — 4 мая 1897, Париж) — принцесса Баварская, герцогиня Баварская, позднее герцогиня Алансонская и Орлеанская.
Question word: Где
Answer: Мюнхен
Transformed Question: Где родилась София Шарлотта Августа, принцесса Баварская?

Sentence: В первой половине XIX века паровозы в Россию, в основном, ввозились из-за рубежа.
Question word: Когда
Answer: XIX век
Transformed Question: Когда паровозы в Россию, в основном, ввозились из-за рубежа?

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: The project's ships monitored NATO naval exercises in the Norwegian and Mediterranean Seas and monitored ship and aircraft carrier groups of the US and British navies.
Question word: Who
Answer: NATO
Transformed Question: Who monitored naval exercises in the Norwegian and Mediterranean seas and monitored ship and aircraft carrier groups of the US and British fleets?

Sentence: On April 1, 1768, Dauvergne was awarded a pension from the Royal Academy of Music in the amount of 1000 livres as the author of music.
Question word: Who
Answer: Royal Academy of Music
Transformed Question: Who, on April 1, 1768, awarded a pension of 1000 livres to Dovergne as the author of music?

Sentence: Sophia Charlotte Auguste (22 February 1847, Munich - 4 May 1897, Paris) - Princess of Bavaria, Duchess of Bavaria, later Duchess of Alençon and Orléans.
Question word: Where
Answer: Munich
Transformed Question: Where was Sophia Charlotte Augusta, Princess of Bavaria born?

Sentence: In the first half of the 19th century, steam locomotives were mainly imported to Russia from abroad.
Question word: When
Answer: 19th century
Transformed Question: When were steam locomotives mainly imported into Russia from abroad?

Prompt E.3: Japanese Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: 2月、竇憲は左校尉の耿夔を遣わし、金微山において北匈奴の単于を包囲しこれを大いに破り、単于の母の閼氏を捕虜とした。
Question word: 誰
Answer: 匈奴
Transformed Question: 2月に金微山で竇憲の遣わした左校尉の耿夔が包囲し大いに破ったのは誰の単于ですか？

Sentence: この町を法人化する法はリチャード・キャズウェルが提出し、キャズウェルはここを本拠地とし、後の1776年から1780年までノースカロライナ州の初代知事となった。
Question word: どこ
Answer: ノースカロライナ州
Transformed Question: リチャード・キャズウェルが初代知事となったのはどこですか？

Sentence: これより以前、司空張華は司馬倫に疎まれて誅殺されていた。
Question word: 誰
Answer: 張華
Transformed Question: 誰がこれより以前に司馬倫に疎まれて誅殺されていたのですか？

Sentence: 魯迅はこの無支祁が孫悟空の先祖・源流ではないかと推測した。
Question word: 誰
Answer: 魯迅
Transformed Question: 誰はこの無支祁が孫悟空の先祖・源流ではないかと推測したのでしょうか？

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: In February, Dou Xian sent Zuo's lieutenant, Geng Kui, to besiege and defeat the Northern Xiongnu Danyu at Jinweishan, and took Danyu's mother, the Yan family, prisoner.
Question word: Who
Answer: Xiongnu
Transformed Question: In February, in Jinweishan, which was the land of Danyu that was besieged and severely defeated by Geng Ku, the commander of the left school sent by Dou Xian?

Sentence: The act to incorporate the town was introduced by Richard Caswell, who made it his home and later became North Carolina's first governor from 1776 to 1780.
Question word: Where
Answer: North Carolina
Transformed Question: Where did Richard Caswell become the first governor?

Sentence: Before this, Zhang Hua was shunned by Sima Lun and killed.
Question word: Who
Answer: Zhang Hua
Transformed Question: Who had been shunned and killed by Sima Lun before this?

Sentence: Lu Xun surmised that this Mujiqi was the ancestor and origin of Sun Wukong.
Question word: Who
Answer: Lu Xun
Transformed Question: Who could have guessed that Mujiqi was the ancestor/origin of Son Goku?

Prompt E.4: Korean Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: 전투에서 승리한 뒤, 오버워치는 10년간 계속해서 평화를 지켰으나 내분으로 인해 해산되었다.
Question word: 누구
Answer: 오버워치
Transformed Question: 누구가 전투에서 승리한 뒤 10년 동안 평화를 지키다가 내분으로 인해 해산되었나요?

Sentence: 그가 구단을 떠난 지 10년이 되는 2013년 4월, 스포르팅 리스본은 호날두를 100,000번째 회원으로 등록해 경의를 표했다.
Question word: 누구
Answer: 스포르팅 리스본
Transformed Question: 누가 2013년 4월 그가 구단을 떠난 지 10년이 되는 해에 호날두를 100,000번째 회원으로 등록해 경의를 표했나요?

Sentence: 19세기 후반에 아일랜드에는 독립과 토지개혁을 요구하는 운동이 크게 확산되었다.
Question word: 어디
Answer: 아일랜드
Transformed Question: 19세기 후반에 독립과 토지개혁을 요구하는 운동이 크게 확산된 나라는 어디입니까?

Sentence: 산탄젤로 다리 () 또는 하드리아누스의 다리는 로마에 있는 다리 가운데 하나이다.
Question word: 어디
Answer: 로마
Transformed Question: 산탄젤로 다리가 있는 곳은 어디인가?

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: After winning the battle, Overwatch continued to maintain peace for 10 years, but was disbanded due to internal strife.
Question word: Who
Answer: Overwatch
Transformed Question: Who won the battle, kept the peace for ten years, and then disbanded due to infighting?

Sentence: In April 2013, 10 years after he left the club, Sporting Lisbon paid tribute to Ronaldo by registering him as their 100,000th member.
Question word: Who
Answer: Sporting Lisbon
Transformed Question: Who paid tribute to Ronaldo by registering him as their 100,000th member in April 2013, marking 10 years since he left the club?

Sentence: In the late 19th century, movements calling for independence and land reform spread widely in Ireland.
Question word: Where
Answer: Ireland
Transformed Question: In which country did the movement calling for independence and land reform spread significantly in the late 19th century?

Sentence: Ponte Sant'Angelo () or Hadrian's Bridge is one of the bridges in Rome.
Question word: Where
Answer: Rome
Transformed Question: Where is the Ponte Sant'Angelo?

Prompt E.5: Arabic Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: ولد كلاي في مقاطعة هانوفر بولاية فرجينيا في عام ١٧٧٧، ولكنه انتقل إلى ليكسينغتون، كنتاكي في عام ١٧٩٧.
Question word: أين
Answer: كنتاكي
Transformed Question: أين انتقل ولد كلاي بعد ميلاده في مقاطعة هانوفر، ولاية فرجينيا في عام ١٧٩٧؟

Sentence: وكان نظام شركة سي بي إس هو الأكثر تقدما وفاز بالمنافسة في كل مرة.
Question word: من
Answer: شركة سي بي إس
Transformed Question: من كان لديه النظام الأكثر تقدما وفاز بالمنافسة في كل مرة؟

Sentence: كما الخطوط الجوية البريطانية تشغل أيضا صالة ذادي التنفيذيين بين البوابات ٢١ و ٢٣.
Question word: من
Answer: الخطوط الجوية البريطانية
Transformed Question: من يشغل صالة ذادي التنفيذيين بين البوابات ٢١ و ٢٣؟

Sentence: وقد افتتح حديثاً في قصر العظم قاعتين تضمان بانوراما لحرب تشرين التحريرية عام ١٩٧٣، وبانوراما لحرب تموز ٢٠٠٦.
Question word: أين
Answer: قصر العظم
Transformed Question: أين بانوراما لحرب تشرين التحريرية عام ١٩٧٣؟

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: Clay was born in Hanover County, Virginia in 1777, but moved to Lexington, Kentucky in 1797.
Question word: Where
Answer: Kentucky
Transformed Question: Where did Clay move after his birth in Hanover County, Virginia in 1797?

Sentence: The CPS system was the most advanced and won the competition.
Question word: Who
Answer: CPS system
Transformed Question: Who had the most advanced system and won the competition?

Sentence: British Airways also operates the Teen Club lounge between gates B21 and B23.
Question word: Who
Answer: British Airways
Transformed Question: Who operates the Executive Club lounge between gates B21 and B23?

Sentence: Two halls were recently opened in Al-Azm Palace containing a panorama of the October Liberation War of 1973, and a panorama of the July War of 2006.
Question word: Where
Answer: Al-Azm Palace
Transformed Question: Where is the panorama of the October Liberation War of 1973?

Prompt E.6: Bengali Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: যাত্রাপথে সবার আগে দ্রৌপদী প্রাণ হারান।
Question word: কে
Answer: দ্রৌপদী
Transformed Question: যাত্রাপথে সবার আগে কে প্রাণ হারান?

Sentence: অপরদিকে কাতাররে রাজধানী দোহাতে রাশিয়ার একটি স্থায়ী দূতাবাস রয়েছে।
Question word: কোথায়
Answer: কাতার
Transformed Question: রাশিয়ার স্থায়ী দূতাবাসটি কোথায় অবস্থিত?

Sentence: ভারতে হাজার হাজার মানুষ অনাহারে মারা যায়, কিন্তু ধর্মপ্রচারকরা তাদের প্রতি উদাসীন।
Question word: কোথায়
Answer: ভারত
Transformed Question: কোথায় হাজার হাজার মানুষ অনাহারে মারা যায় এবং ধর্মপ্রচারকরা তাদের প্রতি উদাসীন থাকে?

Sentence: এটি ওয়াশিংটন -এর সিয়াটল-এ অবস্থিত খোলা জায়গায় একটি মাছের বাজার।
Question word: কোথায়
Answer: সিয়াটল
Transformed Question: এটি ওয়াশিংটন কোথায় খোলা জায়গায় একটি মাছের বাজার?

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: Draupadi was the first to die on the journey.
Question word: Who
Answer: Draupadi
Transformed Question: Who died first on the journey?

Sentence: In addition, Russia has a permanent embassy in Doha, the capital of Qatar.
Question word: Where
Answer: Qatar
Transformed Question: Where is the permanent embassy of Russia located?

Sentence: Thousands of people die of starvation in India, but missionaries are indifferent to them.
Question word: Where
Answer: India
Transformed Question: Where are thousands of people dying of starvation and the missionaries are indifferent to them?

Sentence: It is an open-air fish market located in Seattle, Washington.
Question word: Where
Answer: Seattle
Transformed Question: Where is an open air fish market in Washington?

Prompt E.7: Telugu Example & Translation

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: ఈ గ్రామములో వరి, చెరకు, మామిడి, వేరుశనగ, కూరగాయలు మొదలగునవి ప్రధాన పంటలు.
Question word: ఎవరు
Answer: మామిడి
Transformed Question: ఈ గ్రామములో ప్రధాన పంటలలో ఎవరు ఒకటి?

Sentence: ఈ సమయంలో ప్రపంచంలోని ఉద్గారాల గణనీయమైన పెరుగుదలకు చైనా కారణమైంది.
Question word: ఎక్కడ
Answer: చైనా
Transformed Question: ఈ సమయంలో ప్రపంచంలో ఉద్గారాల గణనీయమైన పెరుగుదలకు ఎక్కడ కారణమైంది?

Sentence: వీటిలో ప్రసిద్ధి చెందిన శ్రీ వాసవి కన్యకా పరమేశ్వరీ దేవి ఆలయం ఆంధ్ర ప్రదేశ్ రాష్ట్రంలో పశ్చిమ గోదావరి జిల్లాలో పెనుగొండ అనే పట్టణంలో ఉంది.
Question word: ఎవరు
Answer: పెనుగొండ
Transformed Question: ఆంధ్ర ప్రదేశ్ రాష్ట్రం పశ్చిమ గోదావరి జిల్లాలోని ప్రసిద్ధి చెందిన శ్రీ వాసవి కన్యకా పరమేశ్వరీ దేవి ఆలయం ఉన్న పట్టణం ఎవరు?

Sentence: సాత్యకిని కృతవర్మ అడ్డుకొనడం చూసిన ద్రోణుడు ధర్మరాజు వైపు వెళ్ళాడు.
Question word: ఎవరు
Answer: ధర్మరాజు
Transformed Question: సాత్యకిని కృతవర్మ అడ్డుకొనడం చూసిన ద్రోణుడు ఎవరు వైపు వెళ్ళాడు?

Rewrite sentences into short and precise questions, using given question words and answers:

Sentence: The main crops in this village are rice, sugarcane, mango, groundnut, vegetables etc.
Question word: Who
Answer: mango
Transformed Question: Which is one of the main crops in this village?

Sentence: China accounted for a significant increase in world emissions during this period.
Question word: Where
Answer: China
Transformed Question: Where in the world has caused the significant increase in emissions during this time?

Sentence: Among these, the famous Sri Vasavi Kanyaka Parameshwari Devi Temple is located in the town of Penugonda in the West Godavari district of the state of Andhra Pradesh.
Question word: Who
Answer: Penugonda
Transformed Question: Which town in West Godavari district of Andhra Pradesh state has the famous Sri Vasavi Kanyaka Parameshwari Devi temple?

Sentence: Seeing Satyaki being stopped by Kritavarma, Drona went towards Dharmaraja.
Question word: Who
Answer: Dharmaraja
Transformed Question: To whom did Drona go when he saw Kritavarma stopping Satyaki?