From Classification to Generation: Insights into Crosslingual Retrieval Augmented ICL

Xiaoqian Li[1,3] **Ercong Nie**[1,2] **Sheng Liang**[†1,2]
[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Germany
[3] Academy of Cyber
Xiaoqian.Li@campus.lmu.de
{nie, shengliang}@cis.lmu.de

Abstract

The remarkable ability of Large Language Models (LLMs) to understand and follow instructions has sometimes been limited by their in-context learning (ICL) performance in low-resource languages. To address this, we introduce a novel approach that leverages cross-lingual retrieval-augmented in-context learning (CREA-ICL). By extracting semantically similar prompts from high-resource languages, we aim to improve the zero-shot performance of multilingual pre-trained language models (MPLMs) across diverse tasks. Though our approach yields steady improvements in classification tasks, it faces challenges in generation tasks. Our evaluation offers insights into the performance dynamics of retrieval-augmented in-context learning across both classification and generation domains.

1 Introduction

In recent years, the field of Natural Language Processing (NLP) has undergone transformative advances, driven primarily by deep transformer techniques [Vaswani et al., 2017, Devlin et al., 2019, Radford et al., 2019]. The emergence of Large Language Models (LLMs) such as GPT-3 [Brown et al., 2020a] and GPT-4 [OpenAI, 2023] has increased the ability of these models to understand and execute instructions, marking a significant milestone in the field of in-context learning (ICL). These models exhibit exceptional skills in tasks like text classification and generation. They cater to a wide array of applications across various languages, with the primary beneficiaries being languages like English [Conneau et al., 2020, Raffel et al., 2020, Radford et al., 2019]. Benchmarks like XTREME [Hu et al., 2020] and BUFFET [Asai et al., 2023] further validate their capabilities. However, several low-resource languages, with Bangla as a notable example, face inherent challenges, primarily due to the limited availability of pretraining corpora [Artetxe and Schwenk, 2019, Hangya et al., 2022, Sazzed, 2020].

Despite its vast native speaker base, Bangla's representation in NLP remains constrained. This limitation is attributed to its linguistic complexity, the dearth of labeled datasets, and issues like data duplication [Das and Bandyopadhyay, 2010, Das and Gambäck, 2014]. While traditional machine learning techniques have made progress in Bangla NLP tasks, the potential of harnessing the capabilities of the latest LLMs for ICL remains to be fully tapped [Bhowmick and Jana, 2021, Wahid et al., 2019, Hoq et al., 2021].

The ongoing shift in ICL with LLMs highlights the value of retrieval augmentation, where sourcing instructions with semantic information has become pivotal [Shi et al., 2023]. In the multilingual ICL domain, methodologies like MEGA [Ahuja et al., 2023] often narrow their focus to task-centric instructions, lacking in-depth semantic insights due to their reliance on random prompt selection. Complementing these advances, the work on CORAC [Asai et al., 2021] represents a significant leap

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

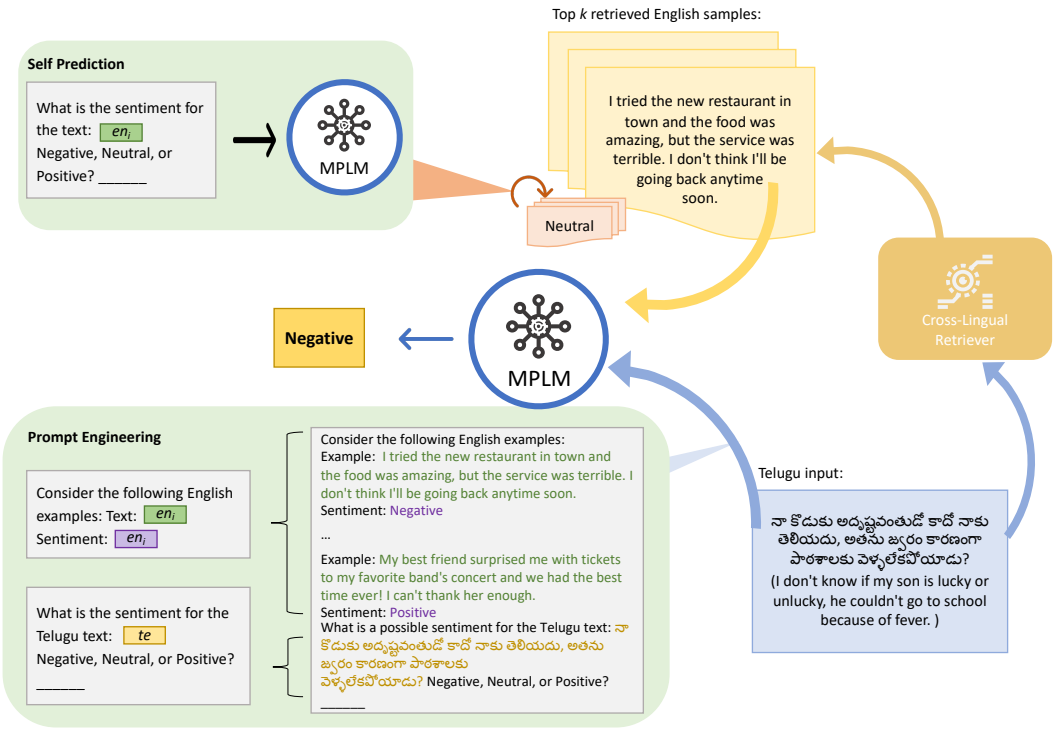

Figure 1: Detailed overview of the CREA-ICL pipeline for LRLs: (a) An LRL input is used as a query for the cross-lingual retriever, which then retrieves the most semantically similar HRL sample from the HRL corpus. The associated label is either taken directly from the corpus (labeled setting) or determined by self-prediction (unlabeled setting). (b) Next, this HRL sample, its label, and the original input are combined to create a retrieval-augmented input for MPLM to make prediction.

forward in multilingual QA models, demonstrating a many-to-many approach that avoids language-specific data and retrieval modules, which is particularly beneficial for low-resource languages.In contrast, strategies like PARC [Nie et al., 2023] propose a more comprehensive methodology by obtaining semantically aligned instructions from high-resource languages.

Building on these methods, our work introduces novel perspectives and aims to bridge gaps. While MEGA provides task-centric instructions, we integrate deeper semantic understanding. We embrace a cross-lingual approach akin to PARC, as depicted in Figure 1. In contrast to PARC's focus on masked language models like mBERT and XLMR, we explore the potential of larger, decoder-only multilingual pretrained language models (MPLMs) - BLOOM and BLOOMZ. Our focus is on addressing both classification and generation tasks in a cohesive generative style, emphasizing instruction execution [Muennighoff et al., 2023, Scao et al., 2022].

This paper explores the application of cross-lingual retrieval-augmented ICL (CREA-ICL) to a specific low-resourced case, the Bangla language-covering text classification and generation tasks. We prioritize the effective execution of instructions. Our main contributions are:

- A comprehensive evaluation of cross-lingual retrieval augmented ICL, highlighting consistent improvements over MPLMs' zero-shot performance on Bangla classification tasks.
- An in-depth analysis revealing the challenges in Bangla generation task, providing insights into the performance dynamics in both the classification and generation domains.
- A pioneering exploration to adapt PARC for generative models, BLOOM and BLOOMZ, providing insights for a unified pipeline of CREA-ICL.

2 Related Work

Retrieval Augmented Prompt External knowledge extracted by information retrieval is often leveraged to solve NLP tasks. Two types of representations have been used for retrieval: (1) sparse bag-of-words representations [Chen et al., 2017, Wang et al., 2018], and (2) dense representation learned by neural networks [Qu et al., 2020]. Dense representations come either from contextual token embeddings [May et al., 2019, Zhang et al., 2020] or from sentence encoders [Conneau et al., 2017, Cer et al., 2018]. Reimers and Gurevych [2019a] propose sentence transformers to create semantically meaningful sentence embeddings by applying siamese and triplet network structures to transformer-based pretrained language models. By using knowledge distillation, sentence transformers can be expanded to support various languages as multilingual sentence transformers [Reimers and Gurevych, 2020], allowing for cross-lingual retrieval.

Brown et al. [2020b] has shown that LLMs like GPT-3 can acquire task-solving abilities by incorporating input-output pairs as context. The in-context learning approach involves concatenating input with randomly selected examples from the training dataset, which is also called the prompting method. Recent research [Gao et al., 2021, Liu et al., 2022, 2023, Shi et al., 2023] has extended this idea by improving prompts for pre-trained models by incorporating semantically similar examples. They apply the retrieval augmented method to discrete prompts, which are represented by tokens instead of vectors in a continuous space. They use them either for finetuning in few-shot settings or for zero-shot learning. Chowdhury et al. [2022] use a similar kNN-based retrieval method for tuning the soft prompts in a continuous space with a standard supervised training setup.

Multilingual In-Context Learning The effectiveness of prompting methods for English models extends to multilingual models in cross-lingual transfer learning as well. Zhao and Schütze [2021] and Huang et al. [2022] investigated the prompt-based learning with multilingual PLMs. Nie et al. [2023] incorporated augmented the prompt with cross-lingual retrieval samples in the multilingual understanding and proposed the PARC pipeline. PARC enhances the zero-shot learning performance for low-resource languages by cross-lingual retrieval from labeled or unlabeled high-resource languages. In the PARC pipeline, the cross-lingual retrieval first uses an low-resource language input sample as a query to find the semantically most similar high-resource language sample in the corpus. The recovered sample's label is received either from the corpus (labeled setting) or by self-prediction (unlabeled setting). The retrieved HRL sample together with its label, and the input sample are reformulated as prompts. Concatenation is used to generate the cross-lingual retrieval-augmented prompt, which is then used by the multilingual PLM to make a prediction. Tanwar et al. [2023] augmented the prompt with not only cross-lingual semantic information but also additional task information. However, previous studies mainly concentrated on the multilingual encoder or encoder-decoder models, while our work extends the PARC pipeline to the decoder-only multilingual LLMs.

Multilingual LLMs In the era of LLMs, BLOOMZ and mT0 [Muennighoff et al., 2023] are two representative newly emerging multilingual models. These two multilingual LLMs are fine-tuned on xP3, a multilingual multitask fine-tuning dataset, and based on the pre-trained models BLOOM [Scao et al., 2022] and mT5 [Xue et al., 2021], respectively. Six different sizes of BLOOMZ models are released from 560M to 176B and 5 different sizes of mT0 models are released from 300M to 13B. These multilingual LLMs open up the possibility for conducting few- and zero-shot cross-lingual in-context learning, as demonstrated by recent benchmarking efforts, for example, MEGA [Ahuja et al., 2023] and BUFFET [Asai et al., 2023].

3 Methodology

Our research extends the work of Nie et al. [2023] by focusing on improving multilingual pre-trained language models (MPLMs) for low-resource languages in a zero-shot setting. Figure 1 illustrates our two-stage pipeline. At its core, this pipeline synergistically combines the strengths of MPLMs with the semantic depth of high-resource languages for both classification and generation tasks.

3.1 Cross-Lingual Retrieval

The foundation of our method lies in the efficient retrieval of semantically relevant samples from high-resource languages, given a low-resource language input. Formally, for a given input sentence

q (the input test for classification and summarization task and the question for QA task) from a low-resource language, the cross-lingual retriever maps it to a vector q_{embed} in a shared embedding space using a function $Embed$:

$$q_{embed} = Embed(q)$$

For each document d_i in the high-resource language corpus, we compute its cosine similarity with q_{embed}:

$$Sim(q_{embed}, d_i) = \cos(q_{embed}, d_i)$$

The top k documents, which are most semantically aligned to the input, are then retrieved:

$$R_{indices} = \arg \max_{i \in \{1,\ldots,|d|\}}^{k} Sim(q_{embed}, d_i)$$

$$R = \{d_i | i \in R_{indices}\}$$

In cases where the retrieved documents d_i are unlabeled, a self-prediction mechanism that feeds d_i to the MPLM provides the necessary annotations.

3.2 Prompt Engineering

Using the semantically-rich retrieved samples R and the original input q, we craft a contextually-enriched input $\hat{q}$ using a predefined prompt template P:

$$\hat{q} = P(q, R)$$

This template P not only encapsulates the task-specific instructions but adeptly combines q and R to maximize the model's comprehension. The transformed input $\hat{q}$ is then processed by the MPLM to produce the desired output.

Depending on the architecture of the chosen MPLM, for autoregressive models like GPT variants, the model naturally generates an output sequence $\mathbf{y}$ for a given input $\hat{q}$:

$$\mathbf{y} = MPLM(\hat{q})$$

, whereas encoder models leverage a mask token prediction mechanism, utilizing a *verbalizer* to map labels to their corresponding linguistic representations, which will be shown in the experiment settings.

4 Experiments

To empirically validate our methodology, we design experiments that include both classification and generation. These experiments provide insights into the effectiveness of integrating MPLMs with high-resource language semantics, to improve ICL for low-resource languages.

4.1 Tasks and Datasets

Vio-Lens The Vio-Lens dataset [Saha et al., 2023] provides a rich collection of YouTube comments related to violent episodes in the Bengal region, structured for classification. Our prompt template P includes:

- Autoregressive models:
  `"Reflecting on the statement {text}, which aggressive level does it resonate with: non-aggressive, slightly aggressive, or highly aggressive?"`
- Mask prediction models: `"The underlying theme in {text} is [MASK]."`
  with the *verbalizer*:
  $v(0) =$ `"assaultive"`, $v(1) =$ `"indirect"`, $v(2) =$ `"peaceful"`

As the retrieval corpora for Vio-Lens, we use the labeled training set of English Sentiment Analysis dataset [Rosenthal et al., 2017], which consists of tweets annotated for sentiment on 2-, 3-, and 5-point scales with labels positive, negative, and neutral.

SentNoB The SentNoB dataset [Islam et al., 2021] is crafted to dissect the sentiment embedded within Bangla texts. Our prompt template P is articulated as:

- Autoregressive models:
  `"Text: {text} What is a possible sentiment for the text given the following options?"`
- Mask prediction models: `"{text} Sentiment: [MASK]"`
  with the *verbalizer*:
  $v(0) =$ `positive`, $v(1) =$ `neural`, $v(2) =$ `negative`

For SentNoB, we resort to the ETHOS dataset [Mollas et al., 2020], a comprehensive repository targeting online hate speech detection as the retrieval sentence pool. This repository provides a dataset designed to identify hate speech on social media, which contains 998 comments, each labeled for the presence or absence of hate speech. Since the labels are inconsistent, we rely on self-prediction to annotate the labels.

XLSum As a typical representative generation task, XLSum [Hasan et al., 2021], the multilingual text summarization dataset consisting of 1.35 million pairs of articles and their corresponding summaries. These pairs have been expertly annotated by the BBC and meticulously extracted through a series of carefully designed heuristic methods. We evaluate its Bangla subset to assess the performance of generation. The English training set is used as the retrieval corpus. The prompt template is defined as follows:

```
"{text} Generate a concise summary of the above text using the same
language as the original text:"
```

XQuAD the XQuAD (Cross-lingual Question Answering Dataset) [Artetxe et al., 2019] was used, which contains topic-diverse, manually-curated question-answer pairs with their respective contexts in 11 languages. Greek and Romanian language subsets are being evaluated. These pairs have been carefully translated from the English SQuAD v1.1 dataset to ensure high quality translations. The prompt template is defined as follows:

```
"context: {context} question: {question} answer:"
```

4.2 Models

BLOOM is an autoregressive Large Language Model trained on a diverse corpus to generate text based on prompts [Scao et al., 2022]. It is capable of generating coherent text in 46 languages.

BLOOMZ takes a novel approach in the MPLM landscape by applying Bloom filters in the context of language models [Muennighoff et al., 2023]. This allows the model to use high-resource languages to improve embeddings for low-resource languages, effectively bridging the gap between languages with different levels of available resources.

mBERT is an early MPLM that extends the original BERT model [Devlin et al., 2018]. It is pretrained on a corpus of 104 languages, using shared WordPiece vocabularies and a unified architecture for all languages.

mT5 or Multilingual T5 [Xue et al., 2021], is an extension of the T5 (Text-to-Text Transfer Transformer) model [Raffel et al., 2020] specifically designed for multilingual capabilities. Pre-trained on mC4, a large multilingual dataset, mT5 demonstrates multilingual capabilities by transforming input text sequences into output sequences.

Cross-Lingual Retriever We followed Nie et al. [2023] to use the multilingual sentence transformer "*paraphrase-multilingual-mpnet-base-v2*" [Reimers and Gurevych, 2019b]. This transformer maps sentences and paragraphs into a 768-dimensional dense vector space. Such a high-dimensional embedding facilitates tasks such as clustering and semantic search. Retrieval sample settings k are meticulously set at 1 and 3 for classification, and confined to 1 for summarization, 3 for QA.

Vio-Lens	zero shot	k=1	k=3	SentNoB	zero shot	k=1	k=3
bloomz-3b	0.19	0.2	0.24	bloomz-3b	0.34	0.44	0.44
bloom-3b	0.00	0.00	0.00	bloom-3b	0.00	0.00	0.00
mbert	0.21	0.28	0.29	mbert	0.30	0.36	0.37

Table 1: F1-scores of the two classification tasks: Bangla zero-shot baseline and our main method CREA-ICL with k retrieval augmented prompts.

	zero shot			k=1			k=3		
bloomz-3b	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
non-violence	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.04	0.08
passive violence	0.36	0.89	0.51	0.36	0.97	0.52	0.36	0.91	0.51
direct violence	0.09	0.10	0.10	0.18	0.06	0.09	0.17	0.07	0.10
accuracy			0.33			0.35			**0.36**
macro avg	0.15	0.33	**0.20**	0.18	0.34	**0.20**	0.26	0.26	0.17
weighted avg	0.14	0.33	0.19	0.15	0.35	0.20	0.42	0.36	**0.24**
mbert	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
non	0.46	0.28	0.35	0.47	0.52	0.49	0.46	0.55	0.50
passive	0.38	0.01	0.02	1.00	0.00	0.00	0.00	0.00	0.00
direct	0.09	0.60	0.16	0.09	0.36	0.14	0.08	0.30	0.13
accuracy			0.22			0.32			**0.33**
macro avg	0.31	0.30	0.18	0.52	0.29	**0.21**	0.18	0.28	**0.21**
weighted avg	0.40	0.22	0.21	0.62	0.32	0.28	0.26	0.33	**0.29**

	zero shot			k=1			k=3		
bloomz-3b	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
Negative	0.66	0.78	0.71	0.61	0.85	0.71	0.62	0.86	0.72
Neutral	0.00	0.00	0.00	0.23	0.02	0.03	0.18	0.01	0.01
Positive	0.57	0.72	0.64	0.60	0.56	0.58	0.59	0.57	0.58
accuracy			**0.61**			0.60			**0.61**
macro avg	0.31	0.37	0.34	0.48	0.48	**0.44**	0.47	0.48	**0.44**
weighted avg	0.51	0.61	**0.55**	0.53	0.60	0.54	0.53	0.61	0.54
mbert	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
negative	0.55	0.48	0.51	0.60	0.38	0.47	0.60	0.45	0.52
neutral	0.18	0.47	0.26	0.19	0.46	0.27	0.20	0.49	0.28
positive	0.40	0.07	0.13	0.43	0.29	0.35	0.46	0.24	0.32
accuracy			0.35			0.37			**0.39**
macro avg	0.38	0.34	0.30	0.40	0.38	0.36	0.42	0.39	**0.37**
weighted avg	0.43	0.35	0.34	0.47	0.37	0.39	0.48	0.39	**0.41**

Table 2: Confusion matrix of CREA-ICL method in Vio-Lens (top) and SentNoB (bottom) test set of BLOOMZ-3b and mBERT.

5 Results

5.1 Results of classification tasks

Table 1 provides a snapshot of the classification results. When we leveraged $k = 3$ retrieved English prompts, Bloomz-3b demonstrated an improvement in F1-scores for both tasks by 5% and 10% respectively. The striking null results from Bloom-3b, in contrast to Bloomz-3b, emphasize the pivotal role instruction tuning plays in retrieval-augmented in-context learning. In comparison, the traditional masked MLM, mBERT, also registered an improvement of 8% and 7%, respectively.

A deeper dive into the confusion matrix, as presented in Table 2, reveals intriguing insights:

1) With a general assessment across micro, macro, and weighted F1 scores, Bloomz-3b and mBERT gained improvement from the retrieval prompts.

2) Comparing the two models, Bloomz-3b's zero-shot setting tends to misclassify "non-violence" and "Neutral", and has a reduced macro F1 compared to its weighted F1, while mBERT has a more balanced distribution of confusion between "non-violence" ("Neutral") and the other classes.

These may indicate that for classification tasks, the generative models struggle more with minority classes compared to masked prediction.

LEAD-64		zero shot			k=1		
		mt5-base	bloomz-1b1	bloomz-3b	mt5-base	bloomz-1b1	bloomz-3b
R-1	18.17	5.01	22.08	22.36	0.97	10.84	6.61
R-2	5.23	0.84	7.11	7.88	0.13	2.80	1.52
R-L	12.73	4.83	18.43	18.60	0.91	9.11	5.56
R-LSum	12.74	4.84	18.44	18.58	0.92	9.12	5.55

		zero-shot		k=3	
		EM	F1	EM	F1
Greek	el	15.29	19.92	8.07	10.34
Romanian	ro	38.66	49.39	15.88	20.93

Table 3: Results of Bangla summarization (top) and QA task (bottom), including the zero-shot baseline and CREA-ICL method with k=1 or k=3 retrieved samples.

5.2 Results of generation task

The summarization task results, as shown in Table 3, provide a comprehensive understanding of the models' capabilities:

LEAD-64 The LEAD-64 baseline operates on a straightforward extractive approach where the first 64 tokens of the input text are considered as its summary. Its commendable performance across various metrics reiterates the often-overlooked significance of the initial segments in articles or documents. These segments frequently encapsulate key points, making them effective summaries. In a zero-shot setting, LEAD-64's results surpass those of the mt5-base model. However, when compared to the Bloomz variants, it finds itself overshadowed. This result emphasizes the effect of instruction tuning that enhances ICL performance.

Zero-Shot Baseline mt5-base's discernible underperformance across the board underscores its struggles in generating high-quality summaries without specialized domain adaptation or data enrichment. In stark contrast, the Bloomz models exhibit remarkable competency, with Bloomz-3b slightly outdoing Bloomz-1b1, especially in the context of the R-2 metric, which evaluates bigram comprehension.

CREA-ICL with k=1 Retrieval augmentation seems to drastically affect the performance of mt5-base, reducing its score considerably. This could be due to noise introduced by the retrieved sample or ineffective use of the additional information. For the Bloomz models, Bloomz-1b1 still retains decent performance, although there's a drop when compared to its zero-shot performance. Surprisingly, Blommz-3b shows a sharper drop, suggesting that the additional retrieval data may be more of a distraction than an advantage for this model configuration in the summarization task.

From the table 3, which shows the performance metrics of the QA task in Greek and Romanian. The CREA-ICL method with k=3 using the Bloomz-3b model seems to be less effective than the zero-shot baseline using the same model for both languages. Both EM and F1-score metrics demonstrate a significant performance advantage, nearly doubling their values compared to CREA-ICL.

5.3 Analysis and Discussion

When analyzing the effectiveness of different models across different tasks, it becomes clear that for classification tasks, advanced models have the ability to identify complex categories. This ability is not limited to a single language or dataset, but extends to a variety of linguistic contexts. Zero-shot learning, a method in which models apply knowledge without direct task-specific training, proves to be versatile and shows remarkable capability even without task-specific fine-tuning. Exploration of different numbers of retrieved samples (k) shows that increasing k does not necessarily improve performance, suggesting strategic information filtering and noise reduction capabilities of the models.

In summarization tasks, maintaining coherence and relevance is critical. In such generative tasks, models designed for generation generally outperform those designed for extraction in the absence of task-specific examples. However, these generative models are sensitive to the addition of retrieved

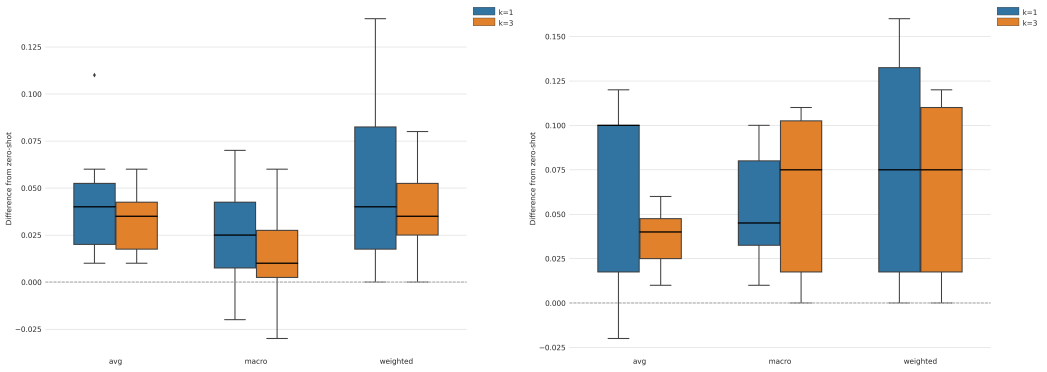

Figure 2: Model performance over differences between zero-shot and CREA-ICL method with k=1 and k=3 demonstrations for Vio-Lens test set using bloomz-3b (left) and mbert (right).

Q: Câte mingi a interceptat Josh Norman?
(How many balls did Josh Norman intercept?)
Referring to the passage above, the correct answer to the given question is:

generated answer using CREA-ICL: patru *(four)*
generated answer using zero-shot: four
Gold answer: patru *(four)*

Figure 3: Example showing the different answer language when answering a Romanian question

information, which can affect their performance. The balance between performance and computational effort becomes clear when analyzing resource allocation, and shows that while models such as Bloomz-3b demonstrate superior performance, larger models do not guarantee better results across all evaluation criteria. This suggests the need for a more selective approach to model selection.

Furthermore, in the question-answering (QA) domain, zero-shot methods have been shown to outperform the CREA-ICL approach. This underscores the robustness of zero-shot strategies in classification contexts, but also points to their limitations in generative tasks, where they may not always be the optimal choice. This finding encourages a more nuanced application of models, depending on the nature of the task at hand.

Our analysis of cross-lingual tasks shows that classification accuracy often has an advantage over that of generative tasks, especially when dealing with multilingual data. Classification tasks typically require the model to produce short, constrained outputs, often limited to a single word or short phrase from a predefined set of options. This specificity in response reduces the likelihood of encountering language confusion.

Conversely, generative tasks require the production of longer sequences of text, which increases the challenge of maintaining language consistency. This is evident in scenarios where a model, while capturing the correct conceptual meaning, outputs in an unexpected language. Figure 3 illustrates this, where the generated output is conceptually accurate but linguistically misaligned. Our methodology, with its dependence on English templates, may inadvertently worsen this problem rather than mitigate it, as the templates may bias the generative process towards English.

Furthermore, the metrics currently used for generative tasks still struggle with evaluating output that is linguistically inconsistent. Our generative tasks have been modified to include a language constraint that requires the output to match the language of the original text. However, as shown in Table 4, this adjustment resulted in only marginal improvements. The implications are significant: when evaluating generative tasks in a cross-lingual context, there remains a significant challenge in accurately capturing the quality of language-specific output. In the future, the development of metrics that can better account for language consistency will be crucial for evaluating the true effectiveness of models in generative multilingual scenarios.

	R-1	R-2	R-L	R-L-Sum
{text} Generate a concise summary of the given text	0.55	0.07	0.53	0.53
...using the same language as the original text({target_lang})	0.97	0.13	0.91	0.92

Table 4: results with and without language constraint in prompt templates of summarization task using mt5-base model with k=1

	k=1			k=3		
	precision	recall	f1-score	precision	recall	f1-score
bangla prompt	"পাঠ্য: {text} নিম্নলিখিত বিকল্পগুলি দেওয়া পাঠ্যের জন্য সম্ভাব্য অনুভূতি কী ?"					
Negative	0.58	0.20	0.29	0.60	0.59	0.60
Neutral	0.20	0.06	0.10	0.16	0.08	0.11
Positive	0.60	0.08	0.15	0.50	0.44	0.47
accuracy			0.14			0.45
macro avg	0.34	0.09	0.13	0.32	0.28	0.29
weighted avg	0.51	0.14	0.21	0.49	0.45	0.46
hindi prompt	"पाठ: {text} निम्नलिखित विकल्पों को देखते हुए पाठ के लिए संभावित भावना क्या है? "					
Negative	0.61	0.45	0.52	0.62	0.74	0.68
Neutral	0.18	0.34	0.24	0.19	0.14	0.16
Positive	0.55	0.31	0.40	0.57	0.48	0.52
accuracy			0.39			0.54
macro avg	0.34	0.28	0.29	0.34	0.34	0.34
weighted avg	0.51	0.39	0.43	0.52	0.54	0.53

Table 5: Results of prompt template in bangla and hindi of CREA-ICL method in SentNoB test of bloomz-3b.

6 Ablation Study

6.1 The Stability across Templates

In our investigation of the Vio-Lens dataset, we evaluated the classification capabilities of Bloomz-3b and mBERT by categorizing text samples. Our goal was to compare the effectiveness of retrieval-augmented prompting to a zero-shot baseline by analyzing performance across different prompt templates.

We applied different templates to both Bloomz-3b and mBERT and generated a boxplot (Figure 2) to show the F1 score variations when using the CREA-ICL method versus the zero-shot baseline. The visual representation confirmed that there was a consistent improvement in performance with the retrieval-augmented English prompts compared to the zero-shot baseline, which did not focus on any specific language. Interestingly, mBERT showed a more pronounced improvement in F1 scores than Bloomz-3b when moving from the zero-shot baseline to the retrieval-augmented prompts. This finding highlights the potential of retrieval augmentation to improve the model's text classification performance beyond the context of a single language.

6.2 Evaluating the Influence of Non-English Prompt Templates

Expanding our scope beyond English, we explored the use of Bangla and its linguistically similar, high-resource counterpart Hindi as prompt templates P, as detailed in Table 5.

We observed that using Hindi as a template language led to precision and recall improvements in some categories. However, it did not surpass the macro average F1 score achieved by the CREA-ICL method with English prompts. The Bangla template, while improving precision in some instances, suffered a drop in recall and overall accuracy, culminating in the least impressive macro average F1 score among the templates examined.

These results suggest that while the Bangla template may improve category-specific performance, it compromises the model's ability to generalize across the spectrum of categories in the SentNoB test.

	k=1			k=3		
bloomz-3b	precision	recall	f1-score	precision	recall	f1-score
Negative	0.58	0.84	0.69	0.59	0.88	0.70
Neutral	0.09	0.00	0.00	0.08	0.00	0.00
Positive	0.55	0.49	0.52	0.58	0.47	0.52
accuracy			0.57			0.58
macro avg	0.41	0.44	0.40	0.42	0.45	0.41
weighted avg	0.48	0.57	0.51	0.49	0.58	0.51
mbert	precision	recall	f1-score	precision	recall	f1-score
Negative	0.48	0.24	0.32	0.48	0.33	0.39
Neutral	0.21	0.34	0.26	0.21	0.28	0.24
Positive	0.27	0.37	0.31	0.25	0.33	0.28
accuracy			0.30			0.32
macro avg	0.32	0.32	0.30	0.31	0.31	0.31
weighted avg	0.36	0.30	0.30	0.36	0.32	0.33

Table 6: Results in SentNoB test of BLOOMZ-3b and mBERT with hindi retrieval corpus.

Similarly, the Hindi template's category-specific improvements in precision and recall do not translate into an increased macro-average F1 score over the CREA-ICL method with English prompts.

In conclusion, the comprehensive F1 score analysis underscores the superiority of the CREA-ICL method coupled with English prompts in terms of overall effectiveness. However, the impact of prompt template language on the performance of specific categories is significant, as evidenced by the Hindi and Bangla templates. This highlights the importance of striking a strategic balance between improving category-focused performance and maintaining overall effectiveness when selecting prompt templates for cross-language retrieval enhancement tasks.

6.3 Impact of Hindi Retrieval Corpus

Comparing the results in Table 6 with the previous experiments, it is clear that neither Bloomz-3b nor mBERT show improvements over the CREA-ICL method using English Retrieval Corpus. This implies that while the use of different retrieval datasets has the potential to improve results within specific sentiment classifications, the selection of retrieval content must be carefully considered to optimize collective performance across different categories in cross-lingual sentiment analysis efforts.

A consistent challenge is the "Neutral" category using the Bloomz-3b model, which suffers from inadequate recall and F1 scores regardless of the retrieval corpus used. This pattern suggests that additional model refinements and strategy adjustments are needed to increase the accuracy of neutral sentiment retrieval.

7 Conclusion

In this study, we presented an innovative methodology CREA-ICL to harness the power of Large Language Models for low-resource languages, with particular emphasis on Bengali. By integrating cross-lingual retrieval-augmented in-context learning, we sought to enhance the capabilities of MPLMs, notably BLOOM and BLOOMZ. Our methods were rigorously evaluated across two classification tasks and two generation tasks.

The empirical outcomes underscore the success of our strategy, as evidenced by the notable F1-scores in classification tasks. A deeper dive into our results reveals the pivotal role the CREA-ICL mechanism plays in strengthening the model's effectiveness.

Our research paves the way for ensuing investigations into cross-lingual retrieval and in-context learning's potential in the realm of low-resource languages. For future work, there's an exciting possibility to adapt and extend this framework to other marginalized languages and to address more complicated NLP challenges, including question-answering and machine translation.

References

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. URL https://api.semanticscholar.org/CorpusID:13756489.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020a.

OpenAI. Gpt-4 technical report, 2023.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080, 2020. URL https://api.semanticscholar.org/CorpusID:214641214.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*, 2023.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.822. URL https://aclanthology.org/2022.emnlp-main.822.

Salim Sazzed. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.wnut-1.8. URL https://aclanthology.org/2020.wnut-1.8.

Amitava Das and Sivaji Bandyopadhyay. Phrase-level polarity identification for bangla. *Int. J. Comput. Linguistics Appl.*, 1(1-2):169–182, 2010.

Amitava Das and Björn Gambäck. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India, December 2014. NLP Association of India. URL https://aclanthology.org/W14-5152.

Anirban Bhowmick and Abhik Jana. Sentiment analysis for Bengali using transformer based models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 481–486, National Institute of Technology Silchar, Silchar, India, December 2021. NLP Association of India (NLPAI). URL https://aclanthology.org/2021.icon-main.58.

Md Ferdous Wahid, Md Jahid Hasan, and Md Shahin Alom. Cricket sentiment analysis from bangla text using recurrent neural network with long short term memory model. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE, 2019.

Muntasir Hoq, Promila Haque, and Mohammed Nazim Uddin. Sentiment analysis of bangla language using deep learning approaches. In *COMS2*, 2021. URL `https://api.semanticscholar.org/CorpusID:236696696`.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*, 2023.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval, 2021.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.528. URL `https://aclanthology.org/2023.findings-acl.528`.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL `https://aclanthology.org/2023.acl-long.891`.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla A. Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph

Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, A. Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100, 2022. URL `https://api.semanticscholar.org/CorpusID:253420279`.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 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, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL `https://aclanthology.org/P17-1171`.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. R3: Reinforced ranker-reader for open-domain question answering. In *AAAI*, 2018.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *North American Chapter of the Association for Computational Linguistics*, 2020.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *ArXiv*, abs/1903.10561, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2020.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL `https://aclanthology.org/D17-1070`.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019a. URL `https://arxiv.org/abs/1908.10084`.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020b.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Yanchen Liu, Timo Schick, and Hinrich Schtze. Semantic-oriented unlabeled priming for large-scale language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 32–38, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sustainlp-1.2. URL https://aclanthology.org/2023.sustainlp-1.2.

Jishnu Ray Chowdhury, Yong Zhuang, and Shuyi Wang. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. In *AAAI*, 2022.

Mengjie Zhao and Hinrich Schütze. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.672. URL https://aclanthology.org/2021.emnlp-main.672.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.790. URL https://aclanthology.org/2022.emnlp-main.790.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.346. URL https://aclanthology.org/2023.acl-long.346.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Sourav Saha, Jahedul Alam Junaed, Arnab Sen Sharma Api, Nabeel Mohammad, and Mohammad Ruhul Amin. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore, dec 2023. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://aclanthology.org/S17-2088.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.278. URL https://aclanthology.org/2021.findings-emnlp.278.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: an online hate speech detection dataset, 2020.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-acl.413.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:204901567.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.