# Crosslingual Retrieval Augmented In-context Learning for Bangla

**Xiaoqian Li**[1] **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
Xiaoqian.Li@campus.lmu.de
{nie, shengliang}@cis.lmu.de

## Abstract

The promise of Large Language Models (LLMs) in Natural Language Processing has often been overshadowed by their limited performance in low-resource languages such as Bangla. To address this, our paper presents a pioneering approach that utilizes cross-lingual retrieval augmented in-context learning. By strategically sourcing semantically similar prompts from high-resource language, we enable multilingual pretrained language models (MPLMs), especially the generative model BLOOMZ, to successfully boost performance on Bangla tasks. Our extensive evaluation highlights that the cross-lingual retrieval augmented prompts bring steady improvements to MPLMs over the zero-shot performance.

## 1 Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed transformative advancements, especially with the advent of deep transformer techniques (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019). The introduction of Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020b) and GPT-4 (OpenAI, 2023), has further revolutionized the landscape. These models showcase unparalleled prowess in tasks like text classification and generation, unified under the umbrella of in-context learning, and cater to a plethora of applications across diverse languages (Conneau et al., 2020; Raffel et al., 2020; Radford et al., 2019). While comprehensive benchmarks like XTREME (Hu et al., 2020) and BUFFET (Asai et al., 2023) underscore their capabilities, languages such as English remain the primary beneficiaries. In stark contrast, several low-resource languages, Bangla being a prime example, grapple with challenges, notably the scarcity of pretraining corpora (Artetxe and Schwenk, 2019; Hangya et al., 2022; Sazzed, 2020).

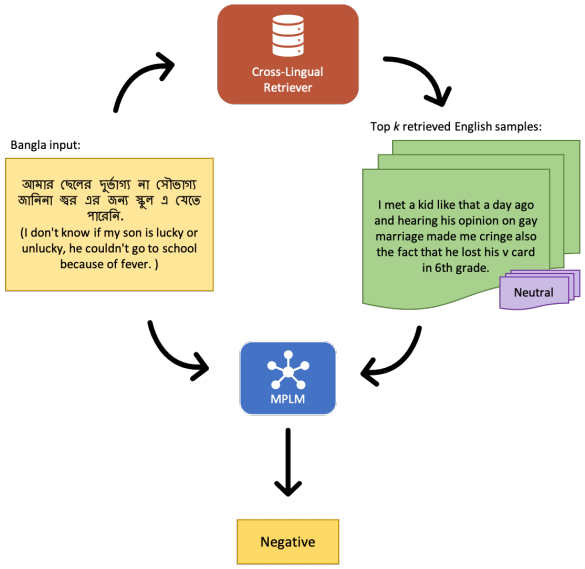


Figure 1: PARC pipeline using decoder-only Multilingual Pretrained Language Models.

Despite having a significant number of native speakers, Bangla remains underrepresented in the NLP arena due to linguistic intricacies, limited labeled datasets, and prevalent issues like data duplication (Das and Bandyopadhyay, 2010; Das and Gambäck, 2014). Although there have been commendable strides using conventional machine learning techniques in Bangla NLP tasks, the untapped potential of the latest LLMs is evident (Bhowmick and Jana, 2021; Wahid et al., 2019; Hoq et al., 2021).

In the evolving landscape of in-context learning with LLMs, the concept of retrieval augmentation, which emphasizes sourcing semantically rich prompts, has gained traction (Shi et al., 2023). However, when it comes to multilingual in-context learning, previous works like MEGA (Ahuja et al., 2023) often limit their scope to task instructions and lack deeper semantic insights due to their

† Corresponding author.



*Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 136–151

approach of random prompt selection. In contrast, strategies like PARC (Nie et al., 2023) pave the way for a more comprehensive methodology, fetching semantically aligned prompts from high-resource languages.

Our work draws inspiration from these methodologies but introduces novel perspectives. While MEGA offers task-level instructions, we infuse semantic understanding into our approach. Similar to PARC, our approach is cross-lingual, ensuring a broader application spectrum. Diverging from PARC's focus on masked language models like mBERT and XLMR, as shown in Figure 1, we venture into uncharted territories by employing larger, decoder-only multilingual pretrained language models (MPLMs) — BLOOM and BLOOMZ — to tackle Bangla NLP tasks in a generative style (Muennighoff et al., 2023; Scao et al., 2022).

In this paper, we explore the application of cross-lingual retrieval augmented in-context learning to Bangla text classification and summarization tasks. Our main contributions encompass:

- An extensive evaluation of cross-language retrieval augmented in-context learning methods in Bangla, achieving steady improvements over the zero-shot performance of MPLMs.
- A pioneering exploration to extend PARC to the generative models, BLOOM and BLOOMZ, providing insights for a unified pipeline of cross-lingual retrieval augmented in-context learning.

## 2 Related Work

**Bangla Natural Language Processing** Bangla is a morphologically rich language with various dialects that belongs to the Indo-Aryan branch of the Indo-European language family. With roughly 270 million speakers concentrating in Bangladesh and some regions of India, Bangla is ranked as the 7th most widely spoken language in the world[1]. However, Bangla is still considered as a low-resource language in the NLP research due to the scarcity of digital text resources and annotated corpora.

Research on Bangla NLP has covered a variety of common NLP subfields since 1990s, such as POS tagging (Dandapat et al., 2004; Ekbal and Bandyopadhyay, 2008b), stemming and lemmatization (Islam et al., 2007; Paik and Parui, 2008), named entity recognition (Ekbal and Bandyopadhyay, 2007, 2008a), sentiment analysis (Das and Bandyopadhyay, 2010; Wahid et al., 2019), news categorization (Mansur, 2006; Mandal and Sen, 2014), etc. However, the research in different areas of Bangla NLP still remains sparse. In the era of deep learning, further progress has been made in Bangla NLP, particularly in terms of the development datasets (Rahman and Kumar Dey, 2018; Islam et al., 2021, 2023) and models (Tripto and Ali, 2018; Ashik et al., 2019; Karim et al., 2020). Pretrained language models have achieved decent performance in a large variety of NLP downstream tasks through the fine-tuning. Under this background, Bhattacharjee et al. (2022) pretrained the BanglaBERT model, a BERT-based language understanding model pretrained on Bangla language corpora. With the advent of the large language models (LLMs), zero- and few-shot prompting methods have gradually gained prominence. Hasan et al. (2023) compared the zero- and few-shot prompting performance of LLMs with the finetuned models for the Bangla sentiment analysis task. Our work explores the application of the retrieval-augmented prompting method in Bangla violence detection and sentiment analysis tasks.

**Multilingual In-context Learning** Brown et al. (2020a) demonstrated 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 expanded on this idea by enhancing prompts for pretrained models through the inclusion of semantically similar examples. 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. Tanwar et al. (2023) augmented the prompt with not only cross-lingual semantic

[1] https://www.ethnologue.com/insights/ethnologue200/

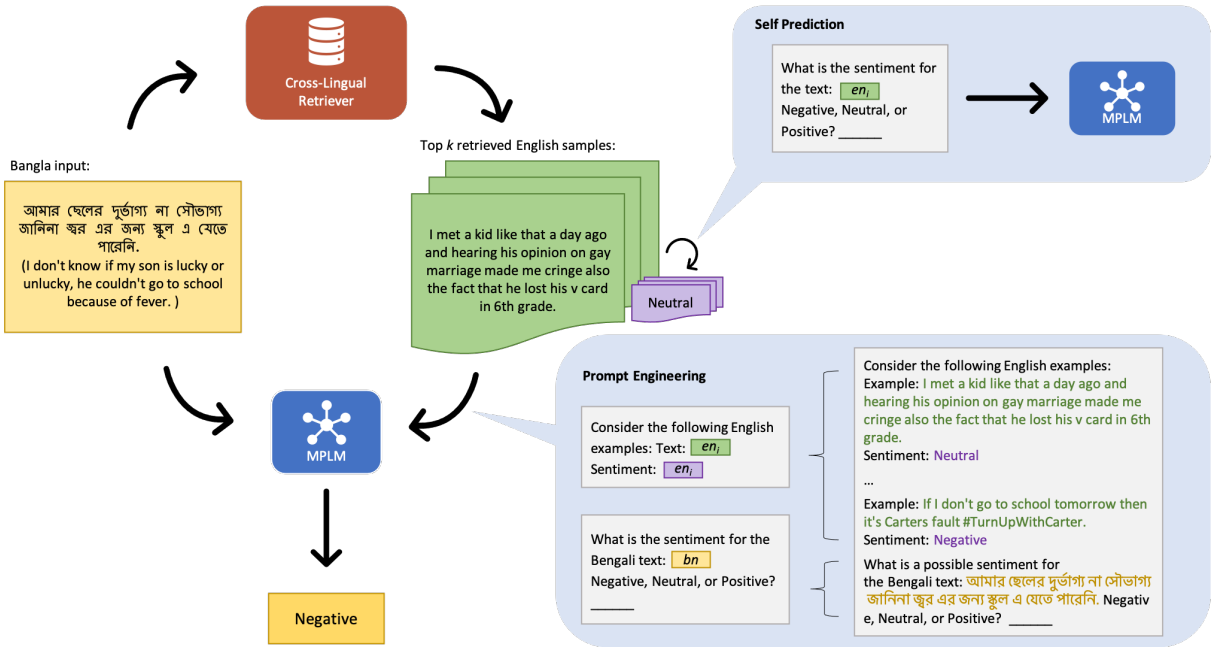


Figure 2: Detailed overview of the PARC pipeline for LRLs using cross-lingual retrieval: (a) An LRL input is used as a query for the cross-lingual retriever, which then retrieves the most semantically similar HRL sample from the HRL corpus. The associated label is either taken directly from the corpus (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-enhanced prompt for MPLM prediction.

information but also additional task information. However, previous studies mainly concentrated on the multilingual encoder or encode-decoder models, while our work extend 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 finetuned on xP3, a multilingual multitask finetuning dataset, and based on the pretrained 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, specifically using retrieved content from high-resource languages such as English.

The backbone of our research approach is a two-stage pipeline consisting of a cross-lingual retriever and a prompt engineering process as shown in Figure 2. This pipeline aims to build on the strengths of MPLMs while mitigating their limitations, especially when dealing with low-resource languages. The first stage of the pipeline uses a cross-lingual retriever that maps the input Bangla text $q$ to a vector $q_{embed}$ in a shared embedding space and uses it as a query. Using semantic similarities with $q_{embed}$, the retriever returns the most similar k examples from high-resource languages either with or without their labels:

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

where $d_i$ means each document in the high-resource language corpus and $|d|$ is the number of documents. If there's no label, it suggests a self-prediction step.

The second stage of the pipeline is the prompt engineering. The input Bangla text and the retrieved pattern are subjected to this process. A prefix prompt template $P$ is used to reformulate the

input to facilitate the model's prediction **y**:

$$\mathbf{y} = MPLM(P(q, R))$$

Depending on the architecture of the chosen MPLM, for decoder-only models, the answer is generated by the model directly. For encoder models, the answer is obtained by first mapping each label to its predefined word using the *verbalizer* and then deducing the label word using mask token prediction.

By integrating cross-lingual content retrieval with prompt-guided prediction, we aim to improve the ability of MPLMs to handle low-resource languages. This synergy not only extracts rich linguistic insights from high-resource languages, but also uses them to improve performance on low-resource language tasks.

## 4 Experiments

In this study, we focused on the tasks of classification and summarization. We refer to our research approach, which uses k retrieved samples for cross-lingual augmented in-context learning methods, as the main method in the following sections.

### 4.1 Baselines

**Zero-shot** The template, when populated with the input sample, is fed directly into the MPLM for prediction. This process bypasses the use of cross-lingual context.

**Lead64** The first 64 tokens of the input text are taken as a summary of the text (For summarization tasks only).

### 4.2 Tasks

#### 4.2.1 Classification

**Vio-Lens** The Vio-Lens dataset (Saha et al., 2023) contains YouTube comments related to violent incidents in the Bengal region, with the goal of highlighting potential threats that could incite further violence. The prompt templates for both main method and zero-shot baseline are defined as follows:

- BLOOMZ-3b and BLOOM-3b:
  `Reflecting on the statement "{text}", which aggressive level does it resonate with: non-aggressive, slightly aggressive, or highly aggressive?`
- mBERT: `The underlying theme in {text} is [MASK].`
  with the verbalizer:
  $v(\textit{Direct Violence}) =$ `assaultive`,
  $v(\textit{Passive Violence}) =$ `indirect`,
  $v(\textit{Non-Violence}) =$ `peaceful`

The 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, serves as the HRL corpora in our study. We use the labeled training set for our experimental sentence pool.

**SentNoB** Designed to capture the sentiment within text, SentNoB classifies content as positive, negative or neutral (Islam et al., 2021). The prompt templates for both main method and zero-shot baseline are defined as follows:

- BLOOMZ-3b and BLOOM-3b:
  `Text: {text} What is a possible sentiment for the text given the following options?`
- mBERT: `{text} Sentiment: [MASK]`
  with the verbalizer:
  $v(0) =$ `positive`, $v(1) =$ `neural`,
  $v(2) =$ `negative`

We use the ETHOS (onlinE haTe speecH detectiON dataSet) (Mollas et al., 2020) as sentence pool in our experiments. This repository provides a dataset designed to identify hate speech on social media. We use the binary variants of the dataset, which contains 998 comments, each labeled for the presence or absence of hate speech. Since the labels are inconsistent, we use the self-prediction method to predict the labels.

#### 4.2.2 Summarization

**XL-Sum** is a large and varied dataset consisting of 1.35 million pairs of articles and their corresponding summaries (Hasan et al., 2021). These pairs have been expertly annotated by the BBC and meticulously extracted through a series of carefully designed heuristic methods. The dataset includes 45 languages, from low to high resource, many of which do not currently have publicly available datasets. The prompt template is defined for all models as follows:

- Main method:
  `{text} Generate a concise summary`

```
of the above text using the same
language as the original text
({target_lang}):
```

- Zero-shot baseline:
```
{text} Generate a concise summary
of the given text:
```

### 4.3 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) designed specifically 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, 2019). 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. In our experiments, the number of retrieval samples $k$ is 1 and 3 for classification task and 1 for summarization task.

## 5 Results

### 5.1 Results of classification tasks

Table 1 provides an overview of the results of classification. With the instructions of $k = 3$ retrieval augmented English prompts, we enhance the F1-scores of Bloomz-3b on the two tasks by 5% and 10% respectively. While Bloom-3b, without instruction tuning compared to Bloomz-3b, cannot generate any meaningful result, suggesting that instruction tuning has a strong impact on retrieval augmented in-context learning. The traditional masked MLM, mBERT, also gained improvement by 8% and 7%.

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

Table 1: F1-scores of the two classification tasks: Bangla zero-shot baseline and with $k$ retrieval augmented prompts.

To facilitate a comprehensive understanding of the performance and discrepancies associated with each task, we present confusion matrices for analysis as follows. Given the confusion matrix in Table 2 , we find that:

1) With a general assessment across micro, macro, and weighted F1 scores, Bloomz-3b and mBERT gained improvement from the retrieval prompts. 2) Compare 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. This may indicate that for classification tasks, the text generation struggles more with minority classes compared to masked prediction.

### 5.2 Results of summarisation task

The Table 3 compares several models and methods for summarization task.

**LEAD-64** As an extractive method, it performs well across all metrics. This indicates that in many cases the first few sentences or tokens of an article or document provide a fairly informative summary. As expected, LEAD-64 outperforms the mt5 base model in the zero-shot setting, but is outperformed by the Bloomz models in the same scenario.

| | zero shot | | | k=1 | | | k=3 | | |
|---|---|---|---|---|---|---|---|---|---|
| bloomz-3b | precision | recall | f1-score | precision | recall | f1-score | precision | recall | f1-score |
| 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 |
| 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 |
| 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 |
| 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 main method in Vio-Lens (top) and SentNoB (bottom) test set of BLOOMZ-3b and mBERT.

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

Table 3: Rouge scores of Bangla summarization.

**Zero-Shot Models** mt5-base produces the lowest scores across all metrics, suggesting that it struggles to produce satisfactory summaries without domain-specific fine-tuning or data augmentation. Both bloomz-1b1 and bloomz-3b show significantly better performance, with bloomz-3b having a slight edge over bloomz-1b1, especially in bigram capture (R-2).

**Retrieval augmentation 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.

### 5.3 Analysis and Discussion

When examining the performance of different models on different tasks, several key observations emerge that are related to linguistic nuances, the underlying language models, and resource allocation.

For classification tasks, it's clear that models with a strong grasp of complex sentence structure and deeper semantics, such as the Bloomz-3b, are more adept at distinguishing nuanced categories like "passive violence" or the more ambiguous "neutral" sentiment. This aptitude likely stems from their ability to understand context better than their simpler counterparts. In parallel, the critical role of zero-shot learning becomes apparent. The ability of a model to generalize a task without specific fine-tuning speaks volumes about its robustness. For example, in our studies, models such as the Bloomz-3b showed commendable performance in a zero-shot setting. Furthermore, as we played around with the variable k (representing the number of samples retrieved), it was instructive to see that a larger value didn't always translate into better performance. This underscores the nuanced ability of a model to sift through information

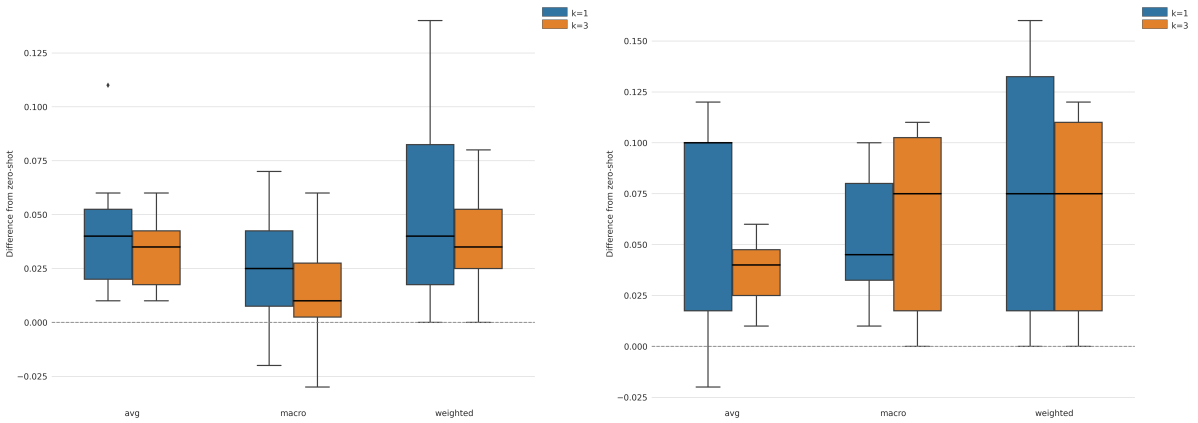


Figure 3: Model performance over differences between zero-shot (represented as '0' on the y-axis) and main method with k=1 and k=3 demonstrations for Vio-Lens test set using bloomz-3b (left) and mbert (right).The y-axis shows the deviations of the main method from the zero-shot values. The statistics are based on 8 and 6 templates, shown in Appendix Table 5 and Table 6, respectively.

and potentially eliminate noise.

Turning to the summarization task, coherence and relevance seem to be the pillars of excellence. Advanced models are more adept at weaving sentences that are not only structurally coherent, but also rich in information. This finesse is evident in the superior Rouge scores of the models. The dichotomy between generative and extractive approaches is also evident. While generative models, including mt5-base and Bloomz-1b1, outperformed the extractive model (LEAD-64) in a zero-shot framework, they seemed a bit sensitive when retrieval augmentation came into play.

Finally, when it comes to resource distribution, there's an undeniable correlation between performance and computational resources. The stellar performance of models like Bloomz-3b likely comes at the cost of intense computational demands. However, one must consider the cost-benefit ratio. In addition, the drop in performance of these models with retrieval augmentation at k=1 suggests a potential sensitivity to the balance or diversity of the dataset.

For the summarization task, an interesting observation is that more extensive models don't always outperform on all metrics, suggesting that we need to be more discriminating in our resource allocation. The significant performance drop with retrieval augmentation further supports this argument.

To conclude this analysis, while modern language models are capable of handling complex tasks, they require careful configuration and thoughtful resource distribution. Unraveling the complexity of these models can pave the way for optimized solutions in both classification and summarization.

## 6 Ablation Study

### 6.1 The Stability across Templates

In our experiment for Vio-Lens, we compared the performance of Bloomz-3b and mbert, in terms of their ability to classify text samples into categories. In order to assess the effectiveness of the retrieval augmented prompting method compared to the zero-shot baseline, we conduct a statistic across different templates.

For Bloomz-3b and mBERT, we test different prompt templates , and created a boxplot (Figure 3) to visualize the difference of F1 scores from our main method to the zero-shot baseline across templates. It's shown that with the retrieval augmented English prompts under different templates, both model achieved a stable improvement compared to the Bangla zeroshot baseline. Aso it's clear that mBERT, on average, shows greater improvements in F1 scores when transitioning from the zero-shot baseline to retrieval augmented prompting, compared to Bloomz-3b.

### 6.2 Impact of Bangla and Hindi Prompt Template

Instead of English, we further explore applying Bangla itself and its linguistically similar high-resource language Hindi as the language of the prompt template , as shown in Table 4.

Main method with English prompt: This configuration yields the highest macro average F1 score

| | k=1 | | | k=3 | | |
|---|---|---|---|---|---|---|
| | precision | recall | f1-score | precision | recall | f1-score |
| bangla prompt | "পাঠ্য: {text} নিম্নলিখিত বিকল্পগুলি দেওয়া পাঠ্যের জন্য সম্ভাব্য অনুভূতি কী ?" | | | | | |
| 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? " | | | | | |
| 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 4: Results of prompt template in bangla and hindi of main method in SentNoB test of bloomz-3b.

of all three prompt templates.

Hindi Prompt Template: While the Hindi prompt template leads to significant improvements in precision and recall for individual categories such as “Neutral”, the macro average F1 score is still lower than that of the main method with the English prompt.

Bangla prompt template: The Bangla prompt template, while showing some improvements in precision for specific categories such as “positive”, experiences a decrease in recall and overall accuracy. As a result, the macro average F1 score is the lowest of the three templates.

This means that while the Bangla prompt template may improve performance for specific categories, it has an overall negative impact on the model’s ability to generalize across all categories in the SentNoB test. Conversely, the Hindi prompt template’s improvements in precision and recall for individual categories don’t translate into a higher macro average F1 score compared to the main method with the English prompt.

In summary, the macro average F1-score results show that the main method with the English prompt template remains the most effective overall. However, the choice of prompt template can significantly affect performance for specific categories, as demonstrated by the Hindi and Bangla templates. This nuanced understanding underscores the need to balance category-specific and overall performance when selecting prompt templates in cross-lingual retrieval augmentation.

### 6.3 Impact of Hindi sentence pool

Comparing the results in Table 7 with the previous experiments, we observe that the Hindi retrieval dataset generally improves the model’s ability to retrieve “Neutral” content in the mBERT model. However, the model continues to struggle with the “Neutral” category, with low recall and F1 scores, regardless of the sentence pool used. This suggests that further refinements may be needed to improve retrieval accuracy for neutral sentiment sentences. The studies with Hindi retrieval data show that both bloomz-3b and mbert don’t show any improvements compared to the main method with the English prompt template. This suggests that while using alternative retrieval datasets can improve performance for specific sentiment categories, the choice of retrieval data may need to be carefully considered to maximize overall performance across categories in cross-lingual sentiment analysis tasks.

## 7 Conclusion

In this paper, we have introduced a novel approach to address the challenges of applying Large Language Models to low-resource languages, with a focus on Bangla. Our methodology employs cross-lingual retrieval-augmented in-context learning, thereby enriching the capabilities of MPLMs, specifically BLOOM and BLOOMZ. We have extensively tested our approach on two classification tasks and one summarization task.

Our experimental results demonstrate the effectiveness of our approach in achieve superior F1 scores for classification tasks.

Upon further analysis, the cross-lingual retrieval mechanism contributes significantly to the model’s performance.

This work lays the foundation for further studies on the application of cross-lingual retrieval and in-context learning methods in low-resource languages. Future work could extend this approach to even more underrepresented languages and potentially adapt it to more complex NLP tasks such as question answering or machine translation.

## Limitations

While our study has yielded promising results, it is not without limitations. The effectiveness of retrieval augmentation is also tied to the model architecture, and its impact on different models remains largely unexplored. In addition, the availability of specific language datasets for sentence retrieval and resource constraints remain practical challenges. Further exploration of prompt design and consideration of external factors could improve our methodology. Acknowledging these limitations is essential for a full interpretation of our results and the direction of future research.

## Acknowledgements

This work was supported by Leibniz Supercomputing Centre (LRZ), Munich Center for Machine Learning (MCML) and China Scholarship Council (CSC).

## References

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

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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

Md Akhter-Uz-Zaman Ashik, Shahriar Shovon, and Summit Haque. 2019. Data set for sentiment analysis on bengali news comments and its baseline evaluation. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Anirban Bhowmick and Abhik Jana. 2021. 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. NLP Association of India (NLPAI).

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

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. 2020b. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sandipan Dandapat, Sudeshna Sarkar, and Anupam Basu. 2004. A hybrid model for part-of-speech tagging and its application to bengali. In *International conference on computational intelligence*, pages 169–172. Citeseer.

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

Amitava Das and Björn Gambäck. 2014. 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. NLP Association of India.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

*for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Asif Ekbal and Sivaji Bandyopadhyay. 2007. A hidden markov model based named entity recognition system: Bengali and hindi as case studies. In *Pattern Recognition and Machine Intelligence: Second International Conference, PReMI 2007, Kolkata, India, December 18-22, 2007. Proceedings 2*, pages 545–552. Springer.

Asif Ekbal and Sivaji Bandyopadhyay. 2008a. Development of bengali named entity tagged corpus and its use in ner systems. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Asif Ekbal and Sivaji Bandyopadhyay. 2008b. A web-based bengali news corpus for named entity recognition. *Language Resources and Evaluation*, 42:173–182.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. 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. Association for Computational Linguistics.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. 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. Association for Computational Linguistics.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. 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. Association for Computational Linguistics.

Muntasir Hoq, Promila Haque, and Mohammed Nazim Uddin. 2021. Sentiment analysis of bangla language using deep learning approaches. In *COMS2*.

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

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. 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. Association for Computational Linguistics.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. 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. Association for Computational Linguistics.

Md Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4207–4218.

Md Zahurul Islam, Md Nizam Uddin, and Mumit Khan. 2007. A light weight stemmer for bengali and its use in spelling checker.

Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. 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. Association for Computational Linguistics.

Yanchen Liu, Timo Schick, and Hinrich Schtze. 2023. 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). Association for Computational Linguistics.

Ashis Kumar Mandal and Rikta Sen. 2014. Supervised learning methods for bangla web document categorization. *arXiv preprint arXiv:1410.2045*.

Munirul Mansur. 2006. *Analysis of n-gram based text categorization for bangla in a newspaper corpus*. Ph.D. thesis, BRAC University.

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

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. 2023. 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. Association for Computational Linguistics.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Jiaul H Paik and Swapan K Parui. 2008. A simple stemmer for inflectional languages. In *Forum for Information Retrieval Evaluation*. Citeseer.

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

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

Md Atikur Rahman and Emon Kumar Dey. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2):15.

Nils Reimers and Iryna Gurevych. 2019. 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. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. 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. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahout, Syed Ishtiaque Ahmed, Nabeel Mohammad, and Mohammad Ruhul Amin. 2023. 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. Association for Computational Linguistics.

Salim Sazzed. 2020. 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. Association for Computational Linguistics.

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 Francois 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. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

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

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. 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. Association for Computational Linguistics.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Md Ferdous Wahid, Md Jahid Hasan, and Md Shahin Alom. 2019. 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.

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.

Mengjie Zhao and Hinrich Schütze. 2021. 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. Association for Computational Linguistics.

# A Appendix

| | zero-shot | k=1 | k=3 |
|---|---|---|---|
| prompt | {text} Direct Aggression, Indirect Aggression, or No Aggression? | | |
| accuracy | 0.53 | 0.54 | 0.54 |
| macro avg | 0.17 | 0.18 | 0.18 |
| weighted avg | 0.38 | 0.38 | 0.38 |
| prompt | Evaluate the text: '{text}'. Would you categorize it as absence of aggression, mild aggression, or strong aggression? | | |
| accuracy | 0.18 | 0.23 | 0.23 |
| macro avg | 0.17 | 0.15 | 0.15 |
| weighted avg | 0.13 | 0.16 | 0.16 |
| prompt | In the context of '{text}', which category best captures its aggression level: absence of aggression, mild aggression, or strong aggression? | | |
| accuracy | 0.12 | 0.15 | 0.16 |
| macro avg | 0.1 | 0.15 | 0.16 |
| weighted avg | 0.06 | 0.11 | 0.12 |
| prompt | For the text: '{text}', ascertain its aggression scale: absence of aggression, mild aggression, or strong aggression? | | |
| accuracy | 0.19 | 0.21 | 0.21 |
| macro avg | 0.13 | 0.14 | 0.14 |
| weighted avg | 0.14 | 0.16 | 0.15 |
| prompt | From the following choices, which resonates with the theme of '{text}'? Options: No Intensity, Low Intensity, High Intensity | | |
| accuracy | 0.13 | 0.24 | 0.19 |
| macro avg | 0.1 | 0.17 | 0.15 |
| weighted avg | 0.12 | 0.26 | 0.2 |
| prompt | From the following choices, which resonates with the theme of '{text}'? Options: no intensity, low intensity, high intensity | | |
| accuracy | 0.23 | 0.28 | 0.27 |
| macro avg | 0.18 | 0.22 | 0.2 |
| weighted avg | 0.22 | 0.31 | 0.26 |
| prompt | In the context of the text '{text}', which of the following best describes its tone? Options: No Intensity, Low Intensity, High Intensity | | |
| accuracy | 0.14 | 0.2 | 0.15 |
| macro avg | 0.11 | 0.15 | 0.12 |
| weighted avg | 0.1 | 0.18 | 0.13 |
| prompt | Reflecting on the statement '{text}', which aggressive level does it resonate with: non-aggressive, slightly aggressive, or highly aggressive? | | |
| accuracy | 0.33 | 0.35 | 0.36 |
| macro avg | 0.2 | 0.2 | 0.17 |
| weighted avg | 0.19 | 0.2 | 0.24 |

Table 5: F1-score results with 8 prompt templates of Vio-Lens test using bloomz-3b model

| | zero-shot | k=1 | k=3 |
|---|---|---|---|
| prompt | The text displays [MASK] aggression: {text} | | |
| verbalizer | direct, indirect, none | | |
| accuracy | 0.36 | 0.35 | 0.36 |
| macro avg | 0.22 | 0.23 | 0.23 |
| weighted avg | 0.31 | 0.31 | 0.31 |
| prompt | Considering aggressive tendencies, this is [MASK]: {text} | | |
| verbalizer | overt, covert, absent | | |
| accuracy | 0.1 | 0.2 | 0.17 |
| macro avg | 0.07 | 0.17 | 0.14 |
| weighted avg | 0.03 | 0.19 | 0.15 |
| prompt | From an aggression perspective, the text is [MASK]: {text} | | |
| verbalizer | overt, covert, absent | | |
| accuracy | 0.12 | 0.22 | 0.2 |
| macro avg | 0.09 | 0.18 | 0.16 |
| weighted avg | 0.06 | 0.21 | 0.18 |
| prompt | The described behavior in {text} is [MASK] aggression. | | |
| verbalizer | explicit, implicit, neutral | | |
| accuracy | 0.24 | 0.36 | 0.35 |
| macro avg | 0.19 | 0.24 | 0.23 |
| weighted avg | 0.23 | 0.31 | 0.3 |
| prompt | The underlying theme in {text} is [MASK] aggression. | | |
| verbalizer | assaultive, indirect, peaceful | | |
| accuracy | 0.22 | 0.32 | 0.33 |
| macro avg | 0.18 | 0.21 | 0.21 |
| weighted avg | 0.21 | 0.28 | 0.29 |
| prompt | {text} is interpreted as [MASK] aggression. | | |
| verbalizer | assaultive, indirect, peaceful | | |
| accuracy | 0.51 | 0.49 | 0.51 |
| macro avg | 0.23 | 0.27 | 0.25 |
| weighted avg | 0.37 | 0.37 | 0.37 |

Table 6: F1-score results with 6 prompt templates of Vio-Lens test using mBert model

| | 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 7: Results in SentNoB test of BLOOMZ-3b and mBERT with hindi retrieval corpus.