ChakmaNMT: Low-resource Machine Translation On Endangered Chakma Language

Anonymous ACL submission

Abstract

The Chakma Language, spoken by the indigenous Chakma community mostly from Bangladesh and India, lacks digitalization and is at risk of extinction due to its limited linguistic resources. Developing a Machine Translation (MT) model for Chakma to Bangla and vice-versa could play a crucial role in preserving the language and bridging the cultural and linguistic gap between the two. In this paper, we have worked on MT between CCP-BN (Chakma-Bangla) and BN-CCP by introducing a novel dataset of 15,021 parallel samples and 42,783 monolingual Chakma samples. We also present a high-quality benchmark dataset of 600 parallel samples across Chakma, Bangla, and English. Additionally, we developed a transliteration system that converts the extremely low-resource Chakma script into Bangla to leverage existing Bangla pretrained models. We experimented with both traditional and state-of-the-art approaches, including statistical machine translation, neural machine translation, and in-context learning with LLMs. In our experiments, fine-tuning BanglaT5 with back-translation using transliterated Chakma achieved the highest BLEU scores of 17.8 (CCP-BN) and 4.41 (BN-CCP). We also found that, with our transliteration system, commercial LLMs achieve near state-of-the-art performance using significantly fewer examples.

Src.	[illegible] ।
Ref.	দেশকে সামনের দিকে এগিয়ে নিয়ে যেতে বেসরকারি সেক্টর খুবই গুরুত্বপূর্ণ । (To move the country forward, private sector is very important.)
NMT	দেশটিকে সামনে এগিয়ে নিয়ে যাওয়ার জন্য বেসিরকারি সেক্টরের যথেষ্ট ভূমিকা রয়েছে । (To move the country forward, there is a significant importance of the private sector.)

Figure 1: An example of translation (CCP-BN) by our best NMT model, BanglaT5 (Iterative Back-translation approach). English reference is shown in brackets.

1 Introduction

The Chakma people are the Indigenous tribes native to Bangladesh, the easternmost regions of India, and western Myanmar, and they speak the Chakma language, which belongs to the Indo-Aryan language family (Mohsin, 2013). There are 230,000 speakers of the Chakma Language in India (censusindia, 2011), 80,000 in Myanmar, and 483,299 in Bangladesh (Statistics, 2023). However, despite many Chakma-speaking people, there is little use of the Chakma writing script. According to Saikia and Haokip (2023), the Chakma language is in the category of "Definitely Endangered" language due to the way they are dispersed among India, Bangladesh, and the surrounding nations. Language is an integral part of identity and culture, and its loss could lead to the erosion of traditions, cultural practices, and community bonds. To revitalize the language, some measurements have been taken by the Bangladesh and Indian governments and NGOs. Pre-primary books have already been printed by the National Curriculum and Textbook Board of Bangladesh (NCT, 2024). In India, some textbooks (CAD, 2024) on the Chakma language can be found that are taught only in some schools.

Thus, the lack of communication between the Chakma community and mainstream Bangladesh exacerbates this cultural-linguistic divide. Our research aims to help bridge this gap through the development of MT models, promoting better linguistic understanding and interaction.

Till now, few works have been made to revitalize the Chakma language. Podder et al. (2023) has incorporated deep learning for Chakma character recognition and Pratap et al. (2023) has imple-

mented a language identification model for speech that can recognize 4000+ languages along with Chakma. However, other computational work such as Machine Translation is not explored yet. Such a system would not only enable automatic translation of content but also facilitate communication, and preserve cultural identity. On the other hand, the commercial large language models like GPT and Grok can recognize Chakma script and phonetics, they often fail to generate semantically correct Chakma sentences, where even simple translation tasks like “I want to go there" result in nonsensical Chakma words. Thus, our work aims to work on MT for the Chakma language and promote mutual understanding, and support the preservation against endangeredness.

Our contributions are as follows:

- Introduced the first Chakma-Bangla parallel corpus having 15,201 sentence pairs, a Chakma monolingual corpus of 42,783 samples, and a benchmark set of 600 triples of Chakma, Bangla, and English.
- Build a rule-based transliteration module enabling the usage of large pre-trained models
- Evaluated SMT, RNN, Transformer, and BanglaT5, where BanglaT5 with back-translation achieves the best BLEU scores of 17.8 and 4.41 in CCP-BN and BN-CCP.
- Analyzed the degraded BLEU score in BN-CCP due to orthographic inconsistency even between the language scholars highlighting the challenges of MT on extremely low-resource languages.
- Demonstrated the potential of in-context learning with GPT, enabled by our transliteration technique. The approach produced promising results even with as few as 400 examples, despite the model having very limited exposure to Chakma.

To our knowledge, this is the first MT system developed for Chakma, and we provide the novel dataset with a strong baseline using fine-tuned BanglaT5 with transliteration. Beyond technical contributions, our work is intended to support the preservation of the Chakma language by digitizing valuable linguistic resources and enabling cross-lingual access.

2 Related Works

As mentioned, there has been no record of working with a Machine Translation system for the Chakma language before. Also, no dataset containing the Chakma texts is done yet. However, works in other fields are found in the Chakma Language. Pratap et al. (2023) has implemented a speech recognition system where the Chakma language can be also detected with other thousands of languages of the world. They created a language identification framework that could recognize 4,017 languages including Chakma. On the other hand, to identify Chakma characters, Podder et al. (2023) created a dataset with 47,000 images for the 47 characters of Chakma. The authors suggested a novel SelfONN-based deep learning model named Self-ChakmaNet, which scored 99.84% accuracy on the test set.

Many efforts were made to work with other dominant Indo-Aryan languages using NLP. To identify Indo-Aryan dialect Subhash et al. (2024) has used the Deep Learning Ensemble Model with data augmentation. Furthermore, Baruah et al. (2021) incorporated Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) models to translate low-resource Assamese language to other Indo-Aryan(Indic) languages. Mumin et al. (2019) has implemented a Phrase-Based Statistical Machine Translation (PBMT) system between English and Bangla languages in both directions and Bangla was a low-resource dataset back then. Until recently, a large dataset containing 2.7M BN-EN pairs was published by Hasan et al. (2020). Transliteration has been shown to improve translation quality between closely related languages with different scripts, such as Hindi and Urdu Durrani et al. (2010). which also motivated us to develop our own rule-based Chakma-to-Bangla transliteration system.

Machine translation has been studied for many years, but the majority of the early research focused on high-resource translation pairs, such as French-English. However, Riza et al. (2016) presented multiple Asian language arrangements with limited resources. Two low-resource translation evaluation benchmarks were presented by Guzmán et al. (2019): Sri Lankan–English and Nepali–English. Existing publications have mostly investigated two approaches to enhance low-resource machine translation: semi-supervised learning by using monolingual (Gulcehre et al., 2015) data and a multilingual

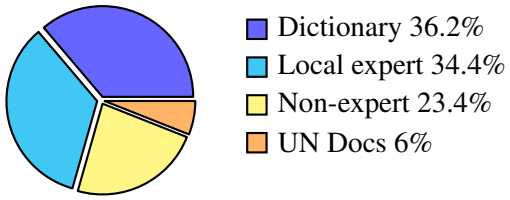

Figure 2: This chart shows the parallel dataset distribution between CCP-BN. Here, UN Docs=Comprised of 2 UN documents(UN-Convention on the Rights of Persons with Disabilities and UN-The Convention on the Rights of the Child), Dictionary=Word pairs collected from a dictionary, Local expert=Direct in-person translation by local experts, and Non-expert=Data collected building a custom website translated by common people. Our total Parallel data set is comprised of 15,021 CCP-BN pairs.

collaboration (Kocmi and Bojar, 2018) or cross-lingual transfer learning. Back-translation is also an effective approach as explored by Sennrich et al. (2016). Xu et al. (2019) discovered that using the right back-translation technique, rather than just adding more synthetic data, enhances translation performance.

3 Dataset Description

To collect a monolingual and parallel dataset for the endangered Chakma language, we visited many first-language Chakma scholars, local organizations, and typists from the hill-tracts region of Bangladesh as well as taking interviews with the scholars. Our data collection procedure is detailed below. The challenges and details of the interviews can be found in the appendix section A.1 and A.2.

3.1 Parallel Data

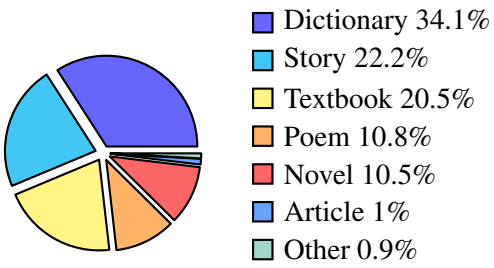

Figure 3: This chart illustrates the distribution of monolingual Chakma data based on content types that were collected from different sources. We collected 42,783 Chakma monolingual samples in total

Parallel Documents: Upon extensive searching, we found two online documents with both Bangla and Chakma translations: UN - Convention on the Rights of Persons with Disabilities (UnitedNation), and UN - The Convention on the Rights of the Child (resolution 44/25, 1989). The Bangla PDF versions were composed of images of each page of the original paper document. We used Tesseract OCR to extract the Bangla sentences, For alignment, we could not perform any automatic alignment similar to Hualign Varga et al. (2005) because they require a rich dictionary which was missing for the Chakma Language. Thus, we did manual alignment for both files and gathered 620 and 291 CCP-BN parallel pairs respectively. Moreover, we incorporated word pairs from the only dictionary as our parallel data which has 5,473 samples, although it was not enough for Hualign.

Manual Translation from experts: With an objective of collecting manually translated data from the local proficient people in the Chakma Language, we prepared some paper forms containing the Bangla sentences of a total of 10,000. Those sentences are collected randomly from BN-EN sentences from Hasan et al. (2020) having a word count between 2 and 8, where the probability of choosing sentences is highest with 4 and 5 words and decreases in both directions. We arranged the volunteering program for 3 days, where 7-10 people participated each day in Dighinala, Khagrachari of Bangladesh, and the participants were mostly young. To note that, the participants struggled with translations as many words from Bangla or English aren't used or don't have direct translations in Chakma. From the program, we successfully gathered 5,203 sentences of CCP-BN-EN pairs.

Manual Translation from crowdsourcing: We have also collected data from common people. First, we created a website where Bangla sentences were shown to people and asked to translate them into Chakma. We shared the website link through social media platforms. These Bangla texts are mostly common dialogues collected from a few sites [1] which already have English translations as well. As most of the people didn't know how to write in Chakma script, they were asked to write the translation using the Bangla transliteration of Chakma. Later, we converted them automatically into Chakma characters by the code that we built to convert CCP-BN and vice versa. After manual

[1] https://www.learnenglishfrombangla.com/2021/07/easily-learn-english-in-bangla-beginner.html, https://www.omniglot.com/language/phrases/bengali.php, and https://en.wikibooks.org/wiki/Bengali/Common_phrases

Dataset	Resource Name	Data Count	Total
Parallel Data	UN - Convention on the Rights of Persons with Disabilities (BN-CCP)	8647	15021
	UN - The Convention on the Rights of the Child (BN-CCP)	291	
	Dictionary app (Word-pairs) (BN-CCP)	5473	
	Translated data from crowdsourcing. (BN-EN-CCP)	3444	
	Translated data by expert (BN-EN-CCP)	5203	
Monolingual Data	Chakma from multiple local sources (CCP)	42783	42783
Evaluation Data	Translation from RisingNews Benchmark by Hasan et al. (2020) (BN-EN-CCP)	600	600

Table 1: Main sources of our parallel, monolingual, and evaluation datasets with their respective data counts.

verification and filtering, we have collected a total of 3,444 CCP-BN-EN paired sentences. Our transliteration codes are available on our github.

In the end, we had a total of 15,021 parallel BN-CCP data sentences, among them 8,647 had CCP-BN-EN language pairs. The overall process of collecting and refining data was very cumbersome and it took us several months to complete. The distribution of our data can be found in Figure 2.

3.2 Monolingual Data

We have managed to collect a good amount of monolingual data in comparison to parallel data, and the soft copies that we have collected mostly comprise poems, articles, stories, a few national textbooks, etc. We have also collected Indian textbooks, a Chakma Folktale app, and a Chakma Dictionary app written by Indian Chakma authors. Then, all of these contents of our sources that we have so far were first copied into separate docx files, which successfully maintained the various Chakma fonts used for those documents, but the fonts were in ASCII format and each of them was mapped with different ASCII encoding. Thus, we build a program that converts a common unified font, RebangUni[2], the first and only UTF-8 font, and we managed to convert for 7 ASCII fonts. Finally, we build a simple segmenter where each line is segmented based on 3 punctuations: '?', '!', and 'l'. After all of the processing, we have gathered 42,783 monolingual samples. In the figure 3, we displayed the distribution of our collected monolingual data. The table 8 and 9 contain all the names and necessary details of the files. The conversion codes are uploaded to our github repository and the ASCII fonts list is given in the appendix table 3. In addition to that, for our training, we gathered 150,000 Bangla and 150,000 English as monolingual data from the dataset by Hasan et al. (2020), where we choose the Bangla and English data in such a way that they are not parallel to each other.

[2]https://github.com/Bivuti/RibengUni.

3.3 Evaluation Data

To evaluate our models, we have meticulously prepared a benchmark dataset as well. We have first selected 500 BN-EN data randomly from the RisingNews Benchmark dataset by Hasan et al. (2020). They processed and filtered their dataset following the approaches of Guzmán et al. (2019), which makes it a standard quality dataset to work on. Moreover, since it has already bi-lingual pairs, additional translation into Chakma can make it possible as a Benchmark between English and Chakma as well. We provided these sentences to 3 different Chakma language researchers to translate who hadn't participated previously in translating our parallel data. We gave each person 200 sentences where 50 sentences were common for each of them. We make these 50 sentences common to each of them so that we can discuss and research further on the variances of translation of the same sentences. Thus, we have 600 samples for our Benchmark Dataset, which we name 'RisingNewsChakma', and it is an out-of-domain compared to our training data. In Table 1 we have shown the counts of all types of our data.

4 Models and Approaches

4.1 SMT and NMT

SMT, once the dominant approach before deep learning, has been shown to outperform NMT in low-resource settings Koehn and Knowles (2017). We therefore employed Phrase-based SMT Koehn et al. (2003) in our experiments to validate its effectiveness. For NMT, we experimented with both RNN and Transformer architectures. RNNs, particularly GRUs (Bahdanau et al., 2014), are more effective than transformers in low-resource settings

due to fewer parameters. We use Luong-style attention (Luong et al., 2015) for the RNN.

4.2 Transfer Learning

We explored transfer learning using pre-trained BanglaT5 (Bhattacharjee et al., 2023), given its linguistic proximity of Bangla to Chakma. As the model does not recognize the Chakma characters, we created a simple rule-based transliteration system that converts between Chakma and Bangla. The transliteration system was easily possible to build since almost all characters, except a very few, have a direct one-to-one mapping between Chakma and Bangla. There are other multilingual models such as mT5 ((Xue et al., 2021)), mBART (Liu et al. (2020)), but due to time and resource constraints, we focused only on BanglaT5.

4.3 Back-Translation

Back-translation is a semi-supervised method that uses monolingual data by translating it in the reverse direction, proving effective when parallel data is scarce (Burlot and Yvon, 2018; Karakanta et al., 2018). We applied the iterative strategy from Hoang et al. (2018) in both Chakma-Bangla and Bangla-Chakma directions.

4.4 Multilingual Training

Multilingual joint training can improve low-resource translation by leveraging cross-lingual patterns (Zhang et al., 2020; Johnson et al., 2017). We adopted a many-to-many strategy, adding 10,000 Bangla-English pairs (Hasan et al., 2020) into the training set and using language tags to guide translation across Chakma, Bangla, and English.

4.5 In-context Learning

We also explored In-Context Learning (ICL) using several of the GPT models via few-shot prompting getting inspired from Agrawal et al. (2022). Our objective of this approach is to investigate the capability of large pretrained LLMs to adapt to extremely low-resource languages with minimal examples.

5 Experimental Setup

We used the Moses toolkit[3] for our phrase-based SMT approach. The NMT experiments are conducted using Pytorch on Google Colab[4] using V100/L4 GPUs. For pre-processing, we adopted the normalization method introduced by Hasan et al. (2020) where we added some adjustments to the normalizer [5]. We applied the Beam search decoding strategy with a beam of width 5 for predictions. The maximum sequence length was capped at 128 tokens and gradient clipping was set at 1.0. We followed the SentencePiece (Kudo and Richardson, 2018) tokenizer for both vocabulary building and tokenization for SMT and NMT. With the SentencePiece we ran the vocabulary sizes(1000, 2000, 5000, 10000, and 20000) as hyper-parameter optimization. Various learning rates were tested(0.001, 0.005, 0.0001, and 0.0005). We considered the batch sizes at 8, 16, and 32. The number of training steps was experimented with 10,000, 15,000, and 20,000 steps. The warmup steps were varied between 0, 2000, and 4000 steps. We also did Label smoothing was various values(0.1, 0.2, 0.3, 0.4, 0.5).

For the RNN, we applied the open implementation of RNN [6] which incorporated the attention mechanism of Luong et al. (2015). Furthermore, we explored our models with 1, 2, and 4 RNN layers. We considered the hidden size and embedding size with values 512 and 1024. Numerous dropout rates are also being tuned(0.1, 0.2, and 0.3). We initialized RNN using a normal distribution with a mean of 0 and a standard deviation of 0.1.

To apply the transformer model, we followed the method introduced in Vaswani et al. (2017). We considered MarianNMT models from Huggingface. We used Glorot's (Glorot and Bengio, 2010) initialization to initialize the weights. We consider models with 1, 2, and 6 layers. Further, we explored models with 1, 2, and 6 attention heads and evaluated dropout rates of 0.1, 0.2, and 0.3. Also, we tested feed-forward hidden dimensions of 512 and 1024.

To experiment with the BanglaT5, we have fine-tuned the model by transliteration of Chakma characters to Bangla fonts and did hyper-parameter optimizations similar to Transformer. We also used the model for multilingual translation between CCP-BN-EN. We added a prefix tag of the target language to the input sentence, and we also oversampled for Chakma pairs to balance between all pairs since it is proven to increase performance (Johnson et al., 2017). For back-translation, we marked the

[3] https://www2.statmt.org/moses/

[4] https://colab.research.google.com/

[5] https://github.com/anonymous_for_now.

[6] https://github.com/bentrevett/pytorch-seq2seq/tree/main.

System	CCP-BN				BN-CCP			
	Dev		Test		Dev		Test	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
SMT	3.20	24.90	0.10	20.40	4.60	26.30	0.10	20.20
RNN	10.62	25.23	0.16	11.19	4.54	24.04	0.09	11.60
Transformer	2.85	26.60	0.37	19.36	1.42	25.75	0.20	25.79
BanglaT5	26.72	45.34	13.44	38.91	10.60	34.27	2.88	28.46
BanglaT5 (IBT +1it)	26.72	45.34	13.44	38.91	11.96	37.92	3.97	30.48
BanglaT5 (IBT +2it)	**28.31**	**49.01**	**17.80**	**49.19**	**11.67**	**38.49**	**4.41**	**31.33**
BanglaT5 (MNMT)	20.43	39.54	12.37	38.17	9.04	31.53	3.18	29.54
GPT 4.1 (ICL)	-	-	16.63	47.91	-	-	2.36	30.46
GPT 4.1 (ICL - non-transliterated)	-	-	0.41	19.09	-	-	0.21	18.24
GPT 4.1-mini (ICL)	-	-	9.74	40.67	-	-	1.81	28.48
GPT o4-mini (ICL)	-	-	11.50	42.85	-	-	1.91	29.80

Table 2: Performance on CCP-BN and BN-CCP translation using SMT, RNN, Transformer, and BanglaT5 with Iterative Back-Translation (IBT) and Multilingual (MNMT) training on parallel data. Test-set results from GPT 4.1, 4.1-mini, and o4-mini using in-context learning with 400 examples are also included.

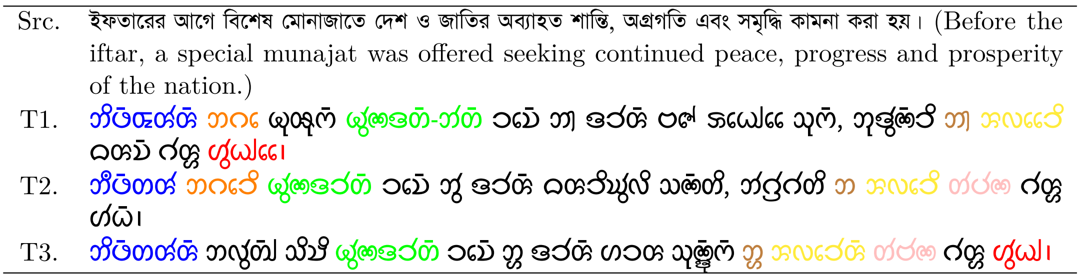

Figure 4: Several variations of spelling same word are shown by marking with same color in Chakma Language from our benchmark which 3 different language scholars translated from Bangla.

Bangla language as the source and Chakma as the target and used 50,000 monolingual samples during back-translation.

For the ICL experiment, we use GPT 4.1, 4.1-mini, and o4-mini. We didn't hyper-tune the decoding parameters of the GPT due to budget constraints, and used the default settings such as temperature as set to 1. We provided 0, 100, 200, or 400 examples of translations between BN and CCP in each prompt, and ask 20 sentences at once to be translated into the other language. A prompt sample is shown in the table 8. Note that we provided transliterated Chakma while using BanglaT5 and GPT models, while for vanilla training using SMT, RNN, and Transformer, we use Chakma characters directly.

We split our parallel set into train and dev sets containing 12,016 and 3,005 respectively. We have used the same split for all of our experiments. We have used our own RisingNewsChakma Benchmark as our Test set, which is an out-of-domain where the dev set is in-domain. We used sacreBleu(Post, 2018) and chrF(Popović, 2015) as the model's performance evaluation but used the sacreBleu as the metric for best model selection. For transliterated experiments, we measured the BLEU and chrF scores on transliterated. Finally, we share our training parameters and settings for NMT in table 4.

6 Results

Vanilla Training: While the models trained from scratch achieve at least a few BLEU scores in the Dev set, they struggle with BLEU scores in the Test set. All of these models: SMT and NMT outputs mostly random and nonsensical words resulting in a BLEU of not even 0.5 in the Test set. A probable reason is the domain mismatch our benchmark data and training set. Moreover, while previous work on endangered languages shows that RNN and SMT outperform Transformers, in our case Transformers beats with the slightest margin

both SMT and RNN in test sets. The results are shown in table 2.

Transfer Learning: Fine-tuned BanglaT5 significantly outperforms the models with vanilla training in both BLEU and chrF metrics across the CCP-BN and BN-CCP. For the CCP-BN translation, BanglaT5 achieves a BLEU score of 26.72 on the development set and a BLEU score of 13.44 on the test set, whereas vanilla models couldn't score even 0.5. In the same way, in the BN-CCP direction, BanglaT5 attains the best performance with a BLEU score of 10.60 and 2.88 on the dev set and test set respectively, which is higher than the vanilla models. The most probable reason behind the BanglaT5's superior performance is that the Chakma language is similar to the Chittagonian dialect(A little different than regular Bangla spoken by the people from Chittagong, Bangladesh) and many words are directly the same as Bangla, which makes the model easily to adapt and understand Chakma. This shows that a simple transliteration into a close dominant language can provide significant benefits for low-resource languages. Moreover, we found that the fine-tuned BanglaT5 can do zero-shot translation between EN-CCP despite the model being trained on BN-CCP data using BanglaT5. Although, we did not do the benchmark testing between EN-CCP, an example shown in table 7 shows that the model-produced translation conveys parts of the original meaning despite inaccuracies.

Back-translation: We did iterative back-translation only on BanglaT5. We see that the performance has increased further after applying back-translation for both metrics after each iteration. Particularly after 2nd iteration, we get the highest BLEU score among all other approaches in both BN-CCP and CCP-BN. We gained more than 1.5 BLEU scores in BN-CCP and even more than 3 BLEU in CCP-BN compared to the simple fine-tuned model. This improvement is coherent with back-translation experiments from most works on low-resource data, although Feldman and Coto-Solano (2020) suggested that back-translation degraded the performance due to out-of-domain monolingual data, where CCP monolingual data was out-of-domain in our case.

Table 5: shows how the performance of different input ratios of our monolingual data to training data can affect the performance of the Back-translation approach. Our outcomes are similar to Hoang et al. (2018): more the ratio of monolingual data, the higher BLEU score. We found that, for the CCP-BN, if we increase the ratio of input data to 1:4, it gives the highest BLEU score of 17.80. Similarly, for the BN-CCP direction, the performance also improves highest at a ratio of 1:4 achieving the best BLEU score of 4.41. These results show that having larger monolingual data is beneficial for low-resource tasks.

Multilingual Training: Table 6 shows our result on multilingual training. For the BN-CCP and CCP-BN translation, the BLEU scores are 3.18 and 12.37 respectively on the Test set, which is slightly higher than fine-tuned approach but lower than the back-translation approaches. This result is similar to findings from Guzmán et al. (2019) between Nepali-English that the multilingual approach performs better than the bilingual approach.

In-context Learning: We showed the results of ICL in table 7. There is a clear increase in BLEU score as we provide more examples in the prompt for both directions. The highest we achieved in BN-CCP in the test set is a BLEU of 2.36 using GPT 4.1. Although the scores are lower than the IBT approach, but particularly, in CCP-BN we get closer to the best scoring IBT with just 400 examples. Between the smaller models: 4.1-min and o4-mini, the reasoning model, o4-mini, tends to perform slightly better in both BN-CCP and CCP-BN. The score is lowest if no examples are provided at all, which clearly shows the GPT lacks understanding the Chakma language. Even when 400 examples are passed in the prompt but using non-transliterated Chakma the scores also falls down even under BLEU 0.5 in BN-CCP (Table 2). This also shows that ICL is ineffective if it doesn't recognize the language. Overall, the results shows the potential of using LLMs using ICL with the fewest data, but it comes with a cost that the model can be very large and commercial. On the other hand, BanglaT5 is significantly smaller but requires more data.

Transliteration Assessment: Since our best approaches rely on transliteration between Chakma and Bangla, it is crucial to assess the quality of transliteration. It can be stated that a transliteration is reliable if we achieve a good score in a round-trip of transliteration. On the benchmark sentences, our transliteration system scores 41.21 BLEU and 79.32 chrF on the round-trip from BN to CCP to BN, and scores BLEU of 34.58 and chrF of 77.74 on CCP to BN to CCP. Some of the scores are lost in the round-trip as few characters don't have a

direct mapping between Chakma and Bangla (we have mapped these few characters based on our intuition and shown in table 5). From our point of view, these scores are reliable while there are still scopes for improving the transliteration system.

BN-CCP vs CCP-BN: We find that there are significant performance gap between BN-CCP and CCP-BN by all the models. The CCP-BN excels in the Test set even when no examples are provided to the GPT models in the ICL experiments scoring BLEU of 14.56. On the other hand, our best models scored highest of 4.41 of BLEU score in BN-CCP. Beside the domain mismatch issue, an obvious reason is that the pre-trained models have already a vast knowledge on the Bangla Language since it at least the models excel while translating into Bangla.

Spelling Incosistency in CCP: The incosistency in spellings found in CCP is a major reason for low getting scores in the BN-CCP. We measured the Inter-Annotator Agreement between the participants for the benchmark set based on the 50 sentences that was common to them. We calculated the BLEU and chrF between each of them in both directions and averaged the score, which turns out also surprisingly as low as 4.48 BLEU but the chrF was 38.82. This high imbalance between the BLEU and chrF indicates the inconsistency of the spellings. Table 4 shows a translation example by those language experts. In the table, we see that different spellings are used for the same word and even the simplest word, marked with brown color, is written differently by each of them. This is a major concern in the Chakma Language and the cause behind these inconsistencies is the lack of established grammar, books, and literature in the Chakma language.

7 Limitations and Future work

Although our work contributes to the revitalization of the Chakma language, there are several limitations. Firstly, the size of our dataset is small compared to the others, limiting the generalization of the models. Moreover, the scarcity of linguistic resources and Chakma language experts made data collection and validation challenging and impacted our dataset's diversity. As mentioned our best approaches are dependent on transliteration, which still has room for improvement. In case of using pre-trained models, we are only experimented with BanglaT5. Many other multilingual pretrained LLMs exist, which opens the case to explore. In case of ICL, we experimented only using GPT models is which due to our limited budget and resources. Finally, future research should focus on expanding datasets, enhancing transliteration techniques, and deeper collaboration with native speakers.

8 Conclusion

In this paper, we introduce the first Machine Translation parallel dataset for the endangered indigenous language Chakma with the help of rare experts in the Chakma language, and by crowdsourcing in the Chakma-speaking communities. We also release Chakma monolingual data collected from Chakma resources from Bangladesh as well as from India. We created a transliteration system that enabled us to take advantage of existing Bangla knowledge in pretrained models. We applied multiple popular machine translation methods on Bangla to Chakma and Chakma to Bangla translation, such as SMT, NMT, and in-context learning. We found that pretrained BanglaT5 with the help of back-translation achieved the highest BLEU score in our carefully curated benchmark dataset. Additionally, our transliteration method opens up the possibility of using pretrained LLMs to create more synthetic training data for Chakma and lift it out of endangered status. We believe our method and dataset will pave the way for other NLP tasks for highly low-resource and endangered languages.

References

2024. Chakma textbooks by chakma autonomous district council(cadc). CADC.

2024. Textbook of small ethenic group(chakma) for pre-primary. NCTB.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *Preprint*, arXiv:2212.02437.

Dzmitry Bahdanau, Kyunghyun Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating

low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

censusindia. 2011. District census handbook lawngtlai. Office of the Registrar General.

Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-Urdu machine translation through transliteration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 465–474, Uppsala, Sweden. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Y. Bengio. 2015. On using monolingual corpora in neural machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin Mubasshir, Md Hasan, Madhusudan Basak, Mohammad Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alina Karakanta, Jon Dehdari, and Josef Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54. Association for Computational Linguistics. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Langauge Technology (HLT-NAACL 2003) ; Conference date: 27-05-2003 Through 01-06-2003.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Amena Mohsin. 2013. Language, identity and state. In Naeem Mohaiemen, editor, *Between Ashes and Hope: Chittagong Hill Tracts in the Blind Spot of*

Bangladesh Nationalism, page 158. Drishtipat Writers' Collective.

Mohammad Mumin, Md Hanif, Muhammed Iqbal, and Mohammed J Islam. 2019. shu-torjoma: An english-bangla statistical machine translation system. *Journal of Computer Science*, 15:1022–1039.

Kanchon Kanti Podder, Ludmila Emdad Khan, Jyoti Chakma, Muhammad E.H. Chowdhury, Proma Dutta, Khan Md Anwarus Salam, Amith Khandakar, Mohamed Arselene Ayari, Bikash Kumar Bhawmick, S M Arafin Islam, and Serkan Kiranyaz. 2023. Self-chakmanet: A deep learning framework for indigenous language learning using handwritten characters. *Egyptian Informatics Journal*, 24(4):100413.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Mamdouh Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1, 000+ languages. *ArXiv*, abs/2305.13516.

General Assembly resolution 44/25. 1989. *Convention on the Rights of the Child*, adopted and opened for signature, ratification and accession by general assembly resolution 44/25 of 20 november 1989 edition. United Nations.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Jonali Saikia and Mary Kim Haokip. 2023. Language endangerment with special reference to chakma. 3.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *Preprint*, arXiv:1511.06709.

Statistics. 2023. Bangladesh bureau of statistics. 2021. "Table A-1.4 Ethnic Population by Group and Sex".

Paliwal Mohan Subhash, Kavitha C.R., Deepa Gupta, and Vani kanjirangat. 2024. Indo-aryan dialect identification using deep learning ensemble model. *Procedia Comput. Sci.*, 235(C):2886–2896.

UnitedNation. *Convention on the Rights of Persons with Disabilities and Optional Protocol*. UN.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. *In Proceedings of the Recent Advances in Natural Language Processing (RANLP 2005)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. *Analysis of Back-Translation Methods for Low-Resource Neural Machine Translation*, pages 466–475.

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.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595, Online. Association for Computational Linguistics.

A Appendix

A.1 Challenges in Chakma Language

Chakma is predominantly an oral language. Very few people can read and write on Chakma Script. This leads to a primary problem: a lack of resources purely written in Chakma. Thus, it is very rare to find any data online, rather we need to visit various language experts, and typists (Usually, language experts are not good at typing. So they pass their handmade copies to typists to make digital copies of their writings), and some private organizations working on language preservation to collect the data. Another great challenge is not having an established grammar. Historically, language experts tend to write their way without publishing their standard of grammar, especially the local shamans,

leading to a variety of forms of spellings and structures. Moreover, from various meetings, disagreements between the language experts acted as a huge obstacle to standard grammar formation. There are a lot of variations in the grammar between Indian and Bangladeshi language scholars as well. The final challenge was that most of the documents were written using various ASCII fonts because the only UTF-8 font (RibengUni) was introduced very recently and usage has increased from offline to social media platforms now. Thus, we need to unify those fonts into a single font, RibengUni.

ASCII Font list of Chakma
BivunabaKhamaC
BijoygiriDPC
Udoy Giri
Alaam
Arjyaban
Chakma(SuJoyan)
Punong Jun

Table 3: ASCII Font list of Chakma documents found in our data sources which were converted into RibengUni Font (UTF-8).

Bangla Characters	Direction	Chakma Characters
শ, ষ	→	𑄪
ড়, ঢ়	→	𑄫
ঋ	→	𑄬
ৎ	→	𑄳
ওআ	←	𑄀

Figure 5: Our mapping of characters between Chakma and Bangla that don't have a direct map.

A.2 Interviews with Chakma Language Experts

We interviewed several scholars in Bangladesh to discuss the variants, for example, the number of characters, diacritics, rules, spelling patterns, etc. The scholars include Arjya Mitra, Injeb Chakma, Ananda Mohon Chakma, and Sugata Chakma. However, almost all of them suggested following the rules maintained by the members of the National Curriculum and Textbook Board of Bangladesh involved in writing the Chakma books for the pre-primary levels because their rules will be followed eventually. The most important rule from them that we followed in our transliteration codes from Bangla to Chakma, is that the core grapheme cannot have more than one diacritic attached to a consonant or a vowel. However, in India, this restriction is not maintained, rather more than one diacritic is seen frequently in their documents.

Parameter	RNN	Trans.	BT5
Max Epochs	-	-	5
Max Train Steps	20000	20000	-
Warmup Steps/Ratio	4000	4000	0.1
Learning Rate	0.0005	0.0001	0.0005
Batch Size	16	32	16
Max Length	128	128	128
Optimizer	adam	adam	adam
Vocab size	2000	10000	-
Beam width	5	5	5
Clip gradient	1.0	1.0	-
Label Smoothing	0.2	0.5	0.3
d_model	-	512	-
dropout	-	0.2	-
layer_dropout	-	0.1	-
att_heads	-	1	-
ffn_dim	-	512	-
blocks	-	6	-
rnn_dropout	0.3	-	-
layer_normalization	True	-	-
layers	1	6	-
word_embedding	512	-	-
hidden_embedding	1024	-	-
weight_decay	-	-	0.01

Table 4: Best hyper-parameter settings and other parameters used for our training of RNN, Trans(Transformer), and BT5(BanglaT5).

Ratio	BLEU	chrF
CCP-BN 1:1	16.4	46.39
CCP-BN 1:2	16.78	48
CCP-BN 1:4	**17.80**	**49.19**
BN-CCP 1:1	3.49	30.16
BN-CCP 1:2	3.6	30.37
BN-CCP 1:4	**4.41**	**31.33**

Table 5: Performance of back-translation with different monolingual-to-parallel data ratios using BanglaT5.

Evaluation Metric	BN-CCP		CCP-BN		EN-CCP	CCP-EN
	Dev	Test	Dev	Test	Test	Test
BLEU	9.04	3.18	20.43	12.37	1.17	6.46
chrF	31.53	29.75	39.54	38.17	23.10	27.69

Table 6: Performance of Multilingual training which includes Chakma, Bangla, and English using BanglaT5 model.

No of Examples	GPT-4.1				GPT-4.1-mini				o4-mini			
	CCP-BN		BN-CCP		CCP-BN		BN-CCP		CCP-BN		BN-CCP	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
0	14.56	46.49	0.48	18.41	8.61	38.91	0.87	22.63	11.29	42.51	0.61	24.02
100	16.28	47.61	1.81	28.58	9.10	39.79	1.28	27.12	11.54	42.72	1.51	28.86
200	17.43	47.86	2.22	29.95	10.24	40.63	1.47	27.59	11.62	42.92	1.71	28.05
400	16.63	47.91	2.36	30.46	9.74	40.67	1.81	28.48	11.50	42.85	1.91	29.80

Table 7: Performance of In-context Learning in Test set using GPT 4.1, 4.1-mini, and o4-mini for by few-shot prompting on translation between Chakma and Bangla.

	Src. 1	জাতিসংঘের সংস্থাগুলো এবং প্রযুক্তি বিশেষজ্ঞদের কাছ থেকে জরুরি প্রস্তুতি
	Ref. 2	[illegible] (UN agencies and technical experts on emergency preparedness)
	Src. 1	জাতিসংঘের সংস্থাগুলো এবং প্রযুক্তি বিশেষজ্ঞদের কাছ থেকে জরুরি প্রস্তুতি , বসবাসযোগ্যতা এবং দ্বীপের সুরক্ষার বিষয়ে কোনো ছাড়পত্রের জন্য অপেক্ষা না করে রোহিঙ্গাদের সেখানে স্থানান্তরিত করায় সরকার তার বারবার করা প্রতিশ্রুতি ভঙ্গ করেছে ।
	Ref. 2	[illegible] (The government has also reneged on repeated promises to await clearance from UN agencies and technical experts on emergency preparedness, habitability, and safety of the island before relocating Rohingya there.)
SMT	Pred. 1	[illegible] । (UN ..... [Rest of the output is gibberish and not semantically meaningful] ...)
	Pred. 2	[illegible] (UN's General Secretary..... [Rest of the output is gibberish and not semantically meaningful] ...)
RNN	Pred. 1	[illegible] (UN children and others)
	Pred. 2	[illegible] (UN all and others and why)
Trans-former	Pred. 1	[illegible] । (UN's UN's UN's UN's help help.)
	Pred. 2	[illegible] (UN's UN's all country country country country autistic person's rights rights rights and and and and their their their own own own country country country)
BanglaT5 (IBT)	Pred. 1	[illegible] (Preparedness required from UN agency and technical expert persons)
	Pred. 2	[illegible] । (The government has reneged on repeated promises to await clearance from UN agencies and technical experts on emergency preparedness, habitability, and safety of the island relocating Rohingya there.)
GPT 4.1 (ICL)	Pred. 1	[illegible] [Note that the translation is almost similar to BanlgaT5 and reference, and have several gibbrish or incorect spellings but semantically meaningful]
	Pred. 2	[illegible] । [Note that the translation is almost similar to BanlgaT5 and reference, and have several gibbrish or incorect spellings but semantically meaningful]

Figure 6: Examples of prediction by different approaches on short and long sentences on BN-CCP.

CCP-BN	Src.	[illegible]
	Ref.	৬ লাখেরও বেশি রোহিঙ্গা যাদের বেশিরভাগই নারী ও শিশু, বাংলাদেশে পালিয়ে গেছে এবং অজানা সংখ্যক অভ্যন্তরীণভাবে বাস্তুচ্যুত হয়েছেন, যাদের খাদ্য, পানি ও আশ্রয়ের অপর্যাপ্ততা রয়েছে। (As over 6 lakh Rohingya, mostly women and children have fled to Bangladesh, and an unknown number remain internally displaced with limited access to food, water, and shelter.)
	Pred.	৬ লাক্ষার উপর রোখিঙ্গা যারা বেশির ভাগ নারী ও শিশু, বাংলাদেশে ধেতে এষৌ ও কুক্ষিগতদের মধ্যে বাড়ির দরজাহারা হয়েছে তাদের খাবার পানি ও অরা সব ধরনের নিরাপত্তার জন্য । (As over 6 lakh Rohingya, mostly women and children have fled to Bangladesh, and internally displaced from home their limited access to food, water, and for every kind of security.)
Zero shot (EN-CCP)	Src.	Bangladesh and Finland have agreed to work together on the issue of world-wide climate change.
	Ref.	[illegible]
	Pred.	[illegible] Finel [illegible] । (Bangladesh and Finel nation together work together on world education exchange)

Figure 7: Showing an example of prediction done by BanglaT5 (IBT) on CCP-BN and a zero-shot translation on EN-CCP trained on BN-CCP. For CCP-BN, we mark the wrong words as red, and in the zero-shot translation, we mark the same context words with the same color.

You are given translation examples from Chakma to Bangla below:

Chakma Example 1: মরে নুয়া গরি পিজোর্ ন গরিবে
Bangla Example 1: আমাকে পুনরায় জিজ্ঞাসা করবে না ।

Chakma Example 2: গির্গিরানা
Bangla Example 2: কাঁপা

...
...

Chakma Example K: এ সভাগুনৎ কোরাম্ পুরেবাৎয়াই সরিক্ রাস্ট্রআনির্ তিনভাগর্ দিভাগ্ হাজির্ থা -পরিব ।
Bangla Example K: অংশগ্রহণকারী রাষ্ট্রের দুই -তৃতীয়াংশ উপস্থিতি ফোরাম হিসাবে বিবেচিত হবে ।

Provide the Bangla Translation of the Chakma provided below. Only provide the translation and do not output anything else.

Chakma Test 1: বলানয়ায়ান্ গুঝিলৌো সুনানু মন্ত্রি কোহেয়দে , ধয় কত্তেদি সুনজুক্কানেদয়ই তে কাতার্ বেরা জেব ।
Chakma Test 2: সুনানু মন্ত্রা সেখ্ খাসিনা জধা ভিচ্চে আরব্ আমিরাদৎ (ইউএই) তিন্ দিনর্ সরাকারি পর্ভাচ্ বিদি এচ্চে রেদোৎ দেঝৎ লুহ্ংগেগি ।
...
...
Chakma Input 20: এ আলহটয়্ সুক্রবার্ পুলিঝর্ বেগ দাঙর্ চোক্ দিয়েবো (আইজিপি) জাবেদ্ পাটোয়ারি কোয়াহেদ , ভুক্ক্ গরিয়ে কিঐো জনি তুল্বিচ্ ন চাহ্না , সালেন্ পুলিসসুনে নিজে গিরোহ্চ্ হৈানে তুল্বিচ্ চাহ্বাক্ ।

Figure 8: The prompt format that we used for our ICL experiment.

Title	Content	Samples
Ajanir dajan firana.docx	Story	206
Amader-Bari-2.pdf	Story	12
Amader-Bari-3.pdf	Story	23
Amader-gaye-dewar-pinon.pdf	Story	10
Amar-Charar-Boi.pdf	Poem	123
Amlokir-Gach.pdf	Story	27
Article 3rd Jamachug.docx	Story	194
Article 4th Furamon.docx	Story	194
Article 5th Pawr Murah.docx	Story	191
Bang-O-Puti-mach.pdf	Story	11
Banor-Berate-Eseche.pdf	Story	35
Banorer-Marfa-khaowa.pdf	Story	10
Bashir-soor.pdf	Story	9
Bie-Bari.pdf	Story	28
Bijhu.pdf	Story	28
Binoy Bikash Talukder20.docx	Poem	647
Binoy Dewan.docx	Poem	2004
Bizute-Berano.pdf	Story	12
Bone-Gie-Gach-Kata.pdf	Story	30
Boner-Mama.pdf	Story	11
Chader-Buri.pdf	Story	28
Chakma Dictionary app	Other	14928
Chakma Folktales app	Story	3765
Chakma Love song Uvagit.docx	Story	13
Chakma Text Book For Class-IV 2010 (IN Govt).docx	Textbook	1088
Chakma Text Book for Class-II 2010 (IN Govt).docx	Textbook	490
Chakma Text Book for Class-III 2010 (IN Govt).docx	Textbook	561
Chakma Text Book for Class-V 2010 (IN Govt).docx	Textbook	940
Chakma Text Book for Class-VI 2010 (IN Govt).docx	Textbook	1543
Chakma Text Book for Class-VII 2010 (IN Govt).docx	Textbook	1858
Chakma.docx	Article	136
Charar Boi-Chakma-Pages.pdf	Poem	31
Cijir Orago Boi-Chakma-Pages.pdf	Other	71
Cijir Talmiloni Kodatara-Chakma-Pages.pdf	Other	45
Cycle-e-Bazare-Jawa.pdf	Story	33
Dhanpudi.doc	Story	1278
Dudur-Kanna.pdf	Story	40
Dui-Bandhobir-Kotha.pdf	Story	16
Ghara Poja pire-Chakma-Pages.pdf	Other	4
H.F.Miller's Rangakura.docx	Story	90
Hotat-Agun.pdf	Story	12
Iskulo Akto-Chakma-Pages.pdf	Other	5
Jhimit-Ekhon-Bhalo.pdf	Story	42
Jhogra-Kora-Valo-Noi.pdf	Story	42
Kalo-and-Forshar-Kotha-1.pdf	Story	22
Kanamachi-Khela.pdf	Story	13
Karo-bipode-hasa-thik-na.pdf	Story	15
Kolar-Kotha-1.pdf	Story	11
Korgosher-sobji-bagan.pdf	Story	12
Lairang-er-nodi-par-howa.pdf	Story	13
Lao-er-Desh-Vromon.pdf	Story	44
Laz-kata-Banor.pdf	Story	12
Lobh-kora-valo-na.pdf	Story	16

Table 8: Names of the sources of our Chakma monolingual data with details (Part 1).

Title	Content	Samples
Mamar-Bari.pdf	Story	19
Mayer-Upadesh-1.pdf	Story	19
Meghla-Akash.pdf	Story	22
Mitar-Fuler-Bagan-1.pdf	Story	10
Moina-Pakhi-1.pdf	Story	16
Monar Sabon-Chakma-Pages.pdf	Story	36
Moni-Malar-Kotha-.pdf	Story	22
Monir-shopno-dekha.pdf	Story	14
Morog-Jhuti-Fool.pdf	Story	25
My Legha by Injeb Chakma.doc	Story	727
Nada-bhet-math for class I (IN Govt Tripura).docx	Textbook	878
Nanarakam-ghor.pdf	Story	14
Nirapod-pani-pan-korbo.pdf	Story	13
Ojhapador Chora-Chakma-Pages.pdf	Poem	30
Paka-Lichu.pdf	Story	19
Porichoy.pdf	Story	16
Projapoti-Ronger-Kotha.pdf	Story	12
Puti-Macher-Fal.pdf	Story	13
Rangdhanu.pdf	Story	20
Ranjuni for Class I (IN Govt) Tripura.docx	Textbook	1459
SRM 1st P. Bargang.docx	Poem	156
SRM 1st R. Krisnachura.docx	Poem	149
SRM 2nd P. Belwa Pawr.docx	Poem	259
SRM 2nd R. Chadarok.docx	Poem	76
Sanye-Pidhe-.pdf	Story	6
Shikkha Boi2017.docx	Poem	722
Shing-Macher-Kata.pdf	Story	36
Shiyal-er-Khang-Garang-Bazano.pdf	Story	19
Shrout.pdf	Story	8
Sial-mamar-school.pdf	Story	14
Sukorer-pat-batha-1.pdf	Story	12
Surjyer-Manush.pdf	Story	21
Tanybi.doc	Story	79
Tarum A Ranjuni-Chakma-Pages.pdf	Other	16
Teen-bondhur-golpo.pdf	Story	13
Text-Book-Chakma-pdf.pdf	Story	1405
Thurong-Barite-Raja.pdf	Story	43
Tin-bondhur-gacher-kotha.pdf	Story	15
Tiya-Pakhi-1.pdf	Story	23
chakma novel hlachinu.docx	Novel	1571
chedon akkan(10).pdf	Article	103
diarrhea-hole-ki-Korbo.pdf	Article	18
ghila khara class 3 p. 62.docx	Story	133
kajer-Kotha.pdf	Story	11
kochpanar rubo rega.docx	Story	151
mle- 2 ananda babu.docx	Poem	174
tin fagala-1.docx	Novel	1765
.doc	Other	170
.docx	Novel	1209

Table 9: Names of the sources of our Chakma monolingual data with details (Part 2).