Exploring Cross-Lingual Knowledge Transfer via Transliteration-Based MLM Fine-Tuning for Critically Low-resource Chakma Language

Adity Khisa
IIT, University of Dhaka
bsse1334@iit.du.ac.bd

Nusrat Jahan Lia
IIT, University of Dhaka
bsse1306@iit.du.ac.bd

Tasnim Mahfuz Nafis
IIT, University of Dhaka
bsse1327@iit.du.ac.bd

Zarif Masud
Toronto Metropolitan University
zarif.masud@gmail.com

Tanzir Pial
Stony Brook University
tpial@cs.stonybrook.edu

Shebuti Rayana
State University of New York at Old Westbury
rayanas@oldwestbury.edu

Ahmedul Kabir
IIT, University of Dhaka
kabir@iit.du.ac.bd

Abstract

As an Indo-Aryan language with limited available data, Chakma remains largely underrepresented in language models. In this work, we introduce a novel corpus of contextually coherent Bangla-transliterated Chakma, curated from Chakma literature, and validated by native speakers. Using this dataset, we fine-tune six encoder-based transformer models, including multilingual (mBERT, XLM-RoBERTa, DistilBERT), regional (BanglaBERT, IndicBERT), and monolingual English (DeBERTaV3) variants on masked language modeling (MLM) tasks. Our experiments show that fine-tuned multilingual models outperform their pre-trained counterparts when adapted to Bangla-transliterated Chakma, achieving up to 73.54% token accuracy and a perplexity as low as 2.90. Our analysis further highlights the impact of data quality on model performance and shows the limitations of OCR pipelines for morphologically rich Indic scripts. Our research demonstrates that Bangla-transliterated Chakma can be very effective for transfer learning for Chakma language, and we release our dataset[1] to encourage further research on multilingual language modeling for low-resource languages.

1 Introduction

Large Language Models (LLMs) have transformed the Natural Language Processing (NLP) world through unsupervised pre-training using large corpora of unlabeled data. Since labeled data are not required, LLMs can take advantage of the huge text corpora available in the public domain. For example, even first-generation language models such as BERT use a corpus of 3.3 billion English words (Devlin et al., 2019), while more recent LLMs use multiple massive corpora such as RedPajama (Weber et al., 2024) scraped from the web with hundreds of trillions of tokens. However, the sheer volume of data required for pre-training LLMs poses a challenge for low-resource languages even without labels, as seen in some recent works using datasets of 15 million words for Māori (James et al., 2022), 332 million tokens for Swahili (Conneau et al., 2020), and 108 million tokens from 11 African languages for AfriBERTa (Ogueji et al., 2021). Compared to the trillions of tokens available in high-resource languages, these million-scale corpora are minuscule. Consequently, training LLMs with low-resource corpora does not yield good results, as upon encountering new vocabulary, expressions, or culturally specific semantics, the models struggle to utilize their training patterns for accurate understanding and generation (Zhong et al., 2024).

To address this limitation, researchers have explored knowledge transfer, from LLMs trained on high-resource languages through Masked Language Model (MLM) fine-tuning on the comparatively lower resource language corpus (Fernando and Ranathunga, 2025). Muller et al. (2020) further showed that we can leverage the same transfer learning benefit through transliteration when the two languages do not share a script. They achieved a significant performance gain for Uyghur (105K sentences) and Sorani Kurdish (380K sentences) transliterated into the Latin script, compared to pre-training on those data in their original script alone.

[1] https://github.com/adity1234567/Chakma-MLM-Dataset.git

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 280–291
December 23, 2025

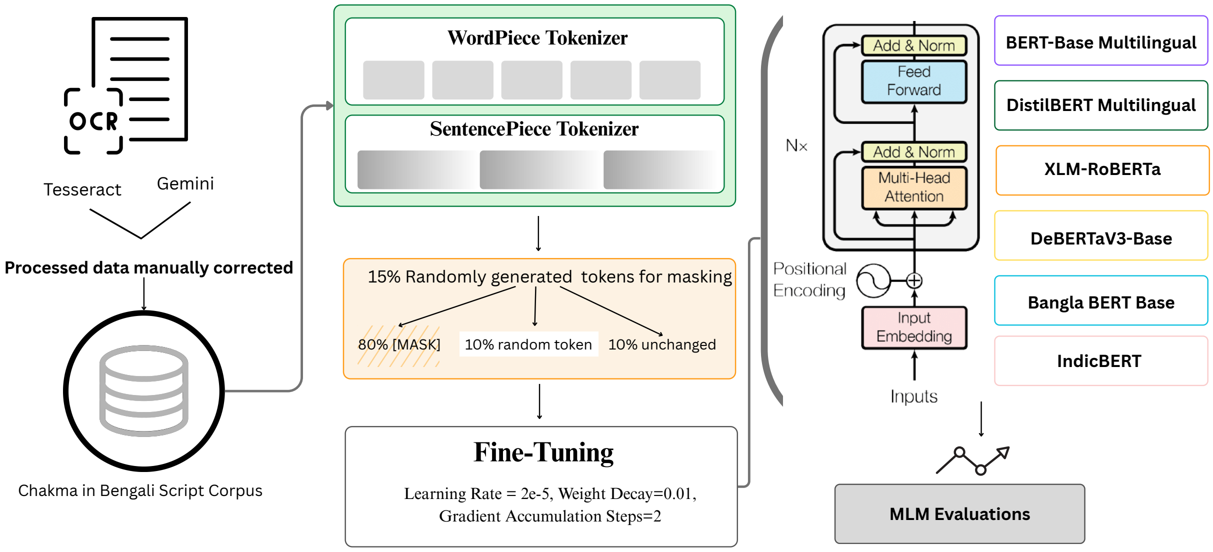

Figure 1: Overall workflow of OCR-based data curation, manual correction, and MLM fine-tuning for Bangla-transliterated Chakma language model

Chakma is an Indo-Aryan language, used as a first language by roughly one million people from the Chakma community living across parts of Bangladesh, India and Myanmar (Chakma Autonomous District Council, 2025). Although Chakma has its own script Ojhā Pāṭh, a considerable portion of Chakma literature is produced in Bangla transliteration (Brandt, 2018). Chakma remains a low-resource language with data scarcity both in its original script and Bangla transliteration (Chakma et al., 2024). At the same time, Bangla script is regularly used in training of multilingual LLMs like mBERT (Pires et al., 2019). In this context, our work shows that Bangla-transliterated Chakma dataset can yield moderately strong performance through MLM fine-tuning. Following Muller et al. (2020)'s idea for transliteration, we use Chakma text transliterated in Bangla for MLM-tuning multiple LLMs that are pre-trained on Bangla. Since, to the best of our knowledge, no contextually coherent Bangla-transliterated Chakma corpus exists, we have curated a novel corpus from Chakma books containing 4,570 manually validated sentences to run our experiments.

Our major contributions are as follows:

- We develop a novel Bangla-transliterated Chakma dataset, curated from images sourced from four books of Chakma literature using Tesseract OCR, comprising a total number of 6,353 sentences, of which 4,570 have been manually corrected.

- We show that language models can learn low-resource languages via MLM fine-tuning on the script of a related language, as demonstrated using Bangla script to fine-tune a Chakma model.

- We demonstrate how data quality impacts model performance, showing that better OCR for Bangla script, compatible with Bangla-transliterated Chakma, can significantly improve transfer learning for the Chakma language.

2 Related Works

In this section, we discuss four key areas that inform our work: multilingual NLP, model adaptation and quantization, tokenization and morphological challenges in Indic scripts, and existing language resources for Bangla and Chakma. These topics collectively highlight the progress and challenges in building effective models for low-resource languages like Chakma.

2.1 Multilingual NLP

The evolution of multilingual models has been driven by the need to extend transformer-based models to low-resource languages, particularly those with limited data or non-Latin scripts (Pakray et al., 2025). Devlin et al. (2019) introduced BERT along with its multilingual variant mBERT, and this marked a turning point. Models like mBERT, pre-trained on Wikipedia data across 104 languages using WordPiece (vocabulary of 110K tokens), and XLM-R, trained on CommonCrawl data from 100 languages with a 250K SentencePiece vocabulary (Conneau et al., 2019), enabled zero-shot cross-lingual transfer. XLM-R, relying solely on MLM pretraining, achieved state-of-the-art performance on multiple benchmarks (Ebrahimi and Kann, 2021).

These multilingual models often show strong zero-shot performance, but disparities remain: languages with less pretraining data or non-Latin scripts typically lag behind high-resource languages (Ebrahimi and Kann, 2021; Marchisio et al., 2024). For example, Wu and Dredze (2020) and Muller et al. (2020) show that mBERT's zero-shot accuracy varies widely by language, with some "hard" languages (often low-resource or using different scripts) remaining poorly served without additional adaptation. These findings spurred research leveraging pretrained transformer models and specialized techniques to handle under-represented languages (Tela et al., 2020; Hangya et al., 2022; Bharadiya, 2023; Pakray et al., 2025). Our work builds on this by fine-tuning a Chakma-specific MLM encoder, addressing data scarcity for this low-resource Indic language.

2.2 Model adaptation techniques and quantization

To address performance disparities in low-resource languages, adaptation strategies emerged to tailor pre-trained models to specific languages or domains. When more data are available, continued monolingual pre-training in target-language data, as demonstrated by Chau et al. (2020), improved zero-shot performance, while domain-adaptive MLM pre-training improves downstream performance even in low-resource settings (Gururangan et al., 2020). Another strategy is to expand the vocabulary of a multilingual model to better cover the target language's lexicon and then additional MLM training improves performance for underrepresented languages (Wang et al., 2020).

2.3 Tokenization and morphology in Indic scripts

Indic languages like Bangla and Chakma are morphologically rich and use complex abugida scripts (Chowdhury, 2025), which raise challenges for subword tokenization. Standard BPE or WordPiece tokenizers can fragment important morphological units, hurting model performance (Pattnayak et al., 2025). Recent work demonstrates that SentencePiece (unigram) tokenization often preserves morphological information better than BPE for Indic languages. For instance, Pattnayak et al. (2025) found that, for zero-shot named entity recognition across several Indic languages, a SentencePiece-based vocabulary outperformed BPE, because it more cleanly segments root words and affixes. Others have noted that vowel forms in abugida scripts (*matras*) attach to consonants and can appear above, below, or beside the base character, which makes character-level segmentation non-trivial (Kashid and Bhattacharyya, 2024; Maung et al., 2025).

2.4 Bangla and Chakma language resources

Although Joshi et al. (2020) categorize Bangla among languages lacking labeled data, Bhattacharjee et al. (2021) developed BanglaBERT, a BERT-base model on the *Bangla2B+* corpus with 2.18 billion tokens from Bangla text, and introduced the Bangla Language Understanding Benchmark (BLUB). BanglaBERT achieves state-of-the-art results on multiple Bangla NLU tasks, outperforming both multilingual baselines (mBERT, XLM-R) and previous monolingual models.

In contrast, NLP work on Chakma is scant. The first known work in Chakma NLP effort is ChakmaNMT (Chakma et al., 2024), which constructed the first parallel corpus (15K sentence-pairs translation, from Chakma to Bangla) and trained a translation model. Using BanglaT5 and transliteration-based back-translation, they achieved a BLEU score of 17.8 for Chakma to Bangla translation. However, this work does not include the Bangla-transliterated Chakma text. The MELD dataset (Mahi et al., 2025) compiled transliterated sentence-level text in Chakma (and Garo, Marma) using the Bangla script. We opted not to use MELD, as its collection of isolated sentences lacks the semantic coherence required for our study. Instead, we focus on Bangla-transliterated Chakma texts extracted via OCR from printed literature.

3 Dataset Creation

Emphasizing the **authenticity** of linguistic resources, particularly in the field where the digitized materials are scarce and under-resourced, we con-

Corrected Dataset	Pytesseract	Gemini
আজবত্র সাপ এম্বা এয়ে ধুরি রঙ্গলালর দুনিয়াদারি উধিচ ন পেইয়্যা এয়ান্ন্যা দুমুর কিয়ই ন দেগন। ধুপছুরি আদাম উধিজে জক্কে তে দুমুর দিল, জিদু তা ঘরান, সেক্কে গদা পিথিমিয়ান তা চেরকিত্ত্যা কালা আন্ধার ভিদিরে লুগি জিয়েগোই। সে ছাবা তা মন উঙুরে পরি তার গম-বজঙ চিদে গোরিভের ছেদামানো ভচ নেজে ফেলেয়্য।	আজবত্র সাপ এম্বা এয়ে ধুরি রঙ্গলালর দুনিয়াদারি উধিচ ন পেইয়্যা এয়ননযাদুমুর কিয়ই ন দেগন। ধূপছুরি আদাম উধিজে জকে তে দুমুর দিল, জিদু তা ঘরান, সেক্কে গদা পিথিমিয়ান তা চেরকিত্ত্যা কালা আম্ধার ভিদিরে লুগি জিয়েগোই। সে ছাবা তা মন উ্জুরে পরি তার গম-বজঙ চিদে গোরিভের ছেদামানো ভচ নেজে ফেলেয়্য।	আজবত্র সাপ missing sentence দেগন। ধূপছুরি আদাম উধিজে জকে তে দুমুর দিল, জিদু তা ঘরান, সেক্কে গদা পিথিমিয়ান তা চেরকিত্ত্যা কালা আম্ধার ভিদিরে লুগি জিয়েগোই। সে ছাবা তা মন উ্জুরে পরি তার গম-বজঙ চিদে গোরিভের ছেদামানো ভচ নেজে ফেলেয়্য।

Figure 2: Sample data illustrating quality comparison across different methods, highlighting **missing sentences** in Gemini and **spelling errors** in other models caused by the misinterpretation of conjunct characters, phonetic signs, vowel diacritics, consonant modifiers, nasalization, and related orthographic features.

Model	**Vocab Size**	**Tokenizer & Special Tokens**	**Tokenization Method**
BERT-Base Multilingual (cased)	∼120k (WordPiece)	[CLS] ... [SEP], [MASK]	WordPiece
DistilBERT Multilingual (cased)	∼120k (WordPiece)	[CLS] ... [SEP], [MASK]	WordPiece
XLM-RoBERTa (XLM-R)	∼250k (SentencePiece)	$\langle s \rangle$ ... $\langle /s \rangle$, $\langle mask \rangle$	SentencePiece
DeBERTaV3-Base	∼128k (WordPiece-style)	[CLS] ... [SEP], [MASK]	WordPiece-style
Bangla BERT Base	∼32k (WordPiece)	[CLS] ... [SEP], [MASK]	WordPiece
IndicBERT	∼200k (SentencePiece)	$\langle s \rangle$ ... $\langle /s \rangle$, $\langle mask \rangle$	SentencePiece

Table 1: Comparison of encoder-based models used in our evaluation. Differences arise in vocabulary size, tokenizer conventions, and tokenization methods. The models include BERT-Base Multilingual (Devlin et al., 2019), DistilBERT Multilingual (Sanh et al., 2019), XLM-RoBERTa (Liu et al., 2019), DeBERTaV3-Base (He et al., 2021), Bangla BERT (Sarker, 2020), and IndicBERT (Kakwani et al., 2020).

struct a novel dataset combining four books (novels and poems) written in the Chakma language, utilizing Bangla script (see Figure 4 and Table 2). To collect these materials, we directly engaged with scholars whose first language is Chakma. The books were gathered from libraries on the basis of their recommendations. However, we acknowledge that most scholars prioritize the preservation and use of their own Chakma script. Chakma et al. (2024) also assigns importance to the Chakma script. Most pretrained models lack support for the complex structure of Chakma scripts. The Bangla-transliterated Chakma script enables the models to process the language effectively using their existing tokenizers.

3.1 Corpus Compilation: Sources and Scale

The data were extracted from the image of the pages of the books using **three primary methods**: Pytesseract (Hoffstaetter et al.), Gemini (Comanici et al., 2025), and manual processing. We used the PyTesseract OCR model and the Gemini 2.5 Pro model API separately to independently assess the quality of text extraction from different systems.

PyTesseract encountered problems with the recognition of Bangla's conjunctive characters and committed frequent spelling errors, as presented in Figure 2. On the other hand, Gemini 2.5 Pro with the free API, posed usage restrictions creating a barrier to scalability as we processed 400 images. Moreover, the Gemini API deviated from correctness in alphabet recognition, often produced incomplete sentences, and sometimes omitted entire sentences, affecting the overall quality of the extracted text. Some examples are presented in Figure 2.

Due to these limitations, we manually fixed one book entirely and another book partially, which we discuss in Section 3.2. The dataset is split into training, testing and validation subsets, as shown in Table 2.

3.2 Manual Curation for Linguistic Fidelity

Both the OCR models and LLMs struggled with accurate processing of conjunct characters, phonetic signs, including vowel diacritics, consonant modifiers, nasalization, silent consonant and incom-

Dataset	Dataset Split (sentences)
Tesseract OCR (Tes OCR)	train: 4,348 eval: 832 test: 1,173
Gemini OCR	train: 3,815 eval: 994 test: 1,173
Manually Fixed Data	train: 2,908 eval: 545 test: 1,118

Table 2: Breakdown of training, evaluation, and test sentence counts for datasets obtained from Tesseract OCR, Gemini OCR outputs, and manually corrected data.

pleteness of sentences (Figure 2). These complex character clusters are fundamental to the Bangla orthography but often are misinterpreted or omitted by OCR systems due to their non-linear composition and script variability (Ali et al., 2023; Guo et al., 2023). After identifying the limitations and to ensure the linguistic fidelity of our dataset, 4,570 sentences of OCR and LLM extracted text underwent a multi-stage manual correction and validation process. Two co-authors of this paper rectified these specific errors to guarantee the high integrity and usability of the final manual dataset. The overall workflow of the paper is presented in Figure 1.

4 MLM-tuning for low-resource languages

We fine-tuned six encoder-based models (including monolingual, multilingual and regional variants) on limited Chakma text written in Bangla script and compared their performance. Table 1 summarizes the models used in our experiments, including their vocabulary sizes and tokenization algorithms. All of these models have been pre-trained on Bangla before.

These LLMS are trained to predict the probability of a masked token/word given the context of surrounding words. This gives the models a foundational understanding of trained languages that can be generalized to other tasks (Wolf et al., 2020). Although each model comes in a similar-sized 12-layer configuration (270M–300M parameters for base models), they vary in vocabulary sizes, tokenizer types, special tokens, and critically, the dataset they were first pre-trained on.

For MLM fine-tuning, we masked 15% of tokens in each input sequence using the standard masking strategy: 80% replaced with the appropriate mask token ([MASK] or <mask>), 10% substituted with random vocabulary tokens, and 10% left unchanged. We maintained strict separation between training, validation, and test datasets across all experiments to prevent data leakage.

After multiple trials and errors, in our final configuration, we use the Adam optimizer, with a learning rate of 2×10^{-5} for all the models. The maximum number of epochs is 20. The dropout rate is 0.01. We keep the batch size at 8. For testing the fine-tuned models, we ensure consistency and reproducibility across the models.

5 Results

We evaluated the performance across three different data processing pipelines (Pytesseract, Gemini and manual processing) and also compared both universal and regional model types. Following the works of Salazar et al. (2019), Rogers et al. (2021) and Ethayarajh (2019), we evaluate our MLM fine-tuned models using perplexity, masked token accuracy, precision, recall, F1(macro), pseudo-log-likelihood (PLL) and predictive entropy.

5.1 Language Modeling Capability

RQ1: How effective are the pre-trained language models at masked language modeling for the (monolingual) Chakma language written in the Bangla script?

Table 3 shows that fine-tuned encoder-based language models consistently outperform their pre-trained counterparts for Bangla-transliterated Chakma. The fine-tuned models achieve accuracies up to 73.54% (XLM-RoBERTa) and perplexity as low as 2.899 (DeBERTaV3-Base) on manually corrected data, underscoring the value of adaptation for low-resource languages. Notably, monolingual models like DeBERTaV3-Base, which start with no prior knowledge of Chakma or Bangla script (0% baseline accuracy), achieve competitive results post-fine-tuning, demonstrating the robustness of adaptation even without cross-lingual pre-training. MLM-tuning also yields a marked reduction in prediction entropy, indicating increased confidence in masked-token predictions in Table 6.

Our best perplexity score of 2.899 is substantially lower (indicating better performance) than

Model	Accuracy (%) ↑			Perplexity ↓		
	Without MLM	With MLM	**Performance**	Without MLM	With MLM	**Performance**
DeBERTaV3-Base	0.00	72.08	+72.08	39329757.5	**2.90**	-39329754.6
XLM-RoBERTa	46.24	**73.54**	+27.30	24.39	3.27	-21.12
BERT-Base mBERT	**48.43**	70.00	+21.57	**13.12**	4.017	-9.103
DistilBERT Multilingual	38.78	65.08	+26.30	24.284	4.3046	-19.978
Bangla BERT Base	29.87	54.52	+24.65	250.09	11.79	-238.3
IndicBERT	17.54	45.36	+27.82	1823.61	16.79	-1806.82

Table 3: Performance comparison of models before and after MLM fine-tuning using manually annotated Chakma corpora. Accuracy (%) and perplexity are reported. Lower perplexity indicates better language modeling performance.

the perplexity scores reported for BERT on English datasets (Salazar et al., 2019). We treat this as an empirical observation rather than definitive evidence of superior absolute performance. This low perplexity may be an artifact of our dataset characteristics, including its relatively small size and the specific nature of the data (potentially featuring simpler or more repetitive linguistic structures compared to diverse English corpora). Hypotheses for this include reduced lexical diversity or script-specific tokenization efficiencies in Chakma, but the exact reasons remain unclear and could be explored in more detail in future work, perhaps by evaluating on larger, more varied Chakma datasets.

Outperforming of universal models over regional encoders: From Figure 3 and Table 4, we observe a consistent advantage for multilingual encoder models (XLM-RoBERTa, BERT-Base mBERT, DistilBERT Multilingual) and the monolingual DeBERTaV3-Base over regional encoder models (BanglaBERT, IndicBERT). Because tokenizers and vocabulary sizes differ across models, masked-language accuracy and perplexity are computed on model-specific tokenizations rather than an identical token sequence. This can potentially advantage models that produce fewer tokens per input, since they evaluate fewer positions and may face fewer rare-subword predictions. However, we argue that the comparison remains informative: the vocabularies are not extremely different, and the underlying dataset is identical for all models, and that accuracy is not simply determined by token count (see Table 1 and Table 3).

We analyze two primary factors that influence

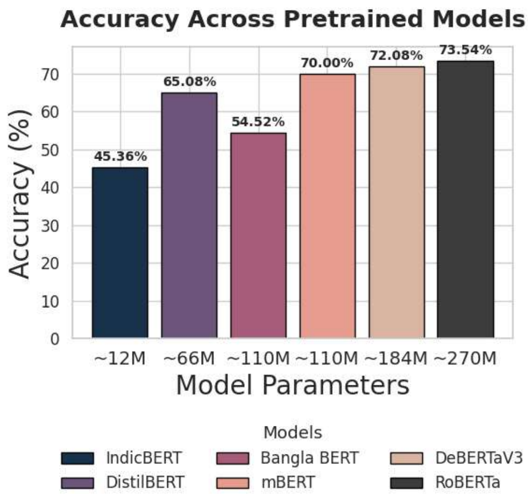

Figure 3: Comparison of universal multilingual and regional encoder models. Each grouped bar chart is showing the accuracy of pre-trained language models fine-tuned on manually fixed data, categorized by their parameter sizes.

model effectiveness: model parameter size and tokenization efficiency.

1. Parameter size → tokenization robustness. Larger multilingual models are trained on broader, more diverse corpora and typically learn richer subword vocabularies. This reduces out-of-vocabulary occurrences, over-fragmentation, and tokenization drift, which can otherwise harm downstream performance. These effects can cause some tokenizers to produce 2–3 times more tokens for the same input (see Figure 3) (Rust et al., 2020).

2. Tokenizer efficiency → evaluation metrics. A smaller number of tokens allows each token to carry more semantic context and reduces prediction

Model	Manual		Tesseract				Gemini			
			Self-Finetuned		Manual-Finetuned		Self-Finetuned		Manual-Finetuned	
	Acc.(↑)	PPL(↓)	Acc.(↑)	PPL(↓)	Acc.(↑)	PPL(↓)	Acc.(↑)	PPL(↓)	Acc.(↑)	PPL(↓)
DeBERTaV3-Base	72.08	2.90	46.94	7.82	46.52	9.75	46.86	9.12	45.77	10.01
XLM-RoBERTa	73.54	3.27	30.28	54.39	29.14	77.74	28.70	80.16	29.67	78.04
BERT-Base mBERT	70.00	4.02	32.91	27.50	31.56	40.59	31.75	44.16	31.23	42.35
DistilBERT Multilingual	65.08	4.30	31.64	29.05	29.63	43.91	30.49	43.20	30.04	41.96
Bangla BERT Base	54.52	11.79	22.17	299.54	20.19	384.28	20.71	483.57	20.86	467.25
IndicBERT	45.36	16.79	23.04	83.67	24.05	80.18	23.64	97.62	23.12	103.10

Table 4: Impact of data quality on model performance. Accuracy (%) and Perplexity (PPL) are reported for each model fine-tuned on manually annotated, Tesseract-processed, and Gemini-processed data. In our table, *Self-Finetuned* refers to training and evaluating each model on the same dataset, while *Manual-Finetuned* involves training on manually corrected data but evaluating on other test datasets like Tesseract or Gemini test sets.

noise for masked positions. Over-fragmentation, by contrast, spreads probability mass across many rare subwords, penalizing sequence-level scoring and hurting pseudo-log-likelihood (PLL) (Kudo and Richardson, 2018).

5.2 Impact of Data Quality

RQ2: In the context of the morphologically rich Bangla-transliterated Chakma, how does the OCR noise of data affect MLM performance?

Building on the findings from RQ1, where fine-tuning encoder-based models on manually corrected Chakma data demonstrated strong improvements in masked language modeling capabilities, we now explore the extent to which OCR-induced noise (stemming from script-specific challenges like transliteration variations and complex conjunct consonants) disrupts the learning of morphological structures in Bangla transliterated Chakma. In each case, the models were fine-tuned and evaluated on their respective dataset, which we refer to as *Self-Finetuned* in our Table 4. Additionally, we evaluated the model fine-tuned on the manually corrected dataset against the Tesseract and Gemini 2.5 Pro test sets, which we denote as *Manually-Finetuned* in the Table 4. Due to transliteration-induced variation with more complex conjunct consonants, the transliterated data (Bangla-transliterated Chakma) appears morphologically heavier than Bangla.

From the Table 4, we can see that models trained with Tesseract and Gemini 2.5 Pro processed data struggled to grasp the Chakma language, showing limited improvements even after fine-tuning, particularly evident in cases where models like DeBERTaV3-Base(He et al., 2021) had no initial understanding of Chakma (Table 3). The fine-tuning with the manually fixed dataset led to substantial gains in accuracy, highlighting that the model learns the affixes, inflections and complex forms of the language in a better way. Meanwhile, these models drop their performance when testing on the noisy test dataset. For instance, the XLM-RoBERTa model achieved its strongest performance with manual data, far surpassing its baseline and revealing that noisy OCR outputs can actually degrade model capabilities compared to their pre-fine-tuned state.

We find a similar pattern when examining perplexity across datasets for individual models. From Table 4, the manual dataset consistently yielded low perplexity, indicating strong language modeling and coherence. However, Tesseract and Gemini data introduced higher perplexity, often worsening it beyond the base model's levels due to inherent noise and errors. This trend holds across all six models in our experiments, emphasizing how high-quality data refines predictions while OCR-generated inaccuracies amplify confusion. Furthermore, when testing manually fine-tuned models on Tesseract or Gemini data, their perplexity suffered slightly compared to self-fine-tuned counterparts, reinforcing the pervasive impact of noise in OCR pipelines on overall model robustness.

Overall, these results show the critical role of preserving morphologically accurate data quality in enhancing model performance for low-resource indigenous languages like Chakma.

6 Conclusion

In this work, we introduced a Bangla-transliterated Chakma dataset, derived from Chakma literature using Tesseract, Gemini 2.5 Pro OCR and manual transcription. We empirically demonstrate that pre-trained multilingual language models can be effectively adapted for the Chakma language through fine-tuning on this data, establishing a strong baseline for Masked Language Modeling for Chakma. Our comprehensive experiments further underscore that model performance is highly sensitive to data quality, and that iterative cleaning directly enhances model performance. To support future research, we publicly release our manually refined dataset. A compelling direction for future work is to investigate the optimal transliteration target for low-resource languages. We hypothesize that for Chakma, which shares significant typological and lexical similarity with Bangla, transliteration into the Bangla script may yield superior performance compared to the English script, despite the generally stronger pre-training of LLMs on English. Systematically evaluating this trade-off between linguistic proximity and model capability remains an open question.

7 Limitations and Future Work

This study focuses on understanding the potential of LLM adaptability to low-resource languages. In our work, we have considered Chakma language as a case study. However, our manually validated Bangla-transliterated Chakma language dataset contains only 4570 sentences. The sentences are collected from story books, which is not sufficient to reflect diverse real-world scenarios, especially in a modern context. So, we aim to expand our Chakma corpus incorporating more diverse text sources, including spoken language transcripts, community-generated contents and parallel translations. Transliteration of Chakma dataset to Latin script is another direction of research following the works of Muller et al. (2020). If such a dataset exists, we can test the hypothesis that transliterating Chakma to a related language (Bangla) as opposed to the strongest language (English) may yield better performance. Inspired by Devlin et al. (2019), we can test our fine-tuned model for Next Sentence Prediction (NSP) accuracy to get a better understanding of how well our model is understanding the Chakma language. Improving OCR accuracy to extract the text with a better performance for conjunct characters, phonetic signs including vowel diacritics, consonant modifiers, nasalization, and others is also a potential direction for improvement.

8 Acknowledgement

The authors would like to thank Bijoy Chakma (Sociology, University of Rajshahi) and Shriswa Chakma (Department of Management Studies, Bangladesh University of Professionals) for their contributions to data collection and manual processing.

References

Hasmot Ali, AKM Shahariar Azad Rabby, Md Majedul Islam, A.k.m Mahamud, Nazmul Hasan, and Fuad Rahman. 2023. Gold standard Bangla OCR dataset: An in-depth look at data preprocessing and annotation processes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 460–470, Singapore. Association for Computational Linguistics.

J Bharadiya. 2023. Transfer learning in natural language processing (nlp). *European Journal of Technology*, 7(2):26–35.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Carmen Brandt. 2018. Writing off domination: the chakma and meitei script movements. *South Asian History and Culture*.

Aunabil Chakma, Aditya Chakma, Soham Khisa, Chumui Tripura, Masum Hasan, and Rifat Shahriyar. 2024. Chakmanmt: A low-resource machine translation on chakma language. *arXiv preprint arXiv:2410.10219*.

Chakma Autonomous District Council. 2025. The chakma people. https://www.cadc.gov.in/the-chakma-people/. Accessed: 2025-10-05.

Ethan C Chau, Lucy H Lin, and Noah A Smith. 2020. Parsing with multilingual bert, a small corpus, and a small treebank. *arXiv preprint arXiv:2009.14124*.

Nilanjana Roy Chowdhury. 2025. Phonological adaptations of bangla words in chakma. *Strength for Today and Bright Hope for Tomorrow Volume 25: 3 March 2025 ISSN 1930-2940*, page 58.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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. *arXiv preprint arXiv:1911.02116*.

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.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. *arXiv preprint arXiv:2106.02124*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Aloka Fernando and Surangika Ranathunga. 2025. Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages. *arXiv preprint arXiv:2501.05700*.

Linsheng Guo, Md Habibur Sifat, and Tashin Ahmed. 2023. Pipeline enabling zero-shot classification for Bangla handwritten grapheme. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 26–33, Singapore. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Samuel Hoffstaetter, Juarez Bochi, Matthias Lee, Lars Kistner, Ryan Mitchell, and Emilio Cecchini. Python-tesseract: Ocr with tesseract in python. `https://github.com/h/pytesseract`,`https://tesseract-ocr.github.io/`.

Jesin James, Vithya Yogarajan, Isabella Shields, Catherine Watson, Peter J Keegan, Peter-Lucas Jones, and Keoni Mahelona. 2022. Language models for code-switch detection of te reo māori and english in a low-resource setting. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 650–660. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Harshvivek Kashid and Pushpak Bhattacharyya. 2024. Roundtripocr: A data generation technique for enhancing post-ocr error correction in low-resource devanagari languages. *arXiv preprint arXiv:2412.15248*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mehraj Hossain Mahi, Anzir Rahman Khan, Mobashsher Hasan Anik, Sheak Rashed Haider Noori, Arif Mahmud, and Mayen Uddin Mojumdar. 2025. Meld: A multilingual ethnic dataset of chakma, garo, and marma in bengali script with english and standard bengali translation. *Data in Brief*, page 111745.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual llms? *arXiv preprint arXiv:2407.03211*.

Aye T Maung, Sumaiya Salekin, and Mohammad A Haque. 2025. A hybrid approach to bangla handwritten ocr: combining yolo and an advanced cnn. *Discover Artificial Intelligence*, 5(1):119.

Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. *arXiv preprint arXiv:2010.12858*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resource languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126. Association for Computational Linguistics.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Priyaranjan Pattnayak, Hitesh Patel, and Amit Agarwal. 2025. Tokenization matters: Improving zero-shot ner for indic languages. In *2025 IEEE International Conference on Electro Information Technology (eIT)*, pages 456–462. IEEE.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. 2020. Transferring monolingual model to low-resource language: The case of tigrinya. *arXiv preprint arXiv:2006.07698*.

Zihan Wang, Stephen Mayhew, Dan Roth, and 1 others. 2020. Extending multilingual bert to low-resource languages. *arXiv preprint arXiv:2004.13640*.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: A reproduction of the llama training dataset. *arXiv preprint arXiv:2411.12372*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, and 1 others. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

A Appendix

Table 5: Model Fine-tuned on Manual Data: Cross-Test Performance

Model	Data	Loss	Perplexity	Accuracy(%) (masked token acc.)	Precision	Recall	F1_macro	Prediction Entropy	Pseudo-log likelihood	Evaluated Tokens (sentences)
manual_roberta	manual	3.1942	24.3903	46.24	0.3432	0.2799	0.2876	2.8258	-3.1910	8092 (431)
	teserrect	4.3534	77.7443	29.14	0.2502	0.1469	0.1603	2.7436	-4.3550	10713 (569)
	gemini-2.5-pro	4.3573	78.0479	29.67	0.253	0.1576	0.1677	2.7591	-4.3559	10645 (569)
manual_bert	manual	1.3905	4.017	70	0.5107	0.4353	0.4512	1.0178	-1.3916	9011 (495)
	teserrect	3.7036	40.5924	31.56	0.2886	0.1837	0.1974	2.123	-3.7023	12198 (657)
	gemini-2.5-pro	3.7461	42.3548	31.23	0.2738	0.1716	0.1880	2.1382	-3.7426	12445 (657)
manual_distilbert	manual	1.4740	4.3668	67.10	0.4803	0.4074	0.4074	1.2122	-1.4742	9180 (495)
	teserrect	3.7822	43.9145	29.63	0.233	0.1415	0.1521	2.3025	-3.7787	12516 (657)
	gemini-2.5-pro	3.7368	41.9629	30.04	0.2357	0.1568	0.1636	2.3025	-3.7411	12424 (657)
manual_debarta	manual	1.0644	2.8991	72.08	0.452	0.3549	0.3724	0.8393	-1.0654	16202 (864)
	teserrect	2.2775	9.7526	46.52	0.1911	0.1402	0.1446	1.3831	-2.2782	20466 (1085)
	gemini-2.5-pro	2.3037	10.0116	45.77	0.1999	0.1460	0.1515	1.3957	-2.3066	20402 (1085)
sagorbangla	manual	2.4675	11.7925	54.52	0.367	0.2663	0.2861	2.6335	-2.4707	18587 (986)
	teserrect	6.1659	476.2059	20.45	0.1542	0.0979	0.1042	3.9245	-6.1638	8216 (434)
	gemini-2.5-pro	6.2530	519.5888	20.51	0.1343	0.0826	0.0886	3.9889	-6.2558	8148 (434)
IndicBERT	manual	2.4492	11.5791	52.75	0.4337	0.2994	0.3297	2.4188	-2.4546	6627 (351)
	teserrect	4.6562	105.2307	23.03	0.2572	0.1140	0.1385	3.7807	-4.6625	8392 (429)
	gemini-2.5-pro	4.5815	97.6632	23.19	0.2964	0.1357	0.1649	3.7785	-4.5871	8251 (429)

Table 6: Model Performance Comparison

Model	Data	Type	Loss	Perplexity	Accuracy (masked token accuracy)	Precision	Recall	F1_macro	Prediction Entropy	Pseudo-log-likelihood	Tokens
BERT-Base Multilingual (cased)	manual	base	2.5744	13.1234	48.43%	0.4173	0.3144	0.3321	2.3547	-2.5763	9126 (495)
		finetuned	1.3905	4.017	70%	0.5107	0.4353	0.4512	1.0178	-1.3916	9011 (495)
	teserrect	base	3.8187	45.5459	25.05%	0.2491	0.1507	0.1640	3.3034	-3.8194	12371 (657)
		finetuned	3.3143	27.5039	32.91%	0.2882	0.1706	0.1889	2.6401	-3.3211	12493 (657)
	gemini-2.5-pro	base	3.7713	43.4369	26.19%	0.2492	0.1543	0.1695	3.2742	-3.7731	12338 (657)
		finetuned	3.7878	44.1608	31.75%	0.2391	0.1821	0.1855	2.1291	-3.7898	12353 (657)
DistilBERT Multilingual (cased)	manual	base	3.1898	24.284	38.78%	0.3030	0.2044	0.2135	3.2877	-3.1989	9106 (495)
		finetuned	1.4597	4.3046	65.08%	0.5070	0.4045	0.4323	1.4180	-1.4636	8796 (439)
	teserrect	base	4.2519	70.2371	20.10%	0.2064	0.1030	0.1090	3.8775	-4.2527	12126 (657)
		finetuned	3.3691	29.052	31.64%	0.2845	0.1699	0.1872	2.6530	-3.3700	1197 (657)
	gemini-2.5-pro	base	4.2025	66.8516	20.30%	0.2175	0.1087	0.1150	3.8539	-4.2038	12459 (657)
		finetuned	3.7658	43.197	30.49%	0.2533	0.1701	0.1811	2.3147	-3.7675	12353 (657)
XLM-RoBERTa (XLM-R)	manual	base	3.1942	24.3903	46.24%	0.3432	0.2799	0.2876	2.8258	-3.1910	8092 (431)
		finetuned	1.1858	3.2732	73.54%	0.5677	0.5114	0.5219	1.0693	-1.1918	8043 (431)
	teserrect	base	4.5293	92.6939	25.79%	0.2294	0.1498	0.1619	4.0378	-4.5321	10653 (569)
		finetuned	3.14	54.39	30.28%	0.2634	0.1554	0.1719	3.2284	-3.9981	10327 (569)
	gemini-2.5-pro	base	4.5354	93.2627	25.72%	0.2213	0.1417	0.1552	4.0543	-4.5400	10762 (569)
		finetuned	4.3840	80.1618	28.70%	0.2421	0.1575	0.1631	2.8692	-4.3910	10620 (569)
DeBERTaV3 Base	manual	base	17.4875	39329757.5	0%	0.0000	0.0000	0.0000	7.8197	-17.4848	16457 (864)
		finetuned	1.0644	2.8991	72.08%	0.4520	0.3549	0.3724	0.8393	-1.0654	16202 (864)
	teserrect	base	14.6398	2280263.447	0%	0.0000	0.0000	0.0000	7.8999	-14.6415	20447 (1085)
		finetuned	2.0572	7.8241	46.94%	0.3072	0.1697	0.1859	1.562	-2.0610	20477 (1085)
	gemini-2.5-pro	base	16.1899	10744904.22	0%	0.0000	0.0000	0.0000	8.1224	-16.1899	20398 (1085)
		finetuned	2.2105	9.1200	46.86%	0.2412	0.1576	0.1688	1.4322	-2.2130	20509 (1085)
Bangla BERT Base	manual	base	5.5218	250.0904	29.87%	0.2338	0.1478	0.1637	5.7929	-5.5302	6451 (344)
		finetuned	2.4675	11.7925	54.52%	0.3670	0.2663	0.2861	2.6335	-2.4707	18587 (986)
	teserrect	base	6.8747	967.5296	15.67%	0.1516	0.0860	0.0981	6.6578	-6.8720	8129 (434)
		finetuned	5.7022	299.5393	22.17%	0.1598	0.1039	0.1116	4.2600	-5.7041	8167 (434)
	gemini-2.5-pro	base	6.8133	909.9072	16.95%	0.1514	0.0847	0.0967	6.5669	-6.8131	8256 (434)
		finetuned	6.1469	467.2527	20.86%	0.1442	0.0888	0.0953	3.9319	-6.1402	8179 (434)
IndicBERT	manual	base	7.5091	1824.6141	17.54%	0.3636	0.1131	0.1489	5.3301	-7.5190	18403 (971)
		finetuned	2.8209	16.7924	45.36%	0.4261	0.2239	0.2693	2.8646	-2.8273	18418 (971)
	teserrect	base	8.0048	2995.1897	12.62%	0.2703	0.0822	0.1060	5.4509	-8.0154	8065 (429)
		finetuned	4.4269	83.6677	23.04%	0.2744	0.1141	0.1434	3.9951	-4.4268	8046 (429)
	gemini-2.5-pro	base	8.0774	3220.7776	12.42%	0.2762	0.0784	0.1026	5.4241	-8.0825	8130 (429)
		finetuned	4.5811	97.6174	23.64%	0.2733	0.1186	0.1470	3.7354	-4.5813	8046 (429)

Figure 4: Details of Chakma Storybooks Used in the Dataset

Book Name	**Author(s)**	**Publication**	**Total Pages**
Book-1: চাকমা ছোটগল্প (Chakma Short Stories)	মৃত্তিকা চাকমা, ঝিমিত ঝিমিত চাকমা, মুক্তা চাকমা (Mrittika Chakma, Jhimita Jhimita Chakma, Mukta Chakma)	জুম ঈসথেটিকস কাউন্সিল (জাক), রাঙ্গামাটি জেলা পরিষদ (Jum Aesthetics Council (JAC), Rangamati Hill District Council)	51
Book-2: আজব সাপ (The Strange Snake)	বিপম চাঙমা (Bipom Changma)	জেড কম্পিউটার এন্ড প্রিন্টার্স, আন্দরকিল্লা, চট্টগ্রাম (J Computer and Printers, Andarkilla, Chattogram)	57
Book-3: হিলট্র্যাক্টসের দুঃখ সুখ CHANGMA KABYE HIL TRAKSAR DUG SUG (The Joys and Sorrows of the Hill Tracts)	জ্ঞানের আলো চাঙমা (Gyaner Alo Changma)	ইসলামিয়া অফসেট প্রেস, খাগড়াছড়ি (Islamiya Offset Press, Khagrachari)	103
Book-4: সাম্মো	সোনা মনি চাঙমা (Sona Moni Changma)	বনযোগীছাড়া, জুরাছড়ি (Bonyogichhara, Jurachhari)	63