ChakmaBridge: A Five-Way Parallel Corpus for Navigating the Script Divide in an Endangered Language

Md. Abdur Rahman[1] **Md. Tofael Ahmed Bhuiyan**[1]
Abdul Kadar Muhammad Masum[1]
[1]Department of Computer Science and Engineering, Southeast University
Dhaka, Bangladesh
abdurrahman.etc@gmail.com, tofael1104@gmail.com, akmmasum@seu.edu.bd

Abstract

The advancement of NLP technologies for low-resource and endangered languages is critically hindered by the scarcity of high-quality, parallel corpora. This is particularly true for languages like Chakma, which also faces the challenge of prevalent non-standard, romanized script usage in digital communication. To address this, we introduce ChakmaBridge, the first five-way parallel corpus for Chakma, containing 807 sentences aligned across English, Standard Bangla, Bengali-script Chakma, Romanized Bangla, and Romanized Chakma. Our dataset is created by augmenting the MELD corpus with LLM-generated romanizations that are rigorously validated by native speakers. We establish robust machine translation baselines across six diverse language and script pairs. Our experiments reveal that a multilingual training approach, combining English and Bangla as source languages, yields a dramatic performance increase, achieving a BLEU score of 0.5228 for Chakma translation, a 124% relative improvement over the best bilingual model. We release ChakmaBridge to facilitate research in low-resource MT and aid in the digital preservation of this endangered language.

1 Introduction

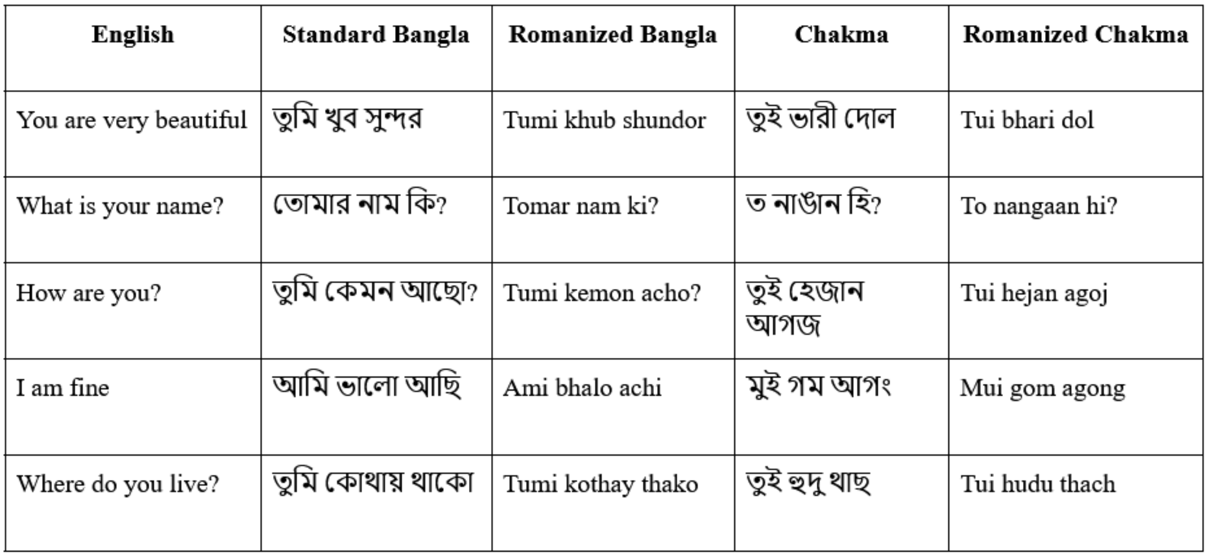
While Large Language Models (LLMs) have driven remarkable progress in Natural Language Processing (NLP), these advancements have not been shared equally across the globe's linguistic landscape (Joshi et al., 2020). Low-resource languages, which are spoken by millions but lack extensive digital corpora, are often excluded from the benefits of modern NLP. This paper addresses this gap by focusing on two such languages from the Indian subcontinent: Bangla and Chakma.

Bangla (Bengali), an Indo-Aryan language with over 270 million native speakers worldwide, is one of the most spoken languages, yet it remains significantly under-resourced in the digital domain compared to languages like English. The Chakma language, also a member of the Indo-Aryan family, is spoken by approximately 800,000 people across communities in Bangladesh, India, and Myanmar (Chakma et al., 2024; Bangladesh Bureau of Statistics, 2023). Despite its significant speaker base, Chakma is classified as "Definitely Endangered" (Saikia and Haokip, 2023), facing immense pressure from dominant regional languages and a critical scarcity of digital resources needed for its preservation and revitalization (Chakma and Maitrot, 2016).

The challenge of digital inclusion is compounded by the widespread use of non-standard scripts in online communication. A prevalent form of text for both Bangla and Chakma on social media and messaging platforms is a romanized version, where the Latin alphabet is used to phonetically represent native words (Moosa et al., 2023). This practice, driven by the ubiquity of QWERTY keyboards, creates a vast but unstructured data source. While this informal text is a valuable linguistic resource, its potential remains largely untapped due to the lack of standardized, parallel datasets that connect it to its native-script counterparts (Roark et al., 2020).

While some foundational computational work exists for Chakma, such as models for character recognition (Podder et al., 2023) and speech identification (Pratap et al., 2024), research in machine translation (MT) remains nascent. Previous studies have highlighted the potential of pre-trained models for Chakma translation but have also underscored the limitations imposed by fragmented and incomplete data (Chakma et al., 2024). we introduce ChakmaBridge, a new, comprehensive parallel dataset for the Chakma language. By releasing this meticulously curated and validated dataset, we aim to empower researchers to develop MT systems that can handle both official scripts and the informal romanized text common in daily dig-

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 259–265
December 23, 2025 ©2025 Association for Computational Linguistics



English	Standard Bangla	Romanized Bangla	Chakma	Romanized Chakma
You are very beautiful	তুমি খুব সুন্দর	Tumi khub shundor	তুই ভারী দোল	Tui bhari dol
What is your name?	তোমার নাম কি?	Tomar nam ki?	ত নাঙান হি?	To nangaan hi?
How are you?	তুমি কেমন আছো?	Tumi kemon acho?	তুই হেজান আগজ	Tui hejan agoj
I am fine	আমি ভালো আছি	Ami bhalo achi	মুই গম আগং	Mui gom agong
Where do you live?	তুমি কোথায় থাকো	Tumi kothay thako	তুই হুদু থাছ	Tui hudu thach

Figure 1: Sample entries from the final five-way parallel ChakmaBridge dataset, illustrating the alignment across English, Standard Bangla, Romanized Bangla, Bengali-script Chakma, and Romanized Chakma.

ital communication. This work seeks to promote linguistic diversity, support the preservation of an endangered cultural heritage, and foster greater digital inclusion for the Chakma-speaking community. Our dataset and the code for our baseline experiments are publicly available on GitHub[1]. Our contributions are two-fold:

1. We introduce ChakmaBridge, the first five-way parallel corpus for Chakma, augmenting the 807 sentences from the MELD dataset (Mahi et al., 2025) with novel, LLM-generated and human-validated columns for `Romanized Bangla` and `Romanized Chakma`.

2. We establish robust translation baselines across six different language and script modalities, demonstrating the dataset's utility and providing a crucial starting point for future research in low-resource MT.

2 Dataset Creation

The creation of our dataset was a multi-stage process designed to build upon an existing resource, augment it with high-quality, semi-automatically generated data, and validate its integrity through a rigorous human-in-the-loop protocol. This section details our data collection and augmentation pipeline, the validation procedure, a statistical analysis of the final corpus, and the experimental setup for establishing baseline translation models.

2.1 Data Collection and Augmentation

Our work commences with the MELD (Multilingual Ethnic Language Dataset), a valuable resource for low-resource languages of Bangladesh (Mahi et al., 2025). The MELD corpus was compiled through structured interviews with native speakers. Crucially, for languages like Chakma that have their own script (Ajhā Pāṭh), the data was phonetically represented using the Bengali script, a common practice for digital communication among the language community. From this resource, we isolated the complete parallel corpus of 807 sentences where Chakma was available, forming a foundational dataset with three aligned columns: `English`, `Standard Bangla`, and `Chakma` (represented in Bengali script).

The primary contribution of our work is the augmentation of this corpus with romanized parallels, addressing the widespread use of Latin script for informal digital communication. To generate the new columns, `Romanized Bangla` and `Romanized Chakma`, we employed a state-of-the-art Large Language Model (LLM), Gemini 2.5 Pro (Comanici et al., 2025). The model was provided with the `Standard Bangla` text as input to produce the `Romanized Bangla` output, and similarly, the Bengali-script `Chakma` text was used to generate the `Romanized Chakma` column. This semi-automated process resulted in ChakmaBridge, a new five-column parallel dataset designed to facilitate research in transliteration, normalization, and multilingual MT between different scripts.

[1] `https://github.com/borhanitrash/ChakmaBridge`

2.2 Validation Protocol

Given the generative nature of the romanization process, we implemented a rigorous, human-centric validation protocol to ensure the quality and authenticity of the LLM-generated text. The complete set of 807 generated sentence pairs underwent an initial quality assurance review. This process was carried out by a team of three: two senior-year undergraduate students in Computer Science and Engineering, and a professor with a PhD in Electronic Human Resource Management. Their role was to ensure the LLM's output accurately corresponded to the source text, verifying the transliteration alignment rather than the semantic content. Any transliteration errors or outputs that were clearly inconsistent with the input were identified and corrected during this review.

Following this initial pass, we performed a quantitative assessment to measure the alignment between the LLM's output and natural human transliterations. We randomly sampled 20 sentences from the dataset. For the `Romanized Bangla` column, the corresponding `Standard Bangla` sentences were provided to two native Bengali speakers (both undergraduate engineering students) to produce their own romanized versions. For the `Romanized Chakma` column, the Bengali-script `Chakma` sentences were given to two native Chakma speakers (an undergraduate engineering student and a registered nurse). The LLM's romanizations were then systematically compared against the human-generated versions using a suite of standard evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and the BERTScore F1 (Zhang et al., 2019) to evaluate semantic similarity. The results of this validation are presented in Table 1, demonstrating a high degree of correlation. Further examples illustrating the natural variations between the LLM and native speaker romanizations for both Bangla and Chakma can be found in Appendix A (Figures 2 and 3).

Lang.	Comparison	BLEU	METEOR	BERTScore
Bangla	LLM vs. Native 1	0.7188	0.8837	0.9751
	LLM vs. Native 2	0.6474	0.8308	0.9608
Chakma	LLM vs. Native 1	0.5962	0.7203	0.9391
	LLM vs. Native 2	0.5143	0.7034	0.9009

Table 1: Validation Scores: LLM vs Two Native Speakers (Bangla & Chakma)

2.3 Dataset Analysis and Statistics

The final ChakmaBridge dataset contains 807 parallel sentences. A sample from the corpus, showcasing the five-way alignment across all languages and scripts, is presented in Figure 1. We partitioned the data into training (564 sentences), validation (81 sentences), and test (162 sentences) splits, ensuring a standard 70/10/20 distribution. Table 2 provides a detailed breakdown of the dataset's composition, including the number of tokens and vocabulary size for each language across the splits.

Further statistical properties of the corpus are detailed in Table 3. This analysis highlights key linguistic characteristics; for instance, `Romanized Bangla` has a higher mean character length (34.08) than its Bengali-script counterpart, `Standard Bangla` (30.81), reflecting phonetic spelling conventions. Furthermore, the vocabulary size for Bengali-script `Chakma` (1501 unique words) is the largest in the corpus, underscoring the lexical diversity captured.

2.4 Experimental Setup for Baselines

To demonstrate the utility of our dataset and establish robust benchmarks for future research, we conducted a series of machine translation experiments. We utilized two powerful, publicly available multilingual models: mBART-50 (Li et al., 2021) and NLLB-200 (Costa-Jussà et al., 2022). For all experiments, the models were fine-tuned for 25 epochs using a learning rate of 5e-5, a batch size of 8, and gradient accumulation over 2 steps. Performance was evaluated using BLEU, METEOR, and BERTScore F1. We designed six distinct translation tasks to evaluate performance across various language and script modalities: (1) English to Chakma, (2) Standard Bangla to Chakma, (3) English to Romanized Chakma, (4) Standard Bangla to Romanized Chakma, (5) Romanized Bangla to Romanized Chakma, and (6) a multilingual task of English + Standard Bangla to Chakma. In these tasks, the 'Chakma' target refers to its Bengali script representation. These experiments serve as comprehensive baselines for future work on this low-resource language pair.

3 Results and Discussion

To establish robust baselines and demonstrate the utility of ChakmaBridge, we evaluated the performance of two powerful multilingual models, mBART-50 and NLLB-200, across the six trans-

Language	Number of Sentences				Total Tokens				Vocab Size			
	Train	Valid	Test	Total	Train	Valid	Test	Total	Train	Valid	Test	Total
English	564	81	162	807	3334	480	996	4810	903	238	380	1139
Standard Bangla	564	81	162	807	2989	434	901	4324	1002	239	413	1273
Romanized Bangla	564	81	162	807	2995	435	905	4335	1008	237	413	1275
Chakma	564	81	162	807	3127	457	932	4516	1190	272	467	1501
Romanized Chakma	564	81	162	807	3123	455	933	4511	1136	267	453	1431

Table 2: Data Statistics per Language: Sentences, Tokens, Vocabulary (Train, Val, Test)

Column	Mean Char. Length	Max Char. Length	Min Char. Length	Mean Word Count	Max Word Count	Min Word Count	Unique Word Count	Unique Sent. Count
English	30.65	102	7	5.96	16	2	1192	775
Standard Bangla	30.81	83	9	5.36	12	2	1273	789
Romanized Bangla	34.08	95	10	5.37	12	2	1316	784
Chakma	28.62	83	6	5.60	14	2	1501	801
Romanized Chakma	31.39	94	7	5.59	14	2	1484	803
Total Dataset	**31.11**	**102**	**6**	**5.58**	**16**	**2**	**5973**	**3952**

Table 3: Detailed character-level and word-level statistics for each column in the dataset.

lation tasks outlined in Section 2.4. The complete results, measured by BLEU, METEOR, and BERTScore F1, are presented in Table 4.

Overall Performance and Model Comparison. Across most tasks, NLLB-200 generally outperforms mBART-50. For instance, in the Standard Bangla → Chakma task, NLLB-200 achieves a BLEU score of 0.2330, surpassing mBART-50's score of 0.2016. This trend suggests that NLLB-200's broader pre-training on a more diverse set of languages, including those from the Indo-Aryan family, provides a better foundation for fine-tuning on low-resource pairs like Bangla-Chakma. While the BLEU and METEOR scores are modest, which is expected for such a low-resource scenario, the high BERTScore F1 values (consistently above 0.80) indicate that the models are successful in capturing the semantic content of the translations, even when n-gram overlap is limited.

Impact of Source Language. Comparing the performance of English vs. Standard Bangla as the source language reveals an interesting pattern. For translation into Bengali-script Chakma, using Standard Bangla as the source yields consistently better results than using English. NLLB-200 achieves a BLEU score of 0.2330 from Bangla, compared to 0.1975 from English. This is likely due to the significant lexical and syntactic similarities between Bangla and Chakma, both being Eastern Indo-Aryan languages. The shared vocabulary and grammatical structures provide a stronger transfer learning signal.

Challenges of Script Transliteratio. The experiments involving romanized targets highlight the added complexity of script conversion. The task of translating from Standard Bangla to Romanized Chakma proves challenging, with NLLB-200 achieving a BLEU score of 0.1867. This task requires the model to perform both semantic translation (Bangla to Chakma) and phonetic transliteration (Bengali script to Latin script) simultaneously. Interestingly, the direct translation from Romanized Bangla to Romanized Chakma (BLEU of 0.2057 with NLLB-200) performs better, suggesting that when the source and target share the same script, the model can focus more effectively on the linguistic translation task.

The Power of Multilingual Training. The most significant finding is the dramatic performance improvement in the multilingual training setting. By jointly training on both English and Standard Bangla source sentences, the NLLB-200 model's performance skyrockets, achieving a BLEU score of 0.5228, a METEOR score of 0.6486, and a BERTScore F1 of 0.9136. This represents a relative improvement of over 124% in BLEU score compared to the next best individual task (Standard Bangla → Chakma). This substantial gain likely stems from several factors. First, the larger and more diverse training data provides a powerful regularization effect, preventing the model from overfitting on the small ChakmaBridge training set. Second, by being exposed to the syntactically distinct English (SVO) alongside the similar Bangla (SOV), the model is forced to learn more abstract and robust cross-lingual representations rather than relying on surface-level pattern matching between the two related Indo-Aryan languages. This encourages the development of a more generalized translation capability.

Translation Direction	Model	BLEU	METEOR	BERTScore F1
English → Chakma	mBART-50	0.1788	0.3449	0.8340
	NLLB-200	0.1975	0.3710	0.8474
Standard Bangla → Chakma	mBART-50	0.2016	0.3640	0.8451
	NLLB-200	0.2330	0.4042	0.8578
English → Romanized Chakma	mBART-50	0.1836	0.3217	0.8794
	NLLB-200	0.1489	0.3038	0.7911
Standard Bangla → Romanized Chakma	mBART-50	0.1664	0.3365	0.7916
	NLLB-200	0.1867	0.3505	0.8857
Romanized Bangla → Romanized Chakma	mBART-50	0.1980	0.3914	0.8247
	NLLB-200	0.2057	0.3817	0.8246
Multilingual (English + Standard Bangla) → Chakma	mBART-50	0.2744	0.4608	0.8646
	NLLB-200	**0.5228**	**0.6486**	**0.9136**

Table 4: Translation Performance Across Directions and Models (BLEU, METEOR, BERTScore F1)

Qualitative Error Analysis. To better understand the models' performance beyond quantitative metrics, we conducted a qualitative analysis of common errors in the output of the best-performing multilingual model (NLLB-200). We identified three primary categories of errors: (1) Lexical Choice Errors, where the model selects a semantically related but contextually incorrect word (e.g., translating "house" as "building"). (2) Grammatical Errors, primarily involving incorrect postpositions or verb conjugations, which are common challenges in morphologically rich languages like Chakma. (3) Omissions, where the model fails to translate a specific noun or adjective from the source sentence, leading to a loss of detail. These errors suggest that while the model effectively captures the overall meaning, as indicated by the high BERTScore, it still struggles with fine-grained lexical and syntactic details, a typical challenge when fine-tuning on a small dataset. Future work could explore data augmentation or constrained decoding techniques to mitigate these specific error types.

4 Conclusion

In this work, we introduced ChakmaBridge, a novel five-way parallel corpus designed to advance NLP research for the low-resource, endangered Chakma language. By augmenting the foundational MELD dataset with human-validated, LLM-generated romanized parallels for both Chakma and Bangla, we have created a unique resource that reflects modern digital communication practices. Our comprehensive baseline experiments demonstrate the dataset's utility for a range of machine translation tasks across different scripts. Notably, our findings reveal that a multilingual training approach, combining both English and the linguistically closer Bangla, dramatically improves translation quality into Chakma, boosting performance by over 124% in BLEU score. We release ChakmaBridge to the community with the hope that it will spur further research, aid in the development of practical translation tools, and contribute to the digital preservation of the Chakma language and its rich cultural heritage. Future efforts should build upon this by focusing on dataset expansion, exploring data augmentation strategies, and incorporating direct support for the native Chakma script.

Limitations

While ChakmaBridge represents a significant step forward, we acknowledge several limitations that offer clear avenues for future research. First, the primary limitation is the modest size of the corpus. At only 807 sentences, it is insufficient for training robust neural models from scratch and restricts the generalizability of our findings. While our multilingual approach serves as a mitigation strategy by leveraging high-resource data, future work should explore targeted data augmentation techniques like back-translation to artificially expand the training set. Second, our romanization methodology has inherent constraints. The LLM-generated text, while validated, is likely more standardized and cleaner than the noisy, orthographically diverse "in-the-wild" romanization found in user-generated content. This could limit the real-world applicability of models trained on this data. Furthermore, our validation protocol, based on a small sample of 20 sentences, provides a preliminary quality check but is not extensive enough to offer definitive statistical confidence. Finally, our work is scoped to Chakma as represented in the Bengali script. A key omission is the lack of evaluation or direct support for the native Chakma (Ajhā Pāṭh) script. Developing resources and models capable of processing and generating the native script is a critical and necessary next step for the comprehensive digital revitalization of the language.

Acknowledgments

This work was supported by Southeast University, Bangladesh. We would like to thank Writu Raj Chakma and Akhi Chakma for their valuable assistance in validating the dataset.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Bangladesh Bureau of Statistics. 2023. Table a-1.4 ethnic population by group and sex. in: *Statistics*, 2023. Based on Bangladesh Bureau of Statistics 2021 data.

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*.

Nikhil Chakma and Mathilde Maitrot. 2016. How ethnic minorities became poor and stay poor in bangladesh: A qualitative enquiry. *EEP/Shiree, July*.

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*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838, Online. Association for Computational Linguistics.

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.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kanchon Kanti Podder, Ludmila Emdad Khan, Jyoti Chakma, Muhammad EH Chowdhury, Proma Dutta, Khan Md Anwarus Salam, Amith Khandakar, Mohamed Arselene Ayari, Bikash Kumar Bhawmick, SM Arafin Islam, and 1 others. 2023. Self-chakmanet: A deep learning framework for indigenous language learning using handwritten characters. *Egyptian Informatics Journal*, 24(4):100413.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

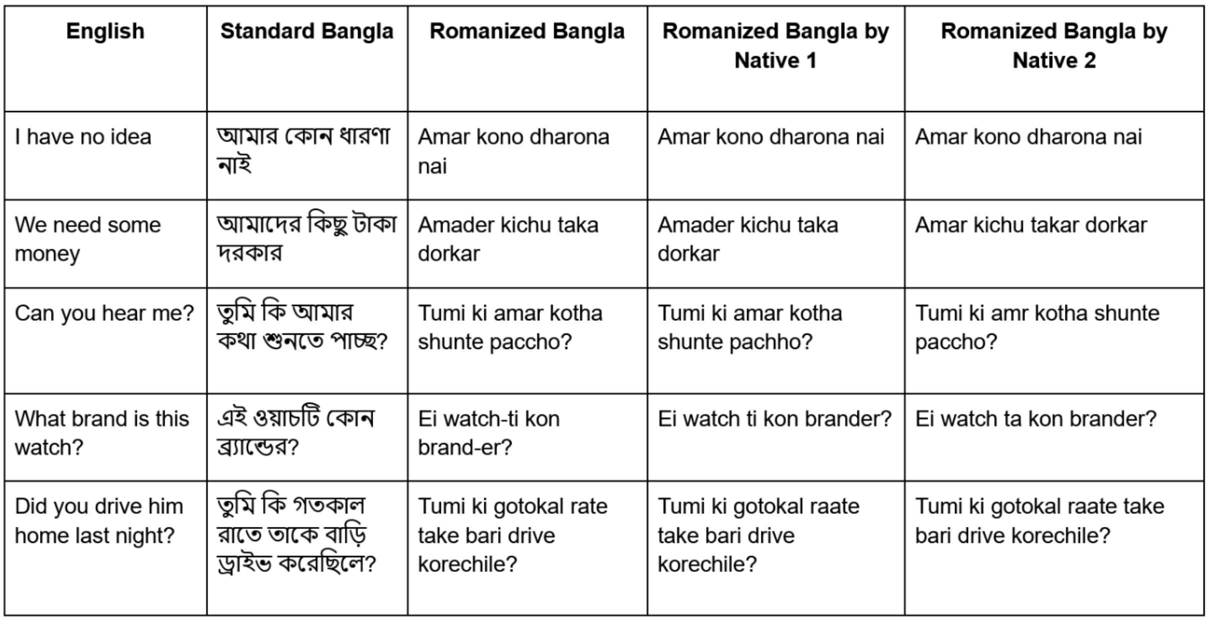
Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Jonali Saikia and Mary Kim Haokip. 2023. Language endangerment with special refrence to chakma. *Journal of English Language and Literature*, 10(3).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

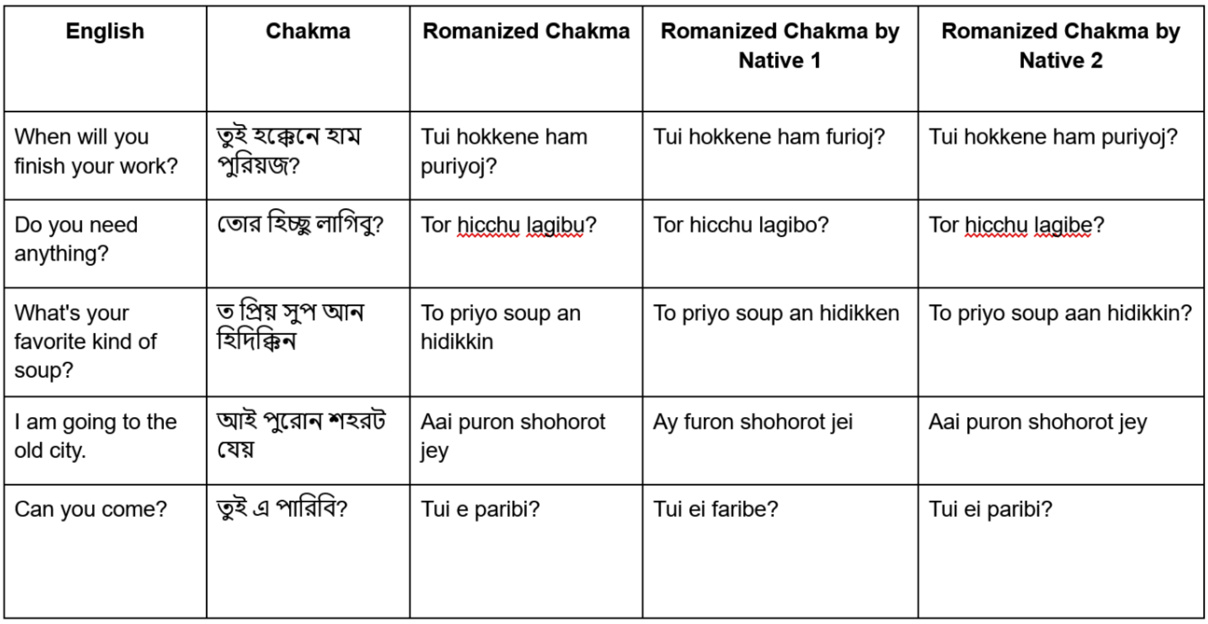
A Dataset and Validation Samples

This appendix provides visual examples from the ChakmaBridge dataset. Figure 2 and Figure 3 present examples from our validation protocol, illustrating the comparison between the LLM-generated romanizations and those produced by native speakers.



English	Standard Bangla	Romanized Bangla	Romanized Bangla by Native 1	Romanized Bangla by Native 2
I have no idea	আমার কোন ধারণা নাই	Amar kono dharona nai	Amar kono dharona nai	Amar kono dharona nai
We need some money	আমাদের কিছু টাকা দরকার	Amader kichu taka dorkar	Amader kichu taka dorkar	Amar kichu takar dorkar
Can you hear me?	তুমি কি আমার কথা শুনতে পাচ্ছ?	Tumi ki amar kotha shunte paccho?	Tumi ki amar kotha shunte pachho?	Tumi ki amr kotha shunte paccho?
What brand is this watch?	এই ওয়াচটি কোন ব্র্যান্ডের?	Ei watch-ti kon brand-er?	Ei watch ti kon brander?	Ei watch ta kon brander?
Did you drive him home last night?	তুমি কি গতকাল রাতে তাকে বাড়ি ড্রাইভ করেছিলে?	Tumi ki gotokal rate take bari drive korechile?	Tumi ki gotokal raate take bari drive korechile?	Tumi ki gotokal raate take bari drive korechile?

Figure 2: Validation samples for Romanized Bangla, comparing the LLM-generated output with versions from two native speakers. The table highlights the natural diversity in phonetic romanization; for instance, the sound ’ch’ is variably rendered as ’cch’ (paccho) or ’chh’ (pachho). Such nuances demonstrate the challenge of standardizing romanized text and the importance of our validation process.



English	Chakma	Romanized Chakma	Romanized Chakma by Native 1	Romanized Chakma by Native 2
When will you finish your work?	তুই হক্কেনে হাম পুরিয়জ?	Tui hokkene ham puriyoj?	Tui hokkene ham furioj?	Tui hokkene ham puriyoj?
Do you need anything?	তোর হিচ্ছু লাগিবু?	Tor hicchu lagibu?	Tor hicchu lagibo?	Tor hicchu lagibe?
What's your favorite kind of soup?	ত প্রিয় সুপ আন হিদিক্কিন	To priyo soup an hidikkin	To priyo soup an hidikken	To priyo soup aan hidikkin?
I am going to the old city.	আই পুরোন শহরট যেয়	Aai puron shohorot jey	Ay furon shohorot jei	Aai puron shohorot jey
Can you come?	তুই এ পারিবি?	Tui e paribi?	Tui ei faribe?	Tui ei paribi?

Figure 3: Validation samples for Romanized Chakma. This figure demonstrates key phonetic distinctions often found in native Chakma romanization that may be missed by automated systems. For example, the alternation between ’p’ and ’f’ sounds (e.g., puriyoj vs. furioj and paribi vs. faribe) reflects a common phonological variation. Capturing these subtle, real-world differences is crucial for developing robust transliteration and translation models.