BhasaBodh: Bridging Bangla Dialects and Romanized Forms through Machine Translation

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

Abstract

While machine translation has made significant strides for high-resource languages, many regional languages and their dialects, such as the Bangla variants Chittagong and Sylhet, remain underserved. Existing resources are often insufficient for robust sentence-level evaluation and overlook the widespread real-world practice of romanization, the common practice of typing native languages using the Latin script in digital communication. To address these gaps, we introduce BhasaBodh, a comprehensive benchmark for Bangla dialectal machine translation. We construct and release a sentence-level parallel dataset for Chittagong and Sylhet dialects aligned with Standard Bangla and English, create a novel romanized version of the dialectal data to facilitate evaluation in realistic multi-script scenarios, and provide the first comprehensive performance baselines by fine-tuning two powerful multilingual models, NLLB-200 and mBART-50, on seven distinct translation tasks. Our experiments reveal that mBART-50 consistently outperforms NLLB-200 on most dialectal and romanized tasks, achieving a BLEU score as high as 87.44 on the Romanized-to-Standard Bangla normalization task. However, complex cross-lingual and cross-script translation remains a significant challenge. BhasaBodh lays the groundwork for future research in low-resource dialectal NLP, offering a valuable resource for developing more inclusive and practical translation systems.

1 Introduction

Impressive advancements have been made in machine translation (MT), with a focus on high-resource languages like Mandarin and English (Costa-Juss‘a et al., 2022; Fan et al., 2021). Many regional languages, which are often spoken by millions of people, are still in the digital shadows, indicating that this development has not been dispersed equally. According to Sultana et al. (2025), this is especially true for the Bangla language family, where significant dialects like Chittagong and Sylhet are linguistically different from Standard Bangla yet do not have specialized NLP resources. The lack of established evaluation criteria for these languages greatly impedes the development and effective assessment of MT systems meant to serve these groups.

To overcome this data shortage, recent initiatives like ONUBAD (Sultana et al., 2025) have started to provide parallel data across many Bangla dialects. However, the typical unit for evaluating translation fluency and coherence is sentence-level MT assessment, for which the available resources are not optimum. Furthermore, they often ignore romanization, a common occurrence in the actual world. Due to input method constraints or convenience, users often utilize the Latin script to type their local languages in informal digital communication, such as social media and messaging applications. This results in a multi-script translation situation that is beyond the capabilities of current models and benchmarks. Our dataset and the code for our baseline experiments are publicly available on GitHub[1].

BhasaBodh is a representative and high-quality benchmark for Bangla dialectal machine translation that we offer in this work. Three significant contributions are made by our work:

- After a thorough filtering and balancing procedure, a sentence-level parallel assessment dataset for the Chittagong and Sylhet dialects is created and made available. It is in line with Standard Bangla and English and is taken from the ONUBAD corpus.
- In order to facilitate assessment under realistic multi-script situations that replicate user-

[*]Authors contributed equally to this work.

[1]https://github.com/borhanitrash/BhashaBodh

generated material, a new romanized version of the benchmark is included, created using Gemini 2.5 Pro (Comanici et al., 2025).

- NLLB-200 (Costa-Juss‘a et al., 2022) and mBART-50 (Tang et al., 2020), two powerful multilingual MT baselines, are thoroughly tested on seven translation tasks, offering the first thorough performance study in this field. Experiments reveal both strengths and unresolved issues in dialectal transfer, cross-lingual, and cross-script settings.

2 Related Work

Our research focuses on the unique difficulties of dialect processing, low-resource NLP, and multilingual machine translation. We discuss pertinent material below.

2.1 Multilingual MT with Little Resources

Considerable progress has been made in creating models that can translate across a variety of languages. Largely multilingual models with remarkable zero-shot with few-shot capabilities include the M2M-100 (Fan et al., 2021) along with NLLB-200 (Costa-Juss‘a et al., 2022). Similarly, by using cross-lingual representations, pre-trained sequence-to-sequence models like mBART (Liu et al., 2020) and its translation-tuned variation mBART-50 (Tang et al., 2020) have shown good performance across a broad variety of languages. However, for really low-resource languages (Lin et al., 2020), a category that appropriately characterizes the majority of regional dialects that are either absent or badly underrepresented in training data, these models' performance often deteriorates dramatically.

2.2 Assessment Standards for MT

Establishing trustworthy standards is essential for tracking MT development. High-quality multilingual test sets are made available by initiatives like TICO-19 (Anastasopoulos et al., 2020) and FLORES-101 (Goyal et al., 2022). The significance of representative as well as linguistically varied standards is emphasized by more recent initiatives like CCEval (Lou et al., 2023), especially for translation that is centered on Chinese. The emphasis on standard, well-written English, often from formal realms like journalism or Wikipedia, unites these standards. In order to better represent real-world use scenarios, our work adds the unique feature of multi-script assessment and applies this idea to the understudied field of dialectal translation.

2.3 Dialect and Code-Switching NLP

The necessity for NLP tools to support dialect speakers and deal with linguistic variance is becoming more widely acknowledged. This includes dialect-specific MT, dialect identification (Zaidan and Callison-Burch, 2011), and dialectal corpora construction (e.g., Sultana et al., 2025 for Bangla). The "noisy" character of user-generated dialectal writing, which sometimes includes code-switching and non-standard spelling, is a major obstacle. In line with studies on transliteration and modeling for social media writing, we directly address this difficulty by introducing a standard for romanized dialects (Baldwin et al., 2015).

2.4 Materials for the Bangla Language and Dialect

Although NLP has given Standard Bangla more attention (Bhattacharjee et al., 2021), there are still few resources available for its dialects. An important addition is the ONUBAD corpus (Sultana et al., 2025), which offers parallel data for a number of Bangla dialects at various linguistic levels. In order to provide deployable technology for dialect communities, our work directly builds upon ONUBAD by improving it for sentence-level MT assessment and extending it to handle the real-world situation of romanized input.

3 Dataset Constructions

3.1 Data Creation and Augmentation

The ONUBAD dataset (Sultana et al., 2025) was used as the starting point since it offers parallel data for Chittagong, Sylhet, and Barisal that are in line with Standard Bangla and English. Barisal was excluded to prioritize depth over breadth, focusing on Chittagong and Sylhet, the most linguistically distant and resource-poor dialects, for a more targeted analysis within our scope. In order to create the BhasaBodh dataset, only sentence-level pairings were used from this source. Filtering to keep only sentence-level alignments, cleaning with tokenization correction, orthographic harmonization, and punctuation normalization, balancing to guarantee equal representation of Chittagong and Sylhet sentences, and organizing the data into a

English Translation	Standard Bangla Language	Romanized Bangla	Chittagong Language	Romanized Chittagong Language	Sylhet Language	Romanized Sylhet Language
Lying on the bed, I was watching TV	বিছানায় শুয়ে টিভি দেখছিলাম	Bichanay shuye TV dekhchilam	চিছানাত লুডি টিভি দেইক্কিলাম	Sisanat ludi TV deikkilam	বিছানায় হুতি টিভি দেখাত আছিলাম	Bisanay huti TV dekhat asilam
Who sent you?	আপনাকে কে পাঠিয়েছে	Apnake ke pathiyeche	অনরে হন ফাদইয়ে	Onore hon fadoiye	আফনারে কে পাঠাইছে?	Afnare ke faṭaisse?
Why do you feel sleepy?	তোমার ঘুম আসে কেন?	Tomar ghum ashe keno?	তুয়ার ঘুম কিল্লায় আয়েদ্দে	Tuar ghum killay aiyedde	তুমিতাইন ঘুমাও কিতার লাগি?	Tumitain gumao kitar lagi?
Bought me a red rose	বন্ধু আমাকে একটি লাল গোলাপ কিনে দিয়েছে	Bondhu amake ekti lal golap kine diyeche	বন্ধু আরে ওগগা লাল গোলাপ কিনি দিয়িএ	Bondhu are ogga lal gulap kini diye	বন্ধু মোর লাগি এখটা লাল গোলাপ কিনিয়া দিছে	Bondhu mor lagi ekhta lal gulap kiniya dise
It might be something beautiful.	সুন্দর কিছু হতে পারে	Shundor kichu hote pare	সুন্দর কিসু অইত ফারে	Shundor kisu oit fare	হুন্দর কিচ্ছু হইতা পারে	Hundor kissu hoita fare

Figure 1: Sample entries from the BhasaBodh dataset, illustrating the multi-way parallel alignment across English, Standard Bangla, Chittagong, Sylhet, and their respective romanized forms.

three-column format with dialect sentences, Standard Bangla, and English were the steps in the preparation process. Gemini 2.5 Pro (Comanici et al., 2025) was used to produce Latin-script versions of Chittagong and Sylhet phrases in order to better enable romanized input. Output only the romanized version. This ensured authenticity by mimicking user-generated content. The dataset was expanded into a multi-script resource that included both native and romanized versions after the model was particularly instructed to generate natural and informal romanizations indicative of social media use.

3.2 Validation

To perform a preliminary quality check on the synthetically generated data, we employed a manual validation process. We acknowledge that this validation is not exhaustive. Specifically, we used a small spot-check of 20 samples that were randomly selected and evaluated by two native validators. The first validator was an undergraduate engineering student, while the second was a Bachelor of Business Administration (BBA) student. While their feedback provides an initial quality signal, we recognize that a larger sample size and a more diverse group of annotators would be necessary to make more generalizable claims about the dataset's overall quality and representativeness. Their validations were then compared against the outputs generated by our LLM model. The quantitative results of this comparison are presented in Table 1.

Dialect	BLEU	METEOR	BERTScore_F1
Sylhet	56.7109	0.7227	0.9519
Chittagong	79.9174	0.7745	0.9626

Table 1: Validation Results: LLM vs. Native for Bangla Dialects

3.3 Dataset Statistics

Each of the two dialects (Chittagong and Sylhet) has 980 sentences in the final dataset, each having references to Standard Bangla and English. For both dialects, romanized versions were created. In order to concentrate on the two dialects with the fewest resources. Details are in Table 2. To provide a concrete example of the dataset's structure, a sample of the multi-way parallel data is presented in Figure 1.

Split	#Sent	Len	Script
English	980	11.4	Latin
Std. Bangla	980	9.9	Bangla
Chittagong	980	10.2	Bangla
Rom. Chitt.	980	9.5	Latin
Sylhet	980	9.8	Bangla
Rom. Sylhet	980	9.7	Latin

Table 2: Dataset statistics (#Sent = number of sentences, Len = avg. tokens).

3.4 Experiments

In order to cover a variety of situations, seven machine translation experiments were created. For

example, there was a cross-lingual baseline from English to Standard Bangla, dialect generation from Standard Bangla to Chittagong and Sylhet, script normalization from Romanized Bangla to Standard Bangla, direct dialect-to-dialect translation between Chittagong and Sylhet, and a difficult cross-lingual, cross-script task from English to Romanized Sylhet. Two multilingual models were trained and assessed for every task. Hugging Face's `Seq2SeqTrainer` was used to fine-tune mBART-50 (`facebook/mbart-large-50-many-to-many-mmt`) with `en_XX` → `bn_IN` language codes, and NLLB-200 (distilled 600M, `facebook/nllb-200-distilled-600M`) with `eng_Latn` → `ben_Beng` language codes. Using a batch size of 8, a learning rate of 5e-5, a weight decay of 0.01, 50 warm-up steps, and BLEU as the checkpointing measure, both models were trained for 25 epochs on Kaggle GPUs.

3.5 Translation Tasks

In order to thoroughly assess cross-lingual, cross-dialect, and cross-script situations, seven essential translation tasks were established. Standard Bangla to Chittagong and Standard Bangla to Sylhet concentrated on dialectal creation, whereas English to Standard Bangla was the baseline high-resource assignment. While Romanized Bangla to Romanized Chittagong allowed for cross-dialect translation inside the romanized space, Romanized Bangla to Standard Bangla was intended as a script standardization effort. Chittagong to Sylhet was used to assess direct dialect-to-dialect translation, whereas English to Romanized Sylhet, the most difficult scenario, combined cross-lingual and cross-script difficulties.

3.6 Baseline Models and Setup

The BhasaBodh dataset was used to directly refine both models, in contrast to previous zero-shot assessments. To optimize training data on the smaller dataset, the training, validation, and test splits were 70/10/20 for mBART-50 and 80/10/10 for NLLB-200. These splits were chosen based on model architecture: the larger mBART-50 (610M parameters) benefits from a higher training proportion (70%) to leverage its capacity without overfitting, while the distilled NLLB-200 (600M parameters) uses 80% training to maximize data utilization on low-resource tasks, as validated in preliminary cross-validation experiments. Using a batch size of 8, a learning rate of 5e-5, a weight decay of 0.01, and 50 warm-up steps, both models were trained for up to 25 epochs. To mitigate the risk of overfitting on our small dataset, we employed an early stopping strategy based on the validation set's BLEU score, using it as the primary checkpointing metric.

3.7 Evaluation Metrics

For every assignment, we report BERTScore-F1, METEOR, and BLEU. BERTScore evaluates semantic similarity, while BLEU and METEOR capture n-gram overlap. We acknowledge that these metrics, particularly those based on n-gram overlap, may not fully capture the nuances of dialectal and orthographic variations. Future work would benefit from incorporating character-level metrics like chrF++ to better handle spelling differences and learned semantic metrics like COMET to provide a more robust assessment of translation quality.

4 Results and Discussion

The outcomes of the experiment are summarized in Table 3. With a BLEU score of 87.44, mBART-50 performed best on the Romanized Bangla → Standard Bangla task, demonstrating that the consistent orthography from the synthetically generated romanization helps reduce variability and simplifies the normalization task. Dialect-to-dialect translation also achieved strong results; for example, mBART-50 reached a BLEU of 74.36 on Chittagong → Sylhet, owing to its denoising pretraining that is effective for noisy, non-standard text.

NLLB-200 produced competitive results on high-resource directions such as English → Standard Bangla (BLEU = 65.97), benefiting from its large-scale multilingual training and efficient inference. However, it consistently underperformed mBART-50 in both BLEU and METEOR on dialectal and romanized tasks.

The most challenging case was English → Romanized Sylhet, where both models achieved BLEU scores below 41, highlighting the difficulty of cross-lingual, cross-script translation. A brief qualitative error analysis reveals that common errors in this task stem from syntactic divergences and the models' inability to translate idiomatic English phrases into a non-standard, romanized dialectal structure. Overall, mBART-50 demonstrates greater robustness to dialectal noise, while NLLB-200 shows advantages in high-resource pairs due

Task	Model	BLEU	METEOR	BERTScore-F1
English → Standard Bangla	NLLB-200	65.97	0.721	0.938
	mBART-50	49.18	0.634	0.911
Standard Bangla → Chittagong	NLLB-200	61.20	0.642	0.923
	mBART-50	64.02	0.651	0.928
Standard Bangla → Sylhet	NLLB-200	65.47	0.713	0.942
	mBART-50	69.04	0.737	0.950
Romanized Bangla → Standard Bangla	NLLB-200	79.13	0.809	0.970
	mBART-50	87.44	0.869	0.976
Romanized Bangla → Romanized Chittagong	NLLB-200	51.14	0.553	0.890
	mBART-50	59.20	0.628	0.943
Chittagong ↔ Sylhet	NLLB-200	62.74	0.665	0.923
	mBART-50	74.36	0.738	0.941
English → Romanized Sylhet	NLLB-200	40.11	0.512	0.909
	mBART-50	39.00	0.503	0.909

Table 3: Experimental results across Standard Bangla, Chittagong, Sylhet, and their romanized variants.

to its architecture and training corpus.

5 Conclusion

Chittagong and Sylhet romanized versions were added to the BhasaBodh dataset, which was initially presented as a sentence-level machine translation benchmark for Bangla dialects. In contrast to previous research, NLLB-200 and mBART-50 were both optimized on this dataset, providing repeatable baselines for seven translation tasks. The findings show that mBART-50 consistently beats NLLB-200 in the majority of dialectal and romanized tasks, and that romanized input greatly facilitates normalization, attaining BLEU scores up to 87.44. NLLB-200 is often less accurate even if it provides quicker inference. Machine translation between languages and scripts is still quite difficult. All things considered, this benchmark lays the groundwork for low-resource Bangla dialectal NLP, facilitating further studies in multi-dialect transfer learning, dialect-aware pretraining, and the application of larger-scale models to these specific tasks.

Limitations

Our work, while establishing an important baseline, has several limitations.

First, the BhasaBodh dataset, though carefully curated, is modest in scale (980 sentences per dialect). While suitable for a low-resource setting, this size may not be large enough to stabilize performance estimates across different random seeds or support robust subgroup analyses. We agree that expanding the dataset would significantly improve the training of more robust models.

Second, our romanized data was synthetically generated using a large language model. This approach likely compresses the natural diversity of community spellings and code-switching patterns found in authentic user-generated text and may not be fully representative. This synthetic uniformity may also explain why the normalization task (Romanized Bangla → Standard Bangla) achieved unusually high scores. Furthermore, our manual validation was conducted on an extremely small sample by native speakers who are not linguistics experts, and without reporting inter-annotator agreement, which impacts the statistical reliability of the quality assessment.

Finally, our experimental scope and evaluation have constraints. Our analysis is limited to two multilingual models and relies solely on automatic metrics. Future work requires a more comprehensive human evaluation to assess nuances in dialectal and romanized outputs. Furthermore, benchmarking larger-scale models (e.g., in the 7B to 13B parameter range and beyond) in both few-shot prompting and full fine-tuning setups would provide deeper insights. The narrow scope of two dialects and primarily asymmetric translation directions also limits the external validity of our findings. These limitations suggest promising directions for future work, including expanding the dataset with authentic romanized text, conducting broader comparisons across model architectures, and reporting variance across multiple training runs.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzm'an, Junjie Hu, Macduff Hughes, Philipp Koehn, and 1 others. 2020. Tico-19: the translation initiative for covid-19. *arXiv preprint arXiv:2007.01788*.

Timothy Baldwin, Marie-Catherine De Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the workshop on noisy user-generated text*, pages 126–135.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Md Saiful Islam, M Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

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-Juss'a, 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*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzm'an, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lianzhang Lou, Xi Yin, Yutao Xie, and Yang Xiang. 2023. Cceval: A representative evaluation benchmark for the chinese-centric multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10176–10184.

Nusrat Sultana, Rumana Yasmin, Bijon Mallik, and Mohammad Shorif Uddin. 2025. Onubad: A comprehensive dataset for automated conversion of bangla regional dialects into standard bengali dialect. *Data in Brief*, 58:111276.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Omar Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.