# BanglaTLit: A Benchmark Dataset for Back-Transliteration of Romanized Bangla

Md Fahim[1,2] Fariha Tanjim Shifat[1] Fabiha Haider[1] Deeparghya Dutta Barua[1]
Md Sakib Ul Rahman Sourove[1] Md Farhan Ishmam[1,3] Md Farhad Alam[1]

[1]Penta Global Limited [2]CCDS Lab, Independent University [3]Islamic University of Technology

## Background

- **Romanized/Transliterated Bangla:** Uses phonetically similar Latin scripts to represent Bangla syllables.
- **Back-transliteration:** The task of generating the native scripts corresponding to the romanized text while closely aligning to the phonetic meaning.

## Motivation

- Due to the phonemic orthography of Bangla, i.e. it is written as it sounds, the same sentence can have multiple transliterations. But, these variations must back-transliterate to the same original sentence *[many to one mapping]*.
- Back-transliteration must adhere to the grammatical rules of the native language.
- Current Language Models (LMs) are trained on limited transliterated Bangla texts.
- No existing large-scale back-transliteration corpus in Bangla to train LMs.

We introduce large-scale romanized Bangla pre-training corpus, back-transliteration dataset, and novel dual-encoder seq2seq models.

## BanglaTLit Datasets

### BanglaTLit-PT

1. **Sourcing:** 245.7k romanized Bangla samples sourced from TrickBd website, Facebook, YouTube, blogs, and Wikipedia.
2. **Cleaning:** Removed BBcodes, hyperlinks, single-word samples, duplicate samples, and redundant white spaces.

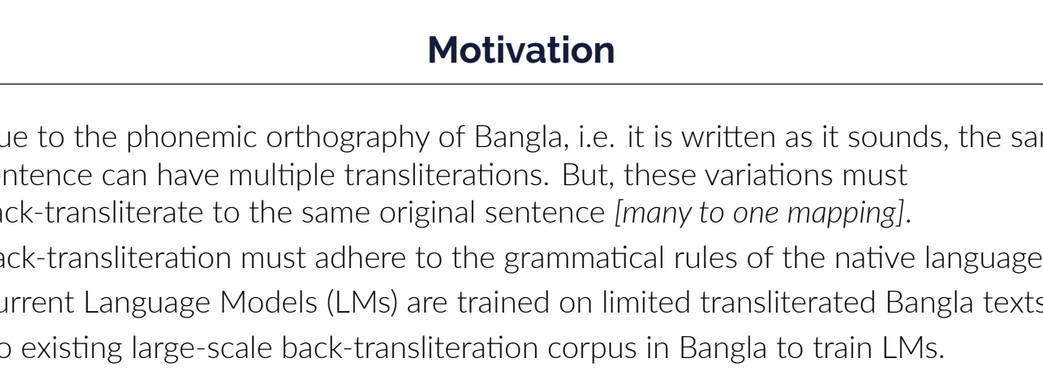


Figure 1. Dataset creation pipeline of BanglaTLit-PT and BanglaTLit.

### BanglaTLit

1. **Filtering:** 42.7k samples filtered from BanglaTLit from our sourced TrickBd dataset, Madhani et al. [2], and Shibli et al. [3].
2. **Annotation:** Conducted by 12 native Bangla speakers with a minimum of undergraduate-level education and proficiency in understanding romanized texts.
3. **Validation:** Three Bangla linguistic experts re-annotated 1,000 random samples. Similarity scores confirm strong alignment between expert and standard annotations.
4. **Key Statistics:** For the romanized and back-transliterated samples, the character length range is [3, 1406] and [2, 212], the word count range is [4, 1347] and [2, 226], and number of unique words is 81.8k and 60.6k respectively.

## Model Architecture

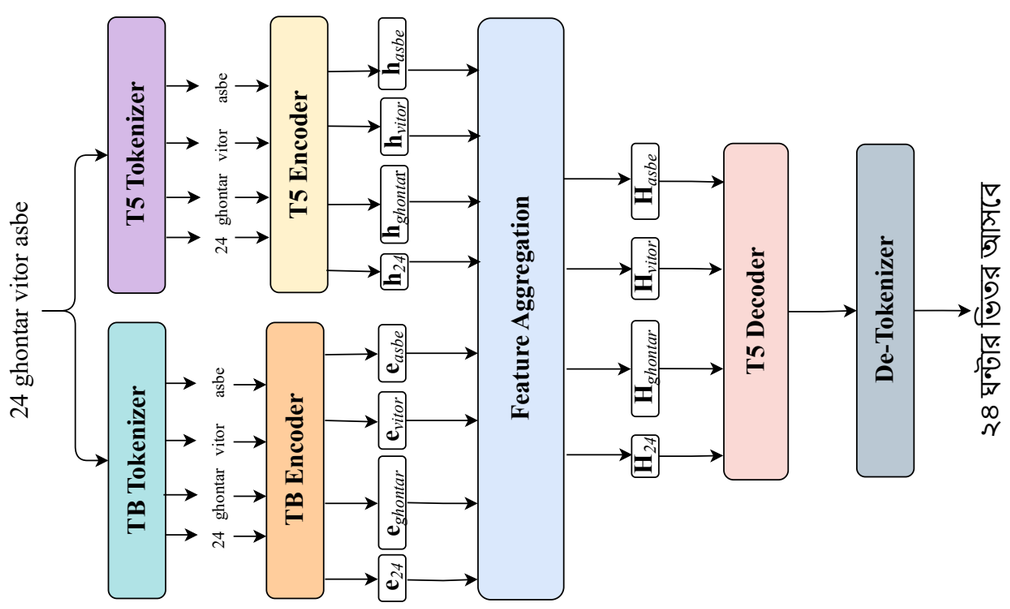


Figure 2. Overview of our dual-encoder-decoder language model.

## Research Question

How can we enhance the representation of romanized Bangla for automatic back-transliteration using seq2seq models?

## Methodology

### Transliterated Bangla (TB) Encoder

We further pretrain the pretrained language models on the BanglaTLit-PT corpus using the Masked Language Modeling loss [1]. For the dataset distribution $\mathcal{D}$, the sentence $S \sim \mathcal{D}$, $S = \{t_1, ..., t_T\}$, mask indices $m \in \mathbb{N}^M$, and training parameters $\theta$, the negative log-likelihood objective is defined as,

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{S\sim\mathcal{D}}\left[\log P_\theta(t_m|t_{\setminus m})\right].$$

### TB Encoder Aggregated T5 Model

For a given text $S$, we obtain representations from the T5 and TB encoders

$$\text{T5}(S) = \mathbf{h} = \{h_1, h_2, \ldots, h_n\}$$

$$\text{TB}(S) = \mathbf{e} = \{e_1, e_2, \ldots, e_n\}$$

The representations $\mathbf{h}$ and $\mathbf{e}$ are then aggregated using either sum-based aggregation, $\mathbf{H}_{\text{sum}} = \mathbf{h} + \mathbf{e}$, or concatenation-based aggregation, $\mathbf{H}_{\text{concat}} = [\mathbf{h}; \mathbf{e}]$, where $\mathbf{H} = \{H_1, H_2, \ldots, H_m\}$. The aggregated representation is passed to the T5 decoder to produce the corresponding Bangla text.

## Experimental Results

The T5-based seq2seq models and GPT Large Language Models (LLMs) perform well on Bangla back-transliteration. Our proposed dual-encoder-decoder architecture outperformed the T5 encoders by a fair margin, with TB-XLM-R + BanglaT5-NMT achieving the best performance in most metrics.

| Model | ROUGE Score | | | BLEU Score | | | BERT Score (F1) | METEOR Score |
|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | BLEU | Brevity Penalty | Length Ratio | | |
| **Encoder-Decoder LM** | | | | | | | | |
| mT5 | 56.02 | 19.83 | 55.90 | 12.48 | 76.13 | 0.82 | 86.43 | 48.71 |
| byteT5 | 15.40 | 1.71 | 14.91 | 6.8e-5 | 11.28 | 0.25 | 72.50 | 6.88 |
| BanglaT5-small | 39.59 | 8.46 | 39.58 | 4.14 | 84.29 | 0.94 | 80.65 | 32.72 |
| BanglaT5 | 73.06 | 33.00 | 73.13 | 31.09 | 91.16 | 0.95 | 92.71 | 69.12 |
| BanglaT5_nmt_en_bn | 75.74 | 34.84 | 76.14 | 36.19 | **98.71** | 1.08 | 94.05 | 74.07 |
| **Prompt-based LLM** | | | | | | | | |
| GPT-3.5 Turbo (0-shot) | 66.21 | 26.18 | 66.64 | 20.73 | 97.94 | 1.11 | 90.06 | 59.97 |
| GPT-4 Turbo (0-shot) | 71.71 | 31.54 | 71.96 | 26.56 | 97.27 | 1.07 | 91.65 | 65.10 |
| GPT-4o (0-shot) | 66.62 | 26.96 | 67.24 | 19.28 | 98.22 | 1.11 | 89.37 | 58.88 |
| LLaMa3-8B (3-shot) | 56.05 | 17.34 | 56.56 | 11.01 | 95.80 | 1.04 | 86.61 | 46.81 |
| **Dual Encoder-Decoder LM (Ours)** | | | | | | | | |
| TB-BanglishBERT + BanglaT5 | 75.14 | 34.65 | 75.13 | 32.82 | 92.25 | 0.96 | 93.83 | 72.34 |
| TB-BanglishBERT + BanglaT5_NMT | 77.27 | 35.98 | 78.32 | 35.18 | 96.58 | **0.97** | 98.22 | 75.37 |
| TB-XLM_R + BanglaT5 | 76.03 | 35.14 | 76.24 | 33.18 | 95.16 | 0.96 | 94.15 | 74.42 |
| TB-XLM_R + BanglaT5_NMT | **78.92** | **36.56** | **79.75** | **36.07** | 98.29 | 1.05 | **98.82** | **78.14** |

Table 1. BanglaTLit test set benchmark results on our baselines.

## Conclusion

BanglaTLit establishes the first automated romanized Bangla back-transliteration system, that can be expanded to include other linguistic variations and dialects.

## Acknowledgments

We express our deepest gratitude to our sponsor, Penta Global Limited, Bangladesh.

**Authors' Note:** *We honor the brave souls of the July student movement, reflecting on their courage, resilience, and fight for justice.*

## References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. Bhasa-abhijnaanam: Native-script and romanized language identification for 22 Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] GM Shahariar Shibli, Md Tanvir Rouf Shawon, Anik Hassan Nibir, Md Zabed Miandad, and Nibir Chandra Mandal. Automatic back transliteration of romanized bengali (banglish) to bengali. *Iran Journal of Computer Science*, 6(1):69–80, 2023.