

JH_NLP_APAShreya2: LoRA Fine-Tuning of NLLB-200 for English→Manipuri Machine Translation at WMT 2026

Aparajita Sinha¹, Monika Agarwal², Shreya Narayana Bhat²,
Shreya Bonal², Aryan Agrahari¹

¹National Institute of Technology, Agartala, Tripura, India

²PES University, Bengaluru, Karnataka, India

aparajitas824@gmail.com, monika.goyal@pes.edu, iamshreyabhat@gmail.com

shreyabonal11@gmail.com, aryanagrahari215@gmail.com

Abstract

This paper describes the submission of team **JH_NLP_APAShreya2** to the English→Manipuri (Bengali-script) track of the WMT 2026 Low-Resource Indic Machine Translation shared task. Our system adapts the multilingual NLLB-200-distilled-600M model using Low-Rank Adaptation (LoRA) on the official English→Manipuri parallel corpus. The corpus contains 23,687 sentence pairs. We used no external parallel or monolingual data, synthetic data, or back-translation. We conducted a progressive sequence of experiments with different data-pool sizes and training durations. The final model was trained for seven epochs and used to translate all 1,000 English source sentences in the official blind test set. The primary submission obtained 8.03 BLEU, 20.15 METEOR, 82.56 TER, 40.93 chrF++, 84.86 BERTScore, and 68.18 COMET. The system ranked within the top three primary systems according to five of the six reported evaluation metrics. These results indicate that parameter-efficient adaptation of multilingual models is a practical approach to low-resource English→Manipuri machine translation.

1 Introduction

Machine translation for low-resource languages continues to present significant challenges because of limited parallel corpora, restricted linguistic resources, and the computational cost associated with adapting large multilingual models. Manipuri, also known as Meiteilon, is a Tibeto-Burman language spoken primarily in the Indian state of Manipur and is one of the 22 scheduled languages of India. Despite its official status, Manipuri remains a low-resource language for Natural Language Processing (NLP), particularly for machine translation. The use of both the Bengali script

and Meitei Mayek further affects the distribution of available linguistic resources. In this work, we participate in the English→Manipuri (Bengali-script) direction of the WMT 2026 Low-Resource Indic Machine Translation shared task (Pakray et al., 2026). The source language is represented using the NLLB language identifier `eng_Latn`, while the target language uses `mni_Beng`. Multilingual machine translation models provide an important basis for translation involving low-resource languages. NLLB-200 (NLLB Team et al., 2022) supports approximately 200 languages, including Manipuri in the Bengali script. However, full fine-tuning of large multilingual models can require considerable memory and computational resources. To reduce these requirements, we adapt the NLLB-200-distilled-600M model using Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA introduces small trainable low-rank matrices into selected model modules while keeping the pretrained model parameters frozen. We investigate a progressive sequence of model configurations involving different training data pool sizes and training durations. The experiments range from a 5,000 pair data pool trained for two epochs to the complete official data pool trained for seven epochs. The final seven epoch model was used to generate translations for all 1,000 English source sentences in the official blind test set. The translations were submitted as the team’s primary system and evaluated by the shared task organizers against hidden references. The main contributions of this work are:

- (i) We develop a parameter-efficient English→Manipuri machine translation system by adapting NLLB-200-distilled-600M using LoRA.
- (ii) We study a progressive sequence of config-

urations involving different data-pool sizes and training durations.

- (iii) We evaluate the final system through the official WMT 2026 blind evaluation, where the primary submission ranked within the top three primary systems on five of the six reported metrics.
- (iv) We examine representative translation outputs to illustrate the behavior and remaining limitations of the final model.

2 Related Work

Neural machine translation (NMT) has achieved substantial improvements through multilingual pretrained models and transformer-based architectures. The Transformer architecture proposed by (Vaswani et al., 2017) established the foundation for modern sequence-to-sequence machine translation systems. Subsequently, multilingual models have demonstrated strong transfer capabilities, particularly for low-resource language pairs (Aharoni et al., 2019; Ranathunga et al., 2021). Among multilingual translation models, NLLB-200 (NLLB Team et al., 2022) significantly expanded language coverage by supporting approximately 200 languages, including Manipuri in the Bengali script. Its pretrained multilingual representations make it well suited for low-resource translation tasks through transfer learning. Parameter-efficient fine-tuning has recently emerged as an effective alternative to full model fine-tuning. LoRA (Hu et al., 2022) adapts large pretrained models by introducing trainable low-rank matrices while keeping most pretrained parameters fixed, substantially reducing computational and memory requirements. This approach has become widely adopted for adapting large multilingual models using limited computational resources. The WMT Low-Resource Indic Machine Translation shared tasks have provided a common benchmark for evaluating translation systems across multiple Indic languages. Previous editions reported strong performance from multilingual pretrained models combined with parameter-efficient adaptation techniques for low-resource translation settings (Pal et al., 2023; Kakum et al., 2023; Pakray et al., 2024,

Data Resource	Size
Official parallel corpus	23,687
Internal diagnostic sample	100
Official blind test source	1,000

Table 1: Data resources used for training, diagnostic comparison, and official evaluation.

2025, 2026). Our work follows this research direction by adapting NLLB-200 using LoRA for the English→Manipuri (Bengali script) translation task while investigating progressive training configurations under the official shared-task constraints.

3 Methodology

3.1 Dataset Description

The experiments use the official English→Manipuri parallel corpus provided for the WMT 2026 Low Resource Indic Machine Translation shared task. The available corpus contains 23,687 parallel sentence pairs. No external parallel data, monolingual data, synthetic data, back translated data, or other data augmentation was used during training. Table 1 summarizes the main data resources used in the study. For each experimental configuration, the selected data pool was divided into training and validation partitions using a 90:10 split with random seed 42. For the 5,000 sentence data pool, this resulted in 4,500 training sentence pairs and 500 validation pairs. For the 10,000 sentence data pool, this resulted in 9,000 training pairs and 1,000 validation pairs. The same 90:10 splitting procedure was applied to the cleaned complete data pool. For internal model comparison, a fixed 100 sentence diagnostic sample was selected using `random_state=123`. The same sample was used across the experimental configurations. Because these 100 sentences were sampled from the original corpus and were not explicitly excluded before training set construction, the resulting BLEU and chrF values are treated only as internal diagnostic scores. They are not interpreted as held out test set performance.

3.2 Task Overview

The system addresses English→Manipuri machine translation in the Bengali script track.

English is represented using the NLLB language identifier `eng_Latn`, while Manipuri in the Bengali script is represented using `mni_Beng`. The official blind test set contains 1,000 English source sentences. Reference Manipuri translations for this test set were not released to participants. Therefore, local reference based evaluation on the official test set was not possible. The final model generated translations for all 1,000 English inputs. These translations were submitted as the team’s primary system for organizer side evaluation against hidden references.

3.3 System Pipeline

The complete system pipeline is illustrated in Figure 1.

The official parallel corpus is first minimally processed and tokenized using the tokenizer associated with NLLB-200. The pretrained multilingual model is adapted using LoRA modules inserted into selected attention projection layers. The adapted model is trained through the progressive experimental sequence and compared using BLEU and chrF on the fixed internal diagnostic sample. The selected seven-epoch model is retained as the final system.

3.4 Data Preprocessing

The study employed minimal preprocessing. Sentence pairs containing missing source or target values were removed using `dropna()`, and the remaining English and Manipuri values were converted to string format. No explicit duplicate removal, Unicode normalization, punctuation normalization, spelling correction, language filtering, synthetic data augmentation, or back translation was performed. The data were tokenized using the NLLB tokenizer with `eng_Latn` as the source language identifier and `mni_Beng` as the target language identifier. Source and target sequences were truncated and padded to a maximum sequence length of 128 tokens.

3.5 NLLB-200 Base Model

Our system uses the pretrained checkpoint: `facebook/nllb-200-distilled-600M`. NLLB-200 (NLLB Team et al., 2022) is a multilingual encoder decoder machine translation model designed to support translation across a large number of languages. The selected distilled

Parameter	Value
LoRA rank (r)	64
LoRA alpha (α)	128
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj

Table 2: LoRA configuration of the final system.

Hyperparameter	Value
Data pool	23,687
Train-validation split	90:10
Split seed	42
Epochs	7
Batch size	4
Learning rate	7×10^{-5}
Optimizer	AdamW
Precision	FP16
Maximum length	128

Table 3: Training configuration of the final system.

checkpoint contains approximately 600 million parameters. The model directly supports both `eng_Latn` and `mni_Beng`, allowing adaptation to the target translation direction without extending the tokenizer vocabulary.

3.6 LoRA Adaptation

We adapt the base model using Low Rank Adaptation (LoRA) (Hu et al., 2022) through the PEFT framework (Mangrulkar et al., 2022). During adaptation, the original pretrained model parameters remain frozen. Small trainable low rank adapters are introduced into selected attention projection modules. The final LoRA configuration is shown in Table 2. The resulting LoRA adapter is approximately 55 MB in size. This demonstrates the storage efficiency of task specific adapter based fine tuning compared with storing a separate fully fine tuned copy of the base model.

3.7 Training Configuration

The final model configuration used the complete available data pool and a 90:10 training validation split. The model was trained for seven epochs. The main training configuration is presented in Table 3.

Experiments were conducted in a Google

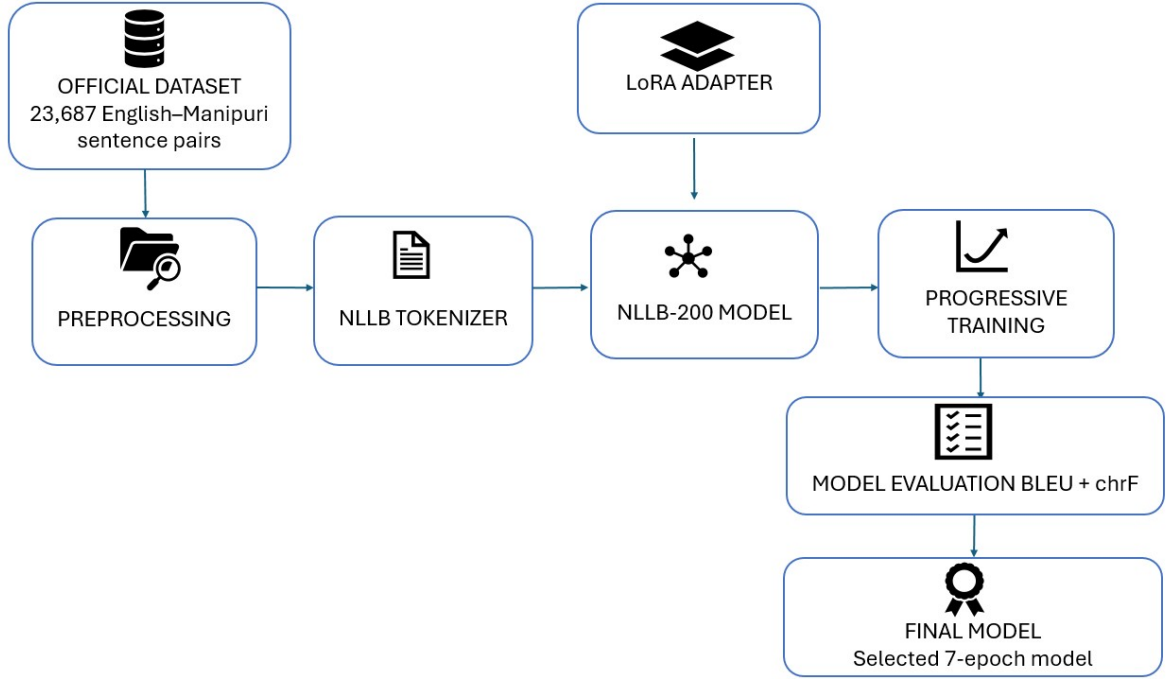


Figure 1: Overview of the proposed English→Manipuri machine translation system, including data preprocessing, NLLB tokenization, LoRA-based model adaptation, progressive training, internal evaluation, and final model selection.

Colab environment using NVIDIA T4 and NVIDIA L4 GPUs, depending on runtime availability. The final seven epoch training run required approximately 3 hours and 34 minutes.

3.8 Inference and Decoding

During inference, the trained LoRA adapter was loaded on top of the same facebook/nllb-200-distilled-600M checkpoint used during training. The source language was configured as `eng_Latn`, and generation was forced toward the target language `mni_Beng`. Translations were generated using beam search with a beam size of 5 and a maximum generation length of 128 tokens. The final system generated translations for all 1,000 source sentences in the official blind test set.

4 Experimental Runs and Primary Submission

We performed a progressive sequence of experiments to examine system behavior under increasing data pool sizes and longer training schedules. All configurations were compared using the same fixed 100 sentence internal diagnostic sample.

4.1 Experiment 1: 5K Data Pool

The first experiment used a data pool containing 5,000 English→Manipuri sentence pairs. After applying the 90:10 split, the model was trained on 4,500 sentence pairs and validated on 500 pairs. The model was trained for two epochs. This configuration obtained an internal diagnostic BLEU score of 12.40 and a chrF score of 48.87.

4.2 Experiment 2: 10K Data Pool

The second experiment increased the available data pool to 10,000 sentence pairs. The 90:10 split resulted in 9,000 training pairs and 1,000 validation pairs. The model was again trained for two epochs. The internal diagnostic scores increased to 13.36 BLEU and 50.56 chrF.

4.3 Experiment 3: Full Data, 3 Epochs

The third experiment used the complete available data pool with the 90:10 training validation splitting procedure. The model was trained for three epochs. This configuration obtained 14.94 BLEU and 52.82 chrF on the fixed internal diagnostic sample.

4.4 Experiment 4: Full Data, 5 Epochs

The fourth experiment retained the complete data pool and extended the training duration to five epochs. The model obtained an internal diagnostic BLEU score of 19.21 and a chrF score 56.51.

4.5 Final Primary System

The final configuration used the complete data pool and extended training to seven epochs. This configuration obtained 22.92 BLEU and 59.87 chrF on the fixed internal diagnostic sample. The resulting seven epoch model was selected as the final system. It was used to translate the complete official 1,000 sentence blind test set. The generated translations were submitted as: JH_NLP_APAShreya2_primary

5 Results

5.1 Evaluation Setup

Internal diagnostic translation quality was measured using the Hugging Face Evaluate library. Corpus level BLEU was computed using the sacrebleu evaluation module (Post, 2018). chrF was computed using the chrF evaluation module (Popović, 2015). The internal chrF configuration used: char_order=6, word_order=0, beta=2. Since word_order=0, this internal metric is reported as chrF and not chrF++. The official WMT evaluation separately reports chrF++ together with BLEU, METEOR, TER, BERTScore, and COMET. COMET is a learned metric designed to assess machine translation quality using neural representations (Rei et al., 2020).

5.2 Internal Diagnostic Results

Table 4 presents the results obtained across the progressive experimental sequence. The diagnostic scores increased across the successive experimental configurations. Under the same two epoch training duration, increasing the data pool from 5,000 to 10,000 sentence pairs increased BLEU from 12.40 to 13.36 and chrF from 48.87 to 50.56. For the complete data pool, extending training from three to seven epochs increased BLEU from 14.94 to 22.92, an absolute increase of 7.98 points. Over the same progression, chrF increased from 52.82 to 59.87, an absolute improvement of 7.05 points. The increase from five to seven epochs was

Run	Pool	Ep.	BLEU	chrF
Exp. 1	5,000	2	12.40	48.87
Exp. 2	10,000	2	13.36	50.56
Exp. 3	23,687	3	14.94	52.82
Exp. 4	23,687	5	19.21	56.51
Final	23,687	7	22.92	59.87

Table 4: Progressive results on the fixed 100 sentence internal diagnostic sample. Pool denotes the available data subset before the 90:10 training validation split.

also substantial. BLEU increased from 19.21 to 22.92, while chrF increased from 56.51 to 59.87. These results indicate that the diagnostic scores had not clearly plateaued within the tested training schedule. However, because the fixed diagnostic sample was not explicitly excluded from the training data, these values are used only for internal configuration comparison and are not treated as held out generalization estimates.

5.3 Official Blind-Test Results

The final model generated translations for all 1,000 English sentences in the official blind test set. Because reference translations were withheld from participants, local reference based evaluation was not possible. The generated translations were submitted to the shared-task organizers, who evaluated them against hidden references. The official scores of our primary system are shown in Table 5. The system ranked among the top three primary systems according to five of the six official evaluation metrics. The system achieved third place according to METEOR, third place according to TER, third place according to chrF++, third place according to BERTScore, and tied second place according to COMET. The system ranked tenth according to BLEU.

6 Discussion

The experimental sequence provides two different forms of evidence: internal diagnostic comparison across model configurations and external evaluation through the official blind shared task test set. The fixed internal diagnostic sample showed increasing BLEU and chrF scores across successive configurations. The progression from smaller data pools to

Metric	Score	Rank
BLEU $\uparrow$	8.03	10th
METEOR $\uparrow$	20.15	3rd
TER $\downarrow$	82.56	3rd
chrF++ $\uparrow$	40.93	3rd
BERTScore $\uparrow$	84.86	3rd
COMET $\uparrow$	68.18	tied 2nd

Table 5: Official results of the JH_NLP_APAShreyas primary submission. Ranks are among the 14 primary systems.

the full available corpus, together with longer training durations, corresponded with higher diagnostic scores. The strongest directly comparable training duration trend was observed on the complete data pool. Increasing training from three to seven epochs corresponded with an increase from 14.94 to 22.92 BLEU and from 52.82 to 59.87 chrF. However, these diagnostic values must be interpreted carefully because the fixed sample was not explicitly excluded from the training corpus. The official blind evaluation provides the actual external evaluation of the submitted system. The primary submission obtained 8.03 BLEU, 20.15 METEOR, 82.56 TER, 40.93 chrF++, 84.86 BERTScore, and 68.18 COMET. The system’s relative ranking differed considerably across metrics. Although it ranked tenth according to BLEU, it ranked within the top three according to METEOR, TER, chrF++, BERTScore, and COMET. Different evaluation metrics measure different properties of translation quality. BLEU emphasizes surface n-gram overlap, while character based and learned metrics provide additional perspectives on translation similarity and semantic adequacy. This divergence may be partly attributable to the morphological richness of Manipuri. As a highly agglutinative Tibeto-Burman language, Manipuri combines multiple grammatical markers (e.g., tense, case, and aspect) within a single word form. A model output that is semantically correct but differs from the reference by a single affix would be penalized heavily under exact n-gram matching (BLEU), while character-level metrics such as chrF++ and embedding-based metrics such as BERTScore and COMET are comparatively more tolerant of such surface-level variation. Script-specific tokenization behavior and syn-

onym variation in the hidden references may also contribute to this gap, though isolating the precise cause would require further controlled analysis. The difference between the internal diagnostic BLEU score of 22.92 and the official BLEU score of 8.03 should not be interpreted as a direct performance drop under identical conditions. The two scores were obtained in different evaluation settings. The internal diagnostic sample originated from the available corpus and was not guaranteed to be independent of training, whereas the official test set was blind and evaluated against hidden references. Furthermore, the best primary-system BLEU score in the official English→Manipuri results was 11.95, indicating that the blind test set was challenging for the participating systems. Further controlled evaluation would be required to determine the precise causes of the internal and official score difference.

6.1 Qualitative Analysis

To complement the automatic evaluation results, we present representative translation examples produced by the final model (Table 6). The examples illustrate both successful translations and challenging cases, highlighting the model’s ability to preserve semantic information while revealing limitations such as lexical substitutions and repetition in complex sentences. The examples are drawn from the internal diagnostic sample and are presented only for qualitative analysis.

Limitations

The experiments are limited by the relatively small size of the available English→Manipuri parallel corpus. The study used only one base model family and conducted a progressive experimental sequence rather than a complete factorial ablation study. Consequently, the contribution of every individual training choice cannot be isolated independently. Model configurations were compared using a fixed 100 sentence internal diagnostic sample. Because this sample was drawn from the original corpus and was not explicitly excluded from training set construction, the diagnostic scores should not be interpreted as held out generalization results. The qualitative examples also originate from the internal diagnostic sample and are used only to illustrate model behavior. Fine

Result	Sample English→Manipuri Translation Outputs
Example 1	
Source	Weekly prizes worth a maximum of Rs. 1 lakh to 7000 consumers.
Reference	কঞ্জুমৰ ৭০০০ত লুপা লাখ ১না মক্‌সমিম ওইবা চয়লৈগী ওইবা প্ৰাইজশীং পীগনি ।
Model	কঞ্জুমৰ ৭০০০দা মক্‌সমিম ওইনা লুপা লাখ ১গী ওইবা চয়লৈগী ওইবা প্ৰাইজশীং ।
Observation	Minor lexical variation while preserving the overall meaning.
Example 2	
Source	Breastfeeding helps to protect babies from pneumonia and other illnesses.
Reference	খমোলাং পীথকপনা অঙাংশীংবু ন্যুমোনিয়া অমসুং অতৈ অনাবশীংদগী ঙাকথোকপদা মতং পাংই ।
Model	খমোলাং ইনবনা অঙাংশীংবু ন্যুমোনিয়া অমসুং অতৈ অনাবশীংদগী ঙাকথোকপদা মতং পাংই ।
Observation	Near-correct translation with only a small lexical substitution.
Example 3	
Source	To trigger processes which lead to holistic development of the identified gram panchayats.
Reference	মশক খংদোকলবা গ্ৰাম পঞ্চায়টশীং অসগী মপুংফাৰা দভিলেপমন্টে লইনগদবা প্ৰদোসেসেশীং মথৌ হাপ্পা ।
Model	মশক খংদোকলবা গ্ৰাম পঞ্চায়টশীংগী মপুংফাৰা দভিলেপমন্টে লইনবনা প্ৰদোসেসেশীং হৌদোকপা ।
Observation	The translation preserves the main meaning with minor lexical differences.

Table 6: Representative English→Manipuri translation examples produced by the proposed system. Highlighted words indicate lexical differences between the reference and model output.

grained lexical and semantic analysis requires validation by fluent Manipuri speakers. Finally, the study did not explore external monolingual data, synthetic data generation, back translation, transliteration, corpus filtering, larger NLLB variants, or extensive hyperparameter optimization.

7 Conclusion

This paper presented the JH_NLP_APAShrea2 system for the English→Manipuri Bengali script direction of the WMT 2026 Low-Resource Indic Machine Translation shared task. The proposed system adapts the multilingual NLLB-200-distilled-600M model using LoRA. The system was trained using only the official English→Manipuri parallel corpus, without external parallel data, monolingual data, synthetic data, or back translation. A progressive sequence of experiments examined different data pool sizes and training durations. The final seven epoch model was used to translate all 1,000 English source sentences in the official blind test set. The primary submission achieved 8.03 BLEU, 20.15 METEOR, 82.56 TER, 40.93 chrF++, 84.86 BERTScore, and

68.18 COMET. The system ranked among the top three primary systems according to five of the six reported metrics. The results demonstrate that parameter-efficient adaptation of a moderately sized multilingual model can provide a practical approach to low-resource English→Manipuri machine translation. Future work will investigate larger NLLB variants, back translation using carefully validated monolingual Manipuri data, domain adaptation techniques, and translation involving the Meitei Mayek script, as well as a comparison of this parameter-efficient adaptation approach against off-the-shelf large language models such as Qwen3.x and closed API-based systems, to assess how compact fine-tuned multilingual models compare with general-purpose LLMs on this task. Future experiments should also use an independently held out development set, conduct more controlled hyperparameter studies, and include systematic human evaluation by fluent Manipuri speakers.

References

- Roei Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of NAACL-HLT*, pages 3874–3884.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. [Neural machine translation for limited resources English-Nyishi pair](#). *Sādhana*, 48(4):237.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. Software library.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Partha Pakray, Reddi Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. [Findings of WMT 2025 shared task on low-resource Indic languages translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553, Suzhou, China. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 2026. Findings of WMT 2026 shared task on low-resource indic languages translation. In *Proceedings of the Eleventh Workshop on Machine Translation (WMT) under EMNLP 2026*, Budapest, Hungary. Association for Computational Linguistics.
- Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.
- Maja Popović. 2015. chrF: Character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.
- Surangika Ranathunga, En-Shiun Annie Lee, et al. 2021. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 54(6):1–36.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.