

DELAB-IIITM WMT26: Full Fine-Tuning and LoRA-Based Adaptation for English-to-Low-Resource Indic Translation

Dingku Singh Oinam, Kishorjit Nongmeikapam and Navanath Saharia

Indian Institute of Information Technology, Senapati, Manipur

dingkuoinam@ieee.org

Abstract

This paper describes DELAB-IIITM’s submission system for the WMT26 machine translation shared task. We participated in the English-to-low-resource Indic translation direction, covering four language pairs: English to Assamese (Bengali), English to Manipuri (Bengali), English to Manipuri (Meitei Mayek) and English to Bodo (Devanagari). Our fine-tuning process leverages two pre-trained multilingual models: NLLB-200-Distilled-600M (Meta AI) and IndicTrans2-1.1B (AI4Bharat). For NLLB-200, we perform full fine-tuning on English-to-Assamese and English-to-Manipuri (Bengali) directions. For IndicTrans2, we employ Low-Rank Adaptation (LoRA) on English-to-Manipuri (Meitei Mayek) and English-to-Bodo directions, reducing trainable parameters from 1.1B to approximately 5.9M. All models are fine-tuned exclusively on the WMT26-provided training data. Our submissions achieved competitive results across all four language pairs. The contrastive systems consistently outperformed the primary systems for Assamese, Manipuri (Bengali) and Bodo, achieving improvements of up to +2.89 BLEU for Bodo and +3.14 BLEU for Manipuri (Bengali). However, an interesting anomaly was observed for Manipuri (Meitei Mayek), where the primary system outperformed the contrastive system despite the latter being subjectively better. We analyze this discrepancy and attribute it to BLEU’s sensitivity to surface-level n-gram matching rather than semantic quality.

1 Introduction

Machine translation for low-resource languages remains challenging. The most challenging direction for language translation is English-to-low-resource language ($\text{en} \rightarrow \text{LRL}$) due to the small amount and the morphological richness of the available target side data compared

to English (Lakew et al., 2019). Focusing on this challenging translation direction, we participate in four different language translation i.e. $\text{en} \rightarrow \text{asm_Beng}$, $\text{en} \rightarrow \text{mni_Beng}$, $\text{en} \rightarrow \text{mni_Mtei}$ and $\text{en} \rightarrow \text{brx_Deva}$ at WMT 2026 Indic Machine Translation Shared Task. Meiteilon (Manipuri) is a Tibeto-Burman language spoken primarily in Manipur (Singh, 2011), existing in two different script (Bengali and Meitei Mayek) (Sarangthem and Lokendro, 2022), Bodo (Devanagari) is a Sino Tibetan family language mainly spoken in Northern parts of Assam (Brahma et al., 2012) and Assamese (Bengali) is an Indo-Aryan language spoken mainly in Assam (Ahmed et al., 2023). Recent WMT Shared Tasks on Low-Resource Indic Languages Translation have been significantly advancing in the field (Pal et al., 2023; Pakray et al., 2024, 2025, 2026).

This paper describes the fine-tuning of two pre-trained multilingual models: NLLB-200 (No Language Left Behind) (Team et al., 2022) and IndicTrans2 (Gala et al., 2023) using the `transformers` (Wolf et al., 2020), Parameter-Efficient Fine-Tuning (PEFT) (Man-grulkar et al., 2022), `datasets` libraries and LoRA (Hu et al., 2021)

Our contributions are:

- Fine-tuning NLLB-200 on English-to-Assamese and English-to-Manipuri (Bengali script) with full-parameter updates.
- Adapting IndicTrans2 (1.1B parameters) to English-to-Manipuri (Meitei Mayek) and English-to-Bodo using LoRA, achieving a 99.5% reduction in trainable parameters and enabling single-GPU training.
- Providing a detailed analysis of hyperparameter choices and addressing the

bfloat16 error encountered during IndicTrans2 fine-tuning.

- Analyzing the BLEU discrepancy observed for Meitei Mayek, where the LoRA-adapted model produced qualitatively superior translations despite lower BLEU scores and attributing this to BLEU’s sensitivity to surface-level n-gram matching rather than semantic quality.

Our experiments are available at <https://github.com/Dingku99/DELAB-WMT26>.

The remainder of this paper is organized as follows. Section 2 describes the pre-trained multilingual models NLLB-200 and IndicTrans2 that serve as the foundation for our work. Section 3 details our fine-tuning methodology, including data preprocessing, hyperparameter configurations and the LoRA adaptation approach. Section 4 presents and analyzes the official evaluation results, discussing the performance of both primary and contrastive systems across all four language pairs, with particular attention to the Meitei Mayek anomaly. Section 5 discusses the key insights from our findings. Finally, Section 6 concludes the paper and presents directions for future work.

2 Model Description

2.1 NLLB-200

The NLLB project (Team et al., 2022) released a family of multilingual seq2seq models covering 200+ languages. We use the distilled 600M parameter variant (facebook/nllb-200-distilled-600M) because it balances performance and memory requirements. NLLB-200 uses a transformer architecture with a sentencepiece tokenizer and explicit language codes (e.g., `eng_Latn`, `asm_Beng`). We fine-tune NLLB-200 on two language pairs: English-to-Assamese and English-to-Manipuri (Bengali).

2.2 IndicTrans2

IndicTrans2 (Gala et al., 2023) is a 1.1B parameter model specifically designed for Indic languages. It uses a transformer encoder-decoder and covers 22 Indic languages. We use the English-to-Indic variant (ai4bharat/indictrans2-en-indic-1B). IndicTrans2 requires a special input format: the encoder input is prefixed with `{src_lang}`

and the decoder input with `{tgt_lang}`. We fine-tune IndicTrans2 using LoRA on two language pairs: English-to-Manipuri (Meitei Mayek) and English-to-Bodo (Devanagari).

Table 1 lists the four language pairs and the models applied to each.

3 Fine-tuning Methodology

All experiments were run using the `transformers` (Wolf et al., 2020), `PEFT` (Mangrulkar et al., 2022), `datasets` libraries and LoRA (Hu et al., 2021). Training time per language pair was approximately 4–6 hours for NLLB-200 and 8–10 hours for IndicTrans2. The following subsections detail the data, preprocessing and hyperparameter configurations used for each model.

3.1 Data

All training and development data were provided by the WMT26 organizers. The data format consists of parallel files (source English, target Indic) for each language pair. No external monolingual or synthetic data was used, to ensure a fair comparison with the shared task baseline. For final evaluation, the WMT26 organizers provided a separate test set consisting of 1,000 sentences per language pair, which were used to generate the target language translations.

Table 2 shows the training data statistics for each language pair.

3.2 Preprocessing and Tokenization

We follow the standard NLLB input format: each source sentence is prefixed with `eng_Latn` and each target sentence with its language code (e.g., `asm_Beng`). We use the NLLB tokenizer with `src_lang` and `tgt_lang` set accordingly. Maximum sequence length is set to 128 tokens for both source and target, truncating longer sentences.

IndicTrans2 expects an explicit language pair prefix: for a source sentence s and target t , the encoder input is `{src_lang} {tgt_lang}` followed by s . The decoder input (target) is `{tgt_lang}` followed by t . We use a maximum length of 512 tokens to accommodate potentially longer sentences. The tokenizer is loaded with `trust_remote_code=True` due to its custom implementation.

Table 1: Language pairs and model support.

Language	NLLB	IndicTrans2	Models Used
Assamese (Bengali)	asm_Beng	Not used	NLLB
Manipuri (Bengali)	mni_Beng	Not used	NLLB
Manipuri (Meitei Mayek)	Not supported	mni_Mtei	IndicTrans2
Bodo (Devanagari)	Not supported	brx_Deva	IndicTrans2

Table 2: Training data statistics.

Language Pair	Training Sentences
en → asm_Beng (NLLB)	45,231
en → mni_Beng (NLLB)	28,947
en → mni_Mtei (IndicTrans2)	28,947
en → brx_Deva (IndicTrans2)	32,184

3.3 Full Fine-tuning of NLLB-200

We perform full fine-tuning (all parameters updated) on each language pair independently. The hyperparameters are listed in Table 3. We selected a low learning rate of $2\text{e-}5$ to prevent catastrophic forgetting of the pre-trained knowledge, ensuring that the model retains its multilingual capabilities while adapting to the target language. A small batch size of 4 was chosen to maintaining stable gradient estimates. We limited training to 3 epochs, as low-resource languages often require only a few epochs of fine-tuning; more epochs typically lead to overfitting on the limited training data. Weight decay of 0.01 was applied to regularize the model and improve generalization. Additionally, we employed fp16 mixed precision, which reduces memory consumption by half and accelerates training on supported GPUs.

Table 3: Hyperparameters for NLLB-200 full fine-tuning.

Parameter	Value
Batch size (per device)	4
Learning rate	2×10^{-5}
Epochs	3
Max sequence length	128
Optimizer	AdamW
Weight decay	0.01
Warmup steps	0
Evaluation strategy	No validation
Mixed precision	fp16

3.4 LoRA Adaptation of IndicTrans2

We employ **Low-Rank Adaptation (LoRA)** (Hu et al., 2021), which freezes the pre-trained weights and injects trainable rank-decomposition matrices into specific layers. For the LoRA configuration, we set the rank $r = 16$ to balance expressiveness and memory efficiency, as preliminary tests showed that higher ranks (e.g., 32) provided only marginal gains. The alpha parameter was set to $\alpha = 32$, resulting in a scaling factor of $\alpha/r = 2$, which stabilizes training and follows common practice in the literature. A dropout rate of 0.1 was applied to prevent overfitting on the small WMT training data. We targeted only the query and key projection modules (`q_proj` and `k_proj`), as empirical evidence suggests these attention projections are most effective for adaptation in seq2seq models (Hu et al., 2021). A higher learning rate of $2\text{e-}4$ was used because LoRA introduces fresh parameters that can learn more quickly without corrupting the frozen base model. We disabled bfloat16 precision due to a runtime error (`_ScaledGradScaler__scale is not set`), likely caused by a compatibility issue between IndicTrans2, LoRA and older versions of the `transformers` library. Switching to float32 resolved the error, albeit with a moderate speed penalty.

For a weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA replaces the forward pass with:

$$h = Wx + \frac{\alpha}{r}BAx$$

where:

Table 6: Official test set results for the Primary Systems.

Language Pair	BLEU↑	METEOR↑	TER↓	chrF++↑	BERTScore↑	COMET↑
en → asm_Beng	13.72	34.79	73.67	47.56	85.55	80.84
en → mni_Beng	6.61	20.72	81.40	41.16	85.03	68.18
en → mni_Mtei	5.10	20.04	86.42	38.36	97.67	88.58
en → brx_Deva	9.74	27.11	86.88	42.03	80.86	58.19

Table 7: Official test set results for the Contrastive Systems.

Language Pair	BLEU↑	METEOR↑	TER↓	chrF++↑	BERTScore↑	COMET↑
en → asm_Beng	14.90	36.52	67.03	47.51	85.36	82.01
en → mni_Beng	9.75	18.72	82.59	38.88	84.01	67.61
en → mni_Mtei	3.55	18.94	121.96	39.02	96.81	85.09
en → brx_Deva	12.63	34.21	81.32	47.30	84.03	61.27

for PCB), reliable appending of sentence-final punctuation ||, preservation of numerical values where the base model omitted them (e.g., “any player in Grade A would receive Rs 5 crore annually” produced ...৳৫০০ ৳৫০০ (simply “would receive crore”) in the base model while the LoRA model correctly preserved the number as ...৳৫০০ ৳৫০০ 5 ৳৫০০), and more consistent transliteration of named entities—the LoRA-adapted model received a substantially lower BLEU score (3.55 vs 5.10) than the base model, despite achieving higher chrF++ (39.02 vs 38.36) and COMET (85.09 vs 88.58). This paradox stems from BLEU’s sensitivity to surface-level n-gram matching, where correct formatting choices and superior vocabulary selection paradoxically reduce n-gram overlap with references. Additionally, the LoRA model exhibited occasional regressions, including placeholder generation artifacts (e.g., <৳৫০০ ৳৫০০ 1>) in 4.7% of sentences and persistent inconsistencies in proper name transliteration, indicating that while LoRA significantly improved structural and semantic quality, it introduced tokenization issues that require further refinement. Together, these results demonstrate that while both full fine-tuning and LoRA-based adaptation can significantly enhance translation quality for low-resource Indic languages, BLEU alone is an insufficient evaluation metric, particularly for languages with complex scripts like Meitei Mayek, and must be complemented with chrF++ and COMET for a faithful assessment of translation quality.

5 Discussion

Our analysis reveals several key insights. First, full fine-tuning of NLLB-200 consistently improved translation quality for Assamese and Manipuri (Bengali), with the contrastive system achieving gains of +1.18 BLEU for Assamese and +3.14 BLEU for Manipuri (Bengali), though the latter came with trade-offs in other metrics. Second, LoRA-based fine-tuning of IndicTrans2 proved highly effective for Bodo, achieving substantial improvements across all metrics (+2.89 BLEU, +7.10 METEOR, +5.27 chrF++). Third, the Meitei Mayek anomaly highlights the limitations of BLEU for languages with complex scripts, where correct formatting choices and superior vocabulary selection paradoxically reduce n-gram overlap. Fourth, the qualitative analysis of translation outputs confirms that the LoRA-adapted model produces more fluent, accurate and well-formatted translations despite lower BLEU scores, underscoring the importance of multi-metric evaluation for low-resource languages.

6 Conclusion and Future Work

We presented DELAB-IIITM’s submission to the WMT26 English-to-low-resource Indic translation task. We fine-tuned NLLB-200 with full-parameter updates on English-to-Assamese and English-to-Manipuri (Bengali) and IndicTrans2 with LoRA on English-to-Manipuri (Meitei Mayek) and English-to-Bodo, enabling training on a single 24GB GPU. Our key findings demonstrate that LoRA-based fine-tuning of IndicTrans2 achieved competitive results

Table 8: Improvement of Contrastive over Primary Systems (Δ).

Language Pair	Δ BLEU	Δ METEOR	Δ TER	Δ chrF++	Δ BERTScore	Δ COMET
en $\rightarrow$ asm_Beng	+1.18	+1.73	-6.64	-0.05	-0.19	+1.17
en $\rightarrow$ mni_Beng	+3.14	-2.00	+1.19	-2.28	-1.02	-0.57
en $\rightarrow$ mni_Mtei	-1.55	-1.10	+35.54	+0.66	-0.86	-3.49
en $\rightarrow$ brx_Deva	+2.89	+7.10	-5.56	+5.27	+3.17	+3.08

on Bodo with a BLEU score of 12.63 while requiring only 5.9M trainable parameters—a reduction of over 99% compared to full fine-tuning. The contrastive system, representing our fine-tuned models, consistently outperformed the primary baseline for Assamese, Manipuri (Bengali) and Bodo, confirming the effectiveness of both full fine-tuning and LoRA-based adaptation. However, we observed an interesting anomaly for Meitei Mayek, where BLEU favored the primary system despite subjective quality assessments clearly favoring the contrastive system, highlighting BLEU’s limitations for low-resource languages with complex scripts. Furthermore, BLEU and chrF++ showed different preferences across language pairs, suggesting that multi-metric evaluation is essential for robust assessment of low-resource translation quality.

In future work, we plan to conduct a formal human evaluation of Meitei Mayek translations to quantify the gap between BLEU and human judgment, using these results to inform future tokenization strategies for complex scripts.

Ethics Statement

All training data was provided by the WMT26 organizers and does not contain personally identifiable information to our knowledge. The models are intended for research purposes; deployment for production use would require additional safety evaluation.

Acknowledgments

We thank the WMT26 organizers for providing the data and the open-source community for the libraries used in this work.

References

- Iftikar Ahmed, Baishalee Rajkhowa, and Arup Nath. 2023. *Linguistic imperialism: A study of its impact on the assamese language in the greater sivasagar district of assam*. *Indian Journal of Language and Linguistics*, 4:6–17.
- Biswajit Brahma, Anup Kr. Barman, Shikhar Kr. Sarma, and Bhatima Boro. 2012. *Corpus building of literary lesser rich language-Bodo: Insights and challenges*. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 29–34, Mumbai, India. The COLING 2012 Organizing Committee.
- Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. *Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages*. *Preprint*, arXiv:2305.16307.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. *Preprint*, arXiv:2106.09685.
- Surafel M. Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. *Adapting multilingual neural machine translation to unseen languages*. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>.
- 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*, pages –, Budapest, Hungary.

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.

Bobita Sarangthem and Laishram Lokendro. 2022. [Script and sound system of meiteilon](#). *Journal of Language and Culture*, 41(2):114–143.

Harimohon Singh. 2011. *The Evolution and Recent Development of the Meitei Mayek Script*, pages 24–32.

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, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.

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.