Transformers: Leveraging OpenNMT and Transfer Learning for Low-Resource Indian Language Translation

Bhagyashree Wagh Harish Bapat Neha Gupta Saurabh Salunkhe
Centre for Development of Advanced Computing (C-DAC)
{bhagyashreew, bharish, nehag, ssalunkhe}@cdac.in

Abstract

This paper describes our submission to the WMT 2025[1] (Pakray et al, 2025) Shared Task on Low-Resource Machine Translation for Indic languages. This task is an extension of the efforts which was originally initiated in WMT 2023[2] (Pal et al., 2023), and further continued to WMT 2024 [3] (Pakray et al, 2024), received significant participation from the global community. We address English ↔ {Assamese, Bodo, Manipuri} translation, leveraging Hindi and Bengali as high-resource bridge languages. Our approach employs Transformer-based Neural Machine Translation (NMT) models, initialized through multilingual pre-training on high-resource Indic languages, followed by fine-tuning on limited parallel data for the target low-resource languages. The pre-training stage provides a strong multilingual representation space, while fine-tuning enables adaptation to specific linguistic characteristics of the target languages. We also apply consistent preprocessing, including tokenization, true casing, and subword segmentation (Sennrich et al., 2016) with Byte-Pair Encoding (BPE), to handle the morphological complexity of Indic languages. Evaluation on the shared task test sets demonstrates that pre-training followed by fine-tuning yields notable improvements over models trained solely on the target language data.

1 Introduction

India is home to an extraordinary linguistic diversity, with 22 scheduled languages, languages written using many scripts and hundreds of regional and tribal languages spoken across its vast geography. While this richness offers immense cultural value, it presents significant challenges for computational linguistics and natural language processing (NLP). Many of these languages are classified as low resource, meaning that the quantity and quality of available digital text, speech, and annotated corpora are insufficient to support the development of robust NLP tools and machine translation systems.

The scarcity of datasets for low-resource Indian languages arises from multiple factors: historical underrepresentation in digital media, limited digitization of printed and oral resources, and the absence of standardized orthographies and lexical resources for certain languages. Moreover, much of the available data is random, noisy, or inconsistently encoded, making it unsuitable for large-scale model training without extensive preprocessing. This lack of data creates a bottleneck for building accurate and inclusive AI systems that can serve speakers of these languages.

Efforts to address these challenges are further complicated by the prevalence of code-mixing with English and other Indian languages in everyday communication. Consequently, the digital divide in language technology is widening, with high-resource languages benefiting from rapid advances in AI, while low-resource languages risk further marginalization.

[1] https://www2.statmt.org/wmt25/indic-mt-task.html

[2] https://www2.statmt.org/wmt23/indic-mt-task.html

[3] https://www2.statmt.org/wmt24/indic-mt-task.html

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 985–990
November 8–9, 2025.

In this paper, we examine the specific dataset challenges faced by low-resource Indian languages, explore their impact on model performance, and results for low resource languages.

2 Data Source

Low-resource languages such as Assamese, Bodo, and Manipuri face significant challenges in neural machine translation (NMT) due to limited parallel corpora. Recent advances in transfer learning have shown that pretrained models on large multilingual datasets can be effectively adapted to such languages, significantly improving translation quality. In this paper, we describe our submission to the WMT 2025 Shared Task, which combines the OpenNMT-py [4] framework, large-scale pretrained models, and fine-tuning on target language pairs. We used a combination of publicly available datasets, including:

2.1 High-resource parallel corpora

English-Hindi, English-Bengali and English-Manipuri from BPCC (Gala et al., 2023). This dataset is used for pre-training for both directions. We have further reduced the corpus size due to computational limitations. The corpus statistics are shown in Table 1. We further cleaned and normalized the Data for training. Before passing the data to the system, we applied the Byte Pair Encoding (BPE) (Sennrich et al., 2016) to the data.

Language Pair	Dataset Source	Size
En-Hindi	BPCC	3933323
En-Bangla	BPCC	33036843
En-Manipuri	BPCC	387084

Table 1: High-Resource Corpora

2.2 Low-resource parallel corpora

- Data released by WMT 2025 for Low-Resource Indic Language Translation for Training (Primary)
- For Bodo only, we have used another approach by using BPCC (Gala et al., 2023) dataset along with the released data and have performed transfer learning (Contrastive)

Language Pair	Size
En-Bodo	22000
En-Bangla	54000
En-Manipuri	387084+23687 (410771)

Table 2: Low-Resource Corpora

3 Methodology

3.1 Base Model

We adopted a Transformer-based encoder–decoder architecture as implemented in an open-source NMT toolkit. The base multilingual model was pre-trained on high resource parallel corpora in both the directions, providing strong shared representations for Indo-Aryan languages.

Language Pair	Model used as Parent model for Transfer Learning
En-Hindi	Yes (for Bodo)
En-Bangla	Yes (for Assamese)
En-Manipuri	Yes (for Manipuri)

Table 3: Base Model Details

3.2 Fine-tuning

The pre-trained model was fine-tuned on the low-resource language pairs.
Fine-tuning involved:

- Continuing training from the pre-trained checkpoint.
- Reducing the learning rate to prevent catastrophic forgetting.
- Applying early stopping based on validation BLEU & Perplexity.

This process enables the model to adapt quickly to the target languages with minimal overfitting
The detail of fine-tuning is given in Table 4.

Language Pair	Fine-Tuning Dataset	Size	Task
En-Bodo	WMT	22000	Primary
En-Bangla	WMT	54000	Primary
En-Manipuri	BPCC+WMT	387084 +23687 (410771)	Contrastive

Table 4: Fine-Tuning Details

[4] https://github.com/OpenNMT/OpenNMT-py

3.3 Training Details

- Batch size: 1024
- Validation Batch size: 512
- Optimizer: Adam
- Validation checkpoints and model averaging

Parameter	Value
Embedding Dimension	512
FFN Dimension	2048
Attention Heads	8
Encoder Layers	6
Decoder Layers	6

Table 5: Architectural Details

- GPUs used: V100

4 Experiments

4.1 Primary Submission

Our primary submission involved training a Transformer model from scratch using the OpenNMT Toolkit (Klein et al., 2017). Individual models were trained for translation, handling forward and backward language directions. The base model English-Bangla was used for Assamese transfer learning and the English-Manipuri base model was used for English- Manipuri finetuning using WMT datasets. We utilized SubWord tokenizer and Transformer architecture. The architectural details are shown in Table 5.

4.2 Contrastive Submission

The contrastive submission explored fine-tuning Base models in language-specific. The Base model English-Hindi was used for Bodo for transfer learning.

4.3 Other Experiments

4.3.1 Deep Decoder Approach

Additionally, we experimented with increasing decoder depth to 12 and 18 layers but observed that validation loss remained flat despite continued decreases in training loss. This is because each decoder layer has two attention sublayers, making it significantly more parameter-heavy than the encoder and prone to overfitting limited target-side data in low-resource settings. To address this, we plan to adopt an asymmetric depth configuration in future work, using a deeper encoder and a shallower decoder to retain strong source representation while limiting autoregressive overcapacity.

4.3.2 Experiments with LLMs

We also explored the use of the Llama model (Dubey et al., 2024) in conjunction with the LoRA (Low-Rank Adaptation) technique. Zero-Shot and Few-Shot Translation Evaluation We tested Zero Shot Translation capabilities of Llama 3-8B, Llama 3-8B-8192, mixtral8x7B-32768, Llama3-8B-instruct and Llama3.1- 8B-instruct. We also tested the few-shot translation capabilities of Llama3.1-8B-instruct with 3-shot, 5-shot, and 10-shot prompting. Supervised Fine-Tuning with LoRA We finetuned a 4-bit quantized (Liu et al., 2023) Llama3 model using the LoRA technique with Supervised Fine-Tuning (SFT), employing the Hugging face framework. We used a prompt based approach for translation, providing the model with a system prompt and a prompt template specifying the source and target languages. The following template was used for fine-tuning the Large Language Models (LLMs): System Prompt : You are an expert translator. Prompt Template : Translate the following English sentence to {target_language} in {target_script} Script:\n{input_sent}

Component	Setting	Rationale
Target Layers	q_proj, k_proj, v_proj, o_proj	Largest impact in Translation task
LoRA Rank (r)	16	Balance between expressiveness and efficiency
Scaling Factor (α)	32	Ensures effective contribution of LoRA updates
Dropout	0.05	Prevents overfitting given small corpus size
Precision	FP16	Improves training efficiency

Table 6: LLM Fine-Tuning details

5 Results

Training from scratch for low-resource languages like Bodo yields moderate performance but transfer learning from high-resource related languages provides significant gains. Using pretrained models trained on BPPCC as a base, we achieved BLEU improvements of over 12 points for Bodo and similar gains for Assamese and Manipuri. Future work will explore multilingual joint fine-tuning and domain adaptation. The evaluation results of three language pair directions NMT system on FLORES dev set is shown in Table

Language Pair	**Approach**	**BLEU Score**
en-brx	Contrastive	21.96
brx-en	Contrastive	33.93
brx-en	Primary	22.63
en-as	Contrastive	23.07
as-en	Contrastive	16.08
en-mni	Contrastive	11.92
mni-en	Contrastive	9.86
en-brx	Contrastive	21.96

Table 7: Evaluation Results

English	**Assamese**
Actor Shah Rukh Khan announces new film with director Rajkumar Hirani.	পৰিচালক ৰামাম হিৰিণৰ সৈতে নতুন ছবি ঘোষণা কৰিলে অভিনেতা শ্ৰুখ খান .
Priyanka Chopra shares adorable photo with daughter on Instagram.	কন্যাৰ সৈতে ইষ্টাগ্ৰাম আৰু প্ৰিয়াংকা চোপাৰ .
English	**Manipuri**
Actor Shah Rukh Khan announces new film with director Rajkumar Hirani.	পরিচালক রাজকুমার হিরানীগা লোয়ননা অনৌবা ফিল্ম ঘোষণা করলেন অভিনেতা শাহরুখ খান .
Priyanka Chopra shares adorable photo with daughter on Instagram.	ইন্টরগ্রাসতা প্রিয়ঙ্কা চোপ্রানা নুপীমচা অদুগী ফোটোগ্রাফ শেয়ার তৌরি
English	**Bodo**
Priyanka Chopra shares adorable photo with daughter on Instagram.	प्रियंका चोपड़ाया फिसाजोजों लोगोसे गोजोनथाव सावगारिखौ इनस्टाग्रामआव फोसावो .
Amitabh Bachchan tests positive for COVID-19, admitted to hospital.	अमिताभ बच्चना कभिड-19 नि थाखाय पजिटिभ आनजाद नायदोंमोन , जायखौ देहा फाहामसालियाव थिसननाय जादोंमोन .

Table 8: Results English-IL

Assamese	**English**
মাইক্ৰ'আৰএনএৰ নোবেল বিজয়ী আৱিষ্কাৰে কেনেকৈ ৰোগ নিৰ্ণয় আৰু চিকিৎসাৰ মুখখন সলনি কৰি আছে	How the Nobel Laureates of Microsoft ' s RNRs have changed the face of disease management and treatment .
ফুসফুসৰোগ বিশেষজ্ঞসকলে বিপদজনক কাৰকসমূহ শ্বেয়াৰ কৰে আৰু ইয়াক কেনেকৈ প্ৰতিৰোধ কৰিব পাৰি তাৰ পৰামৰ্শ দিয়ে	The rash experts shew dangerous chemicals and advise how to resist them
Manipuri	**English**
মাইক্রোআর.এন্.এ.গি নোবেল্ মাইপাকপা অসিনা মতৌ করম্না দাইগ্নোসিস অমসুং থেরাপিসিংগি মওং মতৌ হোংদোক্লিবনো।	how to change the form of diagnosis and therapy when the microRAN model is successful .
পলমোনোলোজিস্তসিংনা রিস্ক্ ফেক্তরসিং সেয়র তৌই অমসুং মখোয়বু করম্না ঙাকথোক্কদগে হায়বগি পাউতাক্ পিরি।	pulmonologystings share the risk factor and explain how to protect them
Bodo	**English**
केन्सारनि अनगायैबो, हादोर नाङैनो बिजिरसंगिरिफोरा अल्जाइमार आरो भाइरेल सन्देरनायखौ सिनायनो आरो फाहामनो थाखाय माइक्र'आर.एन.ए.जों खामानि मावगासिनो दं।	In addition to cancer , researchers nationwide are working with microRNA to identify and treat Alzheimer ' s and viral infections .
बोसोरनि गोजां बोथोरनि समाव बिलाइ गोथां मैगं-थाइगंफोरा मानो बांसिन निउट्रियेन्टफोर पेक खालामो?	Why do leafy greens pack more nutrients during winter ?

Table 9: Results IL-English

7. The output sample of the shared data (blind evaluation) is provided in table 8 & 9.

6 Conclusion

We described the Team submission to the WMT 2025 Shared Task on Low-Resource Indic Language Translation. By combining OpenNMT-py with transfer learning from BPCC (Gala et al., 2023), we achieved competitive results for English–Assamese, English–Bodo, and English–Manipuri, and vice-versa. Future work will explore back-translation, domain adaptation, and multilingual pre-training with additional Indic languages to further enhance low-resource translation performance

Acknowledgments

We acknowledge the organizers of the WMT 2025 Low-Resource Indic Language Translation Shared Task for providing valuable dataset and facilitating this research. We also thank the developers of OpenNMT, AI4BHRAT[5], Llama 3[6], FLORES[7] for making the models & data publicly available.

References

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.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. [Findings of wmt 2024 shared task on low-resource indic languages translation], In Proceedings of the Ninth Conference on Machine Translation, pp. 654-668. 2024, link: aclanthology.org/2024.wmt-1.54.pdf

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul 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. Transactions on Machine Learning Research.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. arXiv preprint arXiv:1508.07909. https://arxiv.org/abs/1508.07909

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2022. High-resource Language-specific Training for Multilingual Neural Machine Translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-2022)*, pages 4461–4467. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/619

Zoph, B., Yuret, D., May, J., and Knight, K. 2016. Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201.

Li, Zhaocong, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2022. ConsistTL: Modeling Consistency in Transfer Learning for Low-Resource Neural Machine Translation. arXiv preprint arXiv:2212.04262. https://arxiv.org/abs/2212.04262.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.

Scalvini, B., Debess, I. N., Simonsen, A., & Einarsson, H. (2025). *Rethinking Low-Resource MT: The Surprising Effectiveness of Fine-Tuned Multilingual Models in the LLM Age*. NoDaLiDa/Baltic HLT 2025.

Zhang, X., Rajabi, N., Duh, K., & Koehn, P. (2023). Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. WMT 2023.

Su, T., Peng, X., Thillainathan, S., Guzmán, D., Ranathunga, S., & Lee, E.-S. (2024). Unlocking Parameter-Efficient Fine-Tuning for Low-Resource Language Translation. Findings of NAACL 2024.

Stap, D., Hasler, E., Byrne, B., Monz, C., & Tran, K. (2024). The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. ACL Long Papers 2024.

Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Kumar Dash, Arnab

[5] https://ai4bharat.iitm.ac.in/
[6] https://www.llama.com/models/llama-3/
[7] https://huggingface.co/datasets/facebook/flores

Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das and Riyanka Manna. ***Findings of WMT 2025 shared task on Low-resource Indic Languages Translation***, In Proceedings of the Tenth Conference on Machine Translation, Suzhou, China, EMNLP 2025.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, Partha Pakray. **Neural machine translation for limited resources English-Nyishi pair**, Sādhanā 48, 237, Springer [2023]