

BRS-NK: Bidirectional Finetuning of NLLB-200 for Low-Resource Indic Languages: Assamese-English and Bodo-English Machine Translation

Bipesh Subedi* and Nischal Karki*

Department of Computer Science and Engineering

Kathmandu University, Dhulikhel, Nepal

{bipeshrajsubedi,nischal3158}@gmail.com*

Abstract

This paper describes our submission to the WMT 2026 Low-Resource Indic Language Translation Shared Task for the English-Assamese and English-Bodo pairs in both translation directions. We fine-tune the pretrained NLLB-200 distilled 600M model for primary as well as contrastive tasks. Since Bodo is not supported by NLLB-200, we use the Hindi (hin_Deva) language tag as a proxy for Bodo during fine-tuning and inference, based on an empirical evaluation of Devanagari-script language tags. The constrained systems are trained exclusively on the official task data, while the contrastive systems additionally incorporate publicly available parallel corpora, including AI4Bharat BPCC and Google Smol. A rigorous preprocessing pipeline, combined with the extra parallel data, yields consistent gains across all automatic metrics, with the largest improvements observed for the lower-resource English-Bodo pair.

1 Introduction

Low-resource machine translation remains challenging due to limited high-quality parallel data, particularly for several Indic languages. Pretrained multilingual models such as NLLB-200 can help by transferring knowledge across languages, but some low-resource languages are not directly supported. Additional parallel data and alternative adaptation strategies can therefore be useful for improving translation in such settings.

This paper describes our submission to the WMT 2026 Low-Resource Indic Machine Translation shared task (Pakray et al., 2026), focusing on English $\leftrightarrow$ Assamese and English $\leftrightarrow$ Bodo. Our systems use the pretrained NLLB-200 distilled 600M model with bidirectional fine-tuning. For Bodo, which is not supported by NLLB-200, we use the Hindi language tag (hin_Deva) as a proxy based

on an empirical evaluation of Devanagari-script language tags. We submit constrained (Primary) systems using only shared-task data and contrastive systems with additional publicly available parallel corpora. A common preprocessing pipeline is applied to all training data.

Results on the official WMT 2026 test sets show that the contrastive systems consistently outperform the constrained systems, with the largest gains for English $\leftrightarrow$ Bodo. The results also demonstrate the effectiveness of supervised fine-tuning and proxy language tags for adapting multilingual models to low-resource and unsupported languages.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the shared task. Section 4 describes data preparation and preprocessing. Section 5 presents the system architecture, and Section 6 describes training and inference. Section 7 reports the experimental results, followed by the conclusion in Section 8.

2 Related Works

A range of approaches have been explored for the WMT shared task and for low-resource Indic translation more broadly, most of them centered on adapting large pretrained multilingual models to the target languages. In previous editions of the shared task, most participating systems followed a similar recipe: fine-tuning a large multilingual pretrained model on the task data, usually with some mix of preprocessing, back-translation, and external corpora (Wei et al., 2024; Pal et al., 2023; Pakray et al., 2024, 2025).

Massively multilingual sequence-to-sequence models like mBART (Liu et al., 2020), M2M-100 (Fan et al., 2021), and NLLB-200 exploit cross-lingual transfer, which makes them a fairly natural starting point for low-resource translation work. For Indic languages specifically, AI4Bharat’s In-

** Equal contribution.

dicTrans2 release, along with large parallel and back-translation corpora covering all 22 scheduled languages, has become something of a go-to resource (Gala et al., 2023)), and further monolingual corpora and benchmarks have further expanded the data and evaluation resources for Indic languages (Doddapaneni et al., 2023). Instruction-tuned LLMs have also been used for this kind of task, such as fine-tuning LLaMA2-8B model for English $\leftrightarrow$ Kokborok and English $\leftrightarrow$ Bodo translation (Kumar et al., 2025).

When a target language is not in a pretrained model’s vocabulary at all, prior work tends to fall back on script similarity or transfer from a related language. HW-TSC, for instance, combined script-aware transfer with IndicTrans2 at WMT 2024 (Wei et al., 2024), and Kakum et al. (2023) built some of the earliest Neural Machine Translation (NMT) systems for English $\leftrightarrow$ Nyishi, an even lower-resource pair than what we work with here. Ghimire et al. (2026) faced a similar issue with Tamang, a Tibeto-Burman language written in Devanagari that does not have its own NLLB language tag. They found that using the Hindi tag as a proxy during NLLB-200 fine-tuning gave the best results. Based on this finding, we tested the same approach for Bodo. These studies motivate our approach of adapting pretrained multilingual models for low-resource Indic languages while addressing the challenges of limited data and unsupported languages.

3 Shared Task Setup

The WMT 2026 Low-Resource Indic Machine Translation shared task evaluates machine translation systems for under-resourced Indic languages. The task is divided into two categories according to the amount of available parallel training data. Category 1 includes moderately resourced language pairs: English $\leftrightarrow$ Assamese, English $\leftrightarrow$ Mizo, English $\leftrightarrow$ Khasi, English $\leftrightarrow$ Manipuri, English $\leftrightarrow$ Nyishi, and English $\leftrightarrow$ Meitei. Category 2 comprises very-low-resource language pairs: English $\leftrightarrow$ Bodo, English $\leftrightarrow$ Kokborok, English $\leftrightarrow$ Karbi, English $\leftrightarrow$ Nagamese, and English $\leftrightarrow$ Tagin.

We participate in one language pair from each category: English $\leftrightarrow$ Assamese (En $\leftrightarrow$ As) and English $\leftrightarrow$ Bodo (En $\leftrightarrow$ Brx). Both language pairs are evaluated in both translation directions, resulting in four translation tasks: Assamese $\Rightarrow$ English, English $\Rightarrow$ Assamese, Bodo $\Rightarrow$ English, and

English $\Rightarrow$ Bodo. We submit systems to both official tracks:

- **Constrained (Primary):** Training uses only the parallel corpora released for the shared task.
- **Contrastive:** Training additionally incorporates publicly available parallel corpora beyond the official shared task data.

4 Data Preparation and Preprocessing

We use the official parallel corpora released for the WMT 2026 Low-Resource Indic Machine Translation shared task (Pakray et al., 2026) as the training data for the primary systems. For the contrastive systems, we additionally use publicly available parallel corpora, including ai4Bharat BPCC-latest (Gala et al., 2023) and Google Smol (gatitos) (Jones et al., 2023). Since data quality is particularly important in low-resource machine translation, we apply the same preprocessing pipeline to the official and external datasets to improve data quality and consistency before fine-tuning. Table 1 summarizes the corpus sizes for each language pair and training track. It reports the number of sentence pairs before (*Org. Pairs*) and after (*Preproc. Pairs*) preprocessing.

4.1 English-Assamese (en-as)

For the English-Assamese language pair, we begin with the official parallel corpus released for the WMT 2026 shared task. We apply several preprocessing operations to improve sentence alignment and textual consistency as shown in Figure 1. First, we normalize text in both languages by removing redundant whitespace and correcting spacing around punctuation marks. We also normalize numeral representations by converting Roman numerals occurring within Assamese sentences into their corresponding Assamese numeral forms. To improve alignment quality, we remove sentence pairs in which a hyperlink appears on only one side, as this can indicate an extraction or alignment problem. We also remove placeholder artifacts such as PRODUCTNAME and PRODUCTVERSION, along with malformed or semantically inconsistent sentence pairs. Finally, we deduplicate the data at the sentence-pair level so that identical English-Assamese pairs occur only once.

Pair (Track)	Source	Org. Pairs	Preproc. Pairs
En $\leftrightarrow$ As (Primary)	Official WMT-2026 (Pakray et al., 2026)	54,000	52,914
En $\leftrightarrow$ As (Contrastive)	Official WMT-2026	54,000	53,992
	ai4Bharat BPCC-latest (Gala et al., 2023)	106,728	99,490
	Google Smol (gatitos) (Jones et al., 2023)	3,975	3,975
	<i>Total</i>	<i>164,703</i>	<i>157,457</i>
En $\leftrightarrow$ Brx (Primary)	Official WMT-2026 (Pakray et al., 2026)	15,215	12,770
En $\leftrightarrow$ Brx (Contrastive)	Official WMT-2026	15,215	12,758
	ai4Bharat BPCC-latest (Gala et al., 2023)	106,223	103,774
	Google Smol (gatitos, smallsent) (Caswell et al., 2025)	4,857	4,851
	<i>Total</i>	<i>126,295</i>	<i>121,383</i>

Table 1: Total training pairs by source for the primary and contrastive systems, before and after preprocessing.

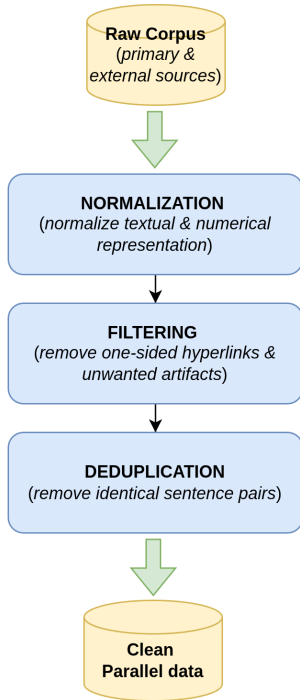


Figure 1: Data Preprocessing Pipeline.

4.2 English-Bodo (en-brx)

Compared with the Assamese corpus, the official English-Bodo dataset requires less cleaning. Nevertheless, we apply the same general preprocessing pipeline to ensure consistency across language pairs. The preprocessing pipeline removes sentence pairs containing asymmetric hyperlinks, placeholder artifacts, malformed translations, and duplicate sentence pairs. Since Bodo text is already represented consistently in the official corpus, no additional numeral normalization is applied.

4.3 Dataset Split and Evaluation Sets

For the primary systems, the preprocessed WMT 2026 parallel data are divided into training, valida-

tion, and internal test sets. The internal test sets contain 501 sentence pairs for English $\leftrightarrow$ Assamese and 504 sentence pairs for English $\leftrightarrow$ Bodo. For the contrastive systems, the preprocessed WMT 2026 data are combined with the additional external parallel corpora, and the resulting data are divided into training and validation sets. The WMT 2026 internal test sets are kept separate and used for internal evaluation. The same training and validation splits are used across the corresponding experiments to ensure consistent comparisons. In particular, we use the internal test sets to evaluate proxy language selection and compare zero-shot and fine-tuned models. The external contrastive data are not used for these comparisons.

For final evaluation, we use the official WMT 2026 test sets, which contain 1,000 source sentences for each translation direction. The target sentences are not provided with the test sets. We generate translations for the provided source sentences and submit the system outputs to the WMT 2026 shared task organizers for evaluation. The organizers evaluate the submitted translations using the official evaluation pipeline and provide the resulting scores. These test sets are kept separate from the training, validation, and internal evaluation data and are used only for the final evaluation of the submitted primary and contrastive systems. Table 2 summarizes the training and validation sizes used for the primary and contrastive experiments.

5 System Architecture

5.1 Overview

We fine-tune the pretrained NLLB-200 (Costa-Jussà et al., 2022) model for each language pair without modifying its architecture. Our approach focuses on data preparation, bidirectional fine-tuning,

Pair (Track)	Train	Validation
En $\leftrightarrow$ As (Primary)	50,413	2,001
En $\leftrightarrow$ As (Contrastive)	150,055	5,400
En $\leftrightarrow$ Brx (Primary)	11,762	504
En $\leftrightarrow$ Brx (Contrastive)	114,099	5169

Table 2: Training and validation sets used in our experiments for the primary and contrastive tracks.

and consistent training and inference procedures. The overall system consists of four stages: data collection, corpus preprocessing, bidirectional fine-tuning, and inference. The dataset sources, preprocessing steps, and data splits are described in Section 4. For the primary systems, only the official parallel corpora released for the WMT 2026 shared task are used for training. The contrastive systems additionally incorporate publicly available parallel datasets, as shown in Table 1. The same preprocessing pipeline is applied to all training data. Following preprocessing, each parallel sentence pair is converted into two training examples by reversing the translation direction. This allows a single model to learn both directions for each language pair. The resulting training data are used to fine-tune the pretrained NLLB-200 model with the same training configuration across Assamese $\leftrightarrow$ English and Bodo $\leftrightarrow$ English. During inference, translations are generated using beam search with language-specific source and target language identifiers.

5.2 Proxy Language Selection for Bodo

Although Bodo is primarily written in the Devanagari script, it is not among the 200 languages supported by NLLB-200. Consequently, NLLB-200 does not provide a dedicated language identifier for Bodo. To fine-tune the model without modifying its vocabulary or architecture, we use a proxy-language tag. We evaluated four NLLB-supported Devanagari-script languages as candidate proxies for Bodo: Hindi (hin_Deva), Nepali (npi_Deva), Maithili (mai_Deva), and Marathi (mar_Deva). For this comparison, we used the WMT 2026 training data and the internally held-out Bodo test set described in Section 4.3. The internal test set contains 504 sentence pairs. The same training data and fine-tuning configuration were used for all candidate proxy languages, with only the proxy language identifier varied. The results are reported in Table 3. Hindi achieved the best performance across all evaluation metrics in both translation directions. It obtained the high-

est SacreBLEU scores of 43.07 for Bodo $\Rightarrow$ English and 30.68 for English $\Rightarrow$ Bodo, while also achieving the best scores for ChrF, ChrF++, METEOR, and COMET. Based on these results, we use hin_Deva as the proxy language tag for all subsequent English–Bodo systems during both fine-tuning and inference. Using a proxy language allows us to use the pretrained NLLB-200 model for Bodo without modifying its vocabulary or architecture. Among the evaluated Devanagari-script candidates, the Hindi language tag provided the best performance for our Bodo experiments.

6 Training and Inference

6.1 Training Configuration

For every parallel sentence pair (x, y) , where x denotes the Indic sentence and y its English translation, we construct two supervised training examples: Indic $\Rightarrow$ English (x, y) and English $\Rightarrow$ Indic (y, x) . Thus, each bilingual sentence pair contributes two training instances, allowing a single model to learn both translation directions. All systems are fine-tuned using the Hugging Face Transformers framework (Wolf et al., 2020), with the pretrained NLLB-200 distilled 600M model as initialization. The fine-tuning hyperparameters used in our experiments are summarized in Table 4. The training and validation sets described in Section 4.3 are used for all fine-tuning experiments. For each experiment, the checkpoint with the lowest validation loss is selected for final inference. The same training configuration is used across language pairs and submission tracks, except for the specified differences in training epochs and warm-up steps for the contrastive systems. Training is performed using the AdamW optimizer (Loshchilov and Hutter, 2019) with mixed-precision (FP16) training to reduce GPU memory consumption and improve training efficiency. Model performance is monitored on the validation set through periodic evaluation and checkpointing. All experiments are conducted on a single NVIDIA RTX Ada Generation 6000 GPU with 48GB of memory.

6.2 Inference Configuration

During inference, translations are generated using beam search with a beam size of five. For each input sentence, the tokenizer is configured with the corresponding source language, while generation is constrained by forcing the beginning-of-sentence (BOS) token to match the desired target

Proxy	Bodo⇒English					English⇒Bodo				
	BLEU	ChrF	ChrF++	METEOR	COMET	BLEU	ChrF	ChrF++	METEOR	COMET
Maithili (mai_Deva)	41.33	64.53	62.78	0.69	0.80	22.30	56.54	51.74	0.44	0.61
Marathi (mar_Deva)	42.19	64.86	63.09	0.69	0.80	27.53	63.81	58.82	0.51	0.65
Nepali (npi_Deva)	41.73	64.56	62.83	0.69	0.80	26.40	62.44	57.47	0.50	0.64
Hindi (hin_Deva)	43.07	65.85	64.09	0.70	0.81	30.68	66.23	61.43	0.54	0.67

Table 3: Comparison of candidate NLLB proxy language tags for Bodo. All models were fine-tuned using identical hyperparameters, differing only in the proxy language identifier. Hindi (hin_Deva) consistently achieved the best performance across all evaluation metrics in both translation directions and was therefore selected for all subsequent experiments.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	5×10^{-5}
Epochs	5
Dropout	0.1
Max length	128
Gradient accumulation steps	2
Effective batch size	64
Warm-up steps	100
Weight decay	0.01
Precision	FP16

Table 4: Fine-tuning hyperparameters used for all experiments, except for contrastive experiments, where the number of epochs is reduced to 3 and warm-up steps is increased to 500.

language. This ensures that the multilingual model generates translations in the intended target language. Input sequences are truncated to a maximum length of 128 tokens, and generated translations are decoded after removing special language identifiers and tokenizer-specific control symbols. To improve inference efficiency, sentences are translated in batches using the PyTorch DataLoader API. The same decoding configuration is used for both the primary and contrastive systems across all translation directions, so that differences in translation quality are primarily due to the training data and fine-tuning process rather than decoding settings.

6.3 Zero-shot Baseline

To measure the effect of task-specific fine-tuning, we also evaluate the pretrained NLLB-200 distilled 600M checkpoint without fine-tuning on the WMT 2026 training data. The zero-shot baseline uses the same decoding configuration described in Section 6.2. Its performance is evaluated on the internally held-out test sets described in Section 4.3.

7 Experimental Results

7.1 Evaluation Protocol

Following the WMT 2026 IndicMT Shared Task protocol, we report the official evaluation scores using SacreBLEU (Post, 2018), ChrF++ (Popović, 2017), METEOR (Banerjee and Lavie, 2005), TER (Snober et al., 2006), COMET (Rei et al., 2020), and BERTScore (Zhang et al., 2020). The official results, released by the shared task organizers, are based on evaluations on the WMT 2026 test sets using the organizer’s evaluation pipeline.

7.2 Results and Discussion

Table 5 summarizes the official automatic evaluation results of our submitted systems on the WMT 2026 Low-Resource Indic Machine Translation Shared Task test sets. Results are reported for both the primary and contrastive tracks across the English-Assamese and English-Bodo language pairs. Across all translation directions, the contrastive systems consistently outperform the corresponding primary systems. These results indicate that additional parallel data provide useful supervision for low-resource translation, with consistent improvements across both language pairs and translation directions. Improvements are observed across lexical metrics such as SacreBLEU and METEOR, character-level similarity measured by ChrF++, and semantic evaluation metrics including BERTScore and COMET. In addition, the contrastive systems achieve lower TER scores, indicating that fewer edits are required to transform the generated translations into the reference translations.

For the English⇔Assamese language pair, the contrastive systems consistently outperform the primary systems in both translation directions. For English⇒Assamese, SacreBLEU improves by 2.57 points (14.96⇒17.53), accompanied by an increase

Trans. Dir.	Track	BLEU $\uparrow$	METEOR $\uparrow$	TER $\downarrow$	ChrF++ $\uparrow$	BERTScore $\uparrow$	COMET $\uparrow$
As $\Rightarrow$ En	Primary	21.27	54.30	69.40	51.33	94.18	84.70
	Contrastive	23.13	56.17	68.93	52.97	94.37	84.97
En $\Rightarrow$ As	Primary	14.96	36.66	66.75	47.81	85.42	80.75
	Contrastive	17.53	40.53	63.22	50.77	86.35	82.18
Brx $\Rightarrow$ En	Primary	19.39	52.43	67.71	48.12	92.83	74.86
	Contrastive	24.29	60.62	62.26	54.23	94.09	78.89
En $\Rightarrow$ Brx	Primary	12.00	28.74	77.71	43.67	82.90	59.94
	Contrastive	17.34	37.95	68.50	50.98	85.10	62.79

Table 5: Official automatic evaluation scores of our submitted systems on the WMT 2026 Low-Resource Indic Machine Translation Shared Task test sets. Results are reported by the shared task organizers using the official evaluation pipeline. Higher scores indicate better performance for BLEU, METEOR, ChrF++, BERTScore, and COMET, whereas lower TER indicates better translation quality.

Trans Dir.	Assamese $\leftrightarrow$ English					Trans. Dir.	Bodo $\leftrightarrow$ English				
	BLEU	ChrF	ChrF++	METEOR	COMET		BLEU	ChrF	ChrF++	METEOR	COMET
En $\Rightarrow$ As (zs)	9.04	35.51	32.15	0.24	0.74	En $\Rightarrow$ Brx (zs)	0.83	15.95	13.13	0.05	0.42
En $\Rightarrow$ As (ft)	14.59	42.28	38.74	0.32	0.78	En $\Rightarrow$ Brx (ft)	30.68	66.23	61.43	0.54	0.67
Δ En $\Rightarrow$ As	5.55	6.77	6.59	0.07	0.04	Δ En $\Rightarrow$ Brx	29.85	50.28	48.30	0.49	0.25
As $\Rightarrow$ En (zs)	23.96	45.34	43.85	0.47	0.77	Brx $\Rightarrow$ En (zs)	2.53	22.55	19.83	0.17	0.42
As $\Rightarrow$ En (ft)	31.92	50.78	49.53	0.53	0.79	Brx $\Rightarrow$ En (ft)	43.07	65.85	64.09	0.70	0.81
Δ As $\Rightarrow$ En	7.96	5.44	5.68	0.06	0.02	Δ Brx $\Rightarrow$ En	40.54	43.30	44.26	0.53	0.39

Table 6: Comparison of zero-shot and fine-tuned NLLB-200 performance on the internal test set for the English $\leftrightarrow$ Assamese and English $\leftrightarrow$ Bodo translation tasks. zs=zero-shot and ft=fine-tuned.

in COMET from 80.75 to 82.18 and a reduction in TER from 66.75 to 63.22. Assamese $\Rightarrow$ English also shows improvements across all reported metrics, with SacreBLEU increasing from 21.27 to 23.13 and COMET from 84.70 to 84.97. These results indicate that the additional parallel corpora provide useful supervision beyond the official shared-task data. The performance gains are more pronounced for the English $\leftrightarrow$ Bodo language pair. For English $\Rightarrow$ Bodo, SacreBLEU improves by 5.34 points (12.00 $\Rightarrow$ 17.34), while METEOR, ChrF++, BERTScore, and COMET improve by 9.21, 7.31, 2.20, and 2.85 points, respectively. TER also decreases from 77.71 to 68.50. For Bodo $\Rightarrow$ English, SacreBLEU increases from 19.39 to 24.29, while COMET improves from 74.86 to 78.89. The larger gains for Bodo suggest that additional parallel data are particularly beneficial for the more resource-constrained language pair. Since the primary and contrastive systems use the same model architecture and decoding strategy, the improvements are largely associated with the additional training data.

To assess the impact of supervised fine-tuning, Table 6 compares the pretrained NLLB-200 model in the zero-shot setting with the corresponding fine-tuned models. This comparison is performed on

the internally held-out test sets described in Section 4.3. Fine-tuning consistently improves translation quality across all translation directions and evaluation metrics. For English $\Rightarrow$ Assamese, SacreBLEU increases from 9.04 to 14.59 (+5.55), while Assamese $\Rightarrow$ English increases from 23.96 to 31.92 (+7.96). Similar improvements are observed across ChrF, ChrF++, METEOR, and COMET.

The improvements are substantially larger for English $\leftrightarrow$ Bodo. English $\Rightarrow$ Bodo improves from a zero-shot SacreBLEU score of 0.83 to 30.68 after fine-tuning, while Bodo $\Rightarrow$ English increases from 2.53 to 43.07. Comparable improvements are observed across the other evaluation metrics. These results demonstrate that task-specific supervised fine-tuning substantially improves translation quality, particularly for Bodo, where the pretrained model has limited zero-shot capability. Together with the proxy-language results in Section 5, these findings suggest that appropriate language-tag selection can facilitate adaptation of multilingual models to unsupported languages.

8 Conclusion

In this paper, we presented our submissions to the WMT 2026 Low-Resource Indic Machine Transla-

tion Shared Task for the English $\leftrightarrow$ Assamese and English $\leftrightarrow$ Bodo language pairs. Our systems are based on the NLLB-200 distilled 600M multilingual model and use bidirectional fine-tuning under both primary and contrastive settings. For Bodo, which is not directly supported by NLLB-200, we used the Hindi language tag (hin_Deva) as a proxy based on empirical evaluation. The official results show that incorporating additional publicly available parallel corpora consistently improves translation quality over the primary systems, with the largest gains observed for English $\leftrightarrow$ Bodo. The zero-shot comparison further demonstrates the importance of supervised fine-tuning for low-resource translation.

Our study has some limitations. Bodo is not directly supported by NLLB-200, so the use of a proxy language tag may not fully capture its linguistic characteristics. The experiments are also limited to two language pairs, and the individual contributions of the additional parallel corpora are not evaluated separately. Finally, we focus on full fine-tuning of NLLB-200 distilled 600M and do not explore parameter-efficient adaptation or larger multilingual models.

In future work, we plan to investigate parameter-efficient adaptation techniques, such as LoRA, to extend multilingual translation models to unsupported low-resource languages while reducing the computational cost of full fine-tuning.

Code and Repository

The source code for this work is publicly available at: [WMT-2026-Low-Resource-Indic-Language-Translation-Shared-Task](#).

References

- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, Solo Farabado Cissé, Edoardo Ferrante, Alessandro Guasoni, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, and 3 others. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). Preprint, arXiv:2502.12301.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Jay Gala, Pranjal A Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.
- Rupak Raj Ghimire, Bipesh Subedi, Balaram Prasain, Prakash Poudyal, Praveen Acharya, Nischal Karki, Rupak Tiwari, Rishikesh Kumar Sharma, Jenny Poudel, and Bal Krishna Bal. 2026. Nep-tam: A nepali-tamang parallel corpus and baseline machine translation experiments. *arXiv preprint arXiv:2603.14053*.
- Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. ["GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation"](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.
- Deepak Kumar, Kshetrimayum Boynao Singh, and Asif Ekbal. 2025. Tackling low-resource nmt with instruction-tuned llama: A study on kokborok and bodo. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1215–1221.
- 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.

- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations (ICLR)*.
- Partha Pakray, Reddi Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, et al. 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.
- 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 the WMT 2026 shared task on low-resource indic languages translation. In *Proceedings of the Eleventh Workshop on Machine Translation (WMT)*, Budapest, Hungary. Held in conjunction with EMNLP 2026.
- 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.
- Maja Popović. 2017. [chrF++: Words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. 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.
- 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.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.
- Bin Wei, Zheng Jiawei, Zongyao Li, Zhanglin Wu, Jiaxin Guo, Daimeng Wei, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, et al. 2024. Machine translation advancements of low-resource indian languages by transfer learning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 775–780.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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](#). *Journal of Machine Learning Research*, 21(60):1–6.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.