ANVITA : A Multi-pronged Approach For Enhancing Machine Translation Of Extremely Low-Resource Indian Languages

Sivabhavani J, Daneshwari Kankanwadi†, Abhinav Mishra, Biswajit Paul
{jsbhavani.cair, abhinavmishra.cair, biswajit.cair}@gov.in, †daneshwarikankanwadi1599@gmail.com
Centre for Artificial Intelligence and Robotics,
CV Raman Nagar, Bangalore, India

Abstract

India has a rich diverse linguistic landscape including 22 official languages and 122 major languages. Most of these 122 languages fall into low, extremely low resource categories and pose significant challenges in building robust machine translation system. This paper presents ANVITA Indic LR machine translation system submitted to WMT 2025 shared task on Low-Resource Indic Language Translation covering three extremely low-resource Indian languages Nyshi, Khasi, and Kokborok. A transfer learning based strategy is adopted and selected suitable public pretrained models (NLLB, ByT5), considering aspects such as language, script, tokenization and fine-tuned with the organizer provided dataset. Further, to tackle low-resource language menace better, the pretrained models are enriched with new vocabulary for improved representation of these three languages and selectively augmented data with related-language corpora, supplied by the organizer. The contrastive submissions however made use of supplementary corpora sourced from the web, generated synthetically, and drawn from proprietary data. On the WMT 2025 official test set, ANVITA achieved BLEU score of 2.41-11.59 with 2.2K to 60K corpora and 6.99-19.43 BLEU scores with augmented corpora. Overall ANVITA ranked first for {Nyishi, Kokborok} ↔ English and second for Khasi ↔ English across evaluation metrics including BLEU, METEOR, ROUGE-L, chrF and TER.

1 Introduction

India is home to 22 official languages, 30 languages with a million plus native speakers and 122 languages with more than 10,000 speakers [1] . Most of these 122 languages fall into low and extremely low resource categories, where availability of parallel corpora is very limited. Moreover, many of these languages do not only suffer from scarcity of parallel corpora but also from quality monolingual corpora and lack of language processing resources and tools, making the development of NLP solutions, including machine translation, particularly challenging.

WMT 2025 shared task on Low-Resource Indic Language Translation[2] presented a novel challenge of building robust machine translation system for low-resource Indic languages from diverse language families. The task focused on translation of seven North-Eastern languages to and from English and comprises of (i) Assamese (*an Indo-Aryan language spoken mainly in the northeastern Indian state of Assam*), (ii) Mizo (*a Sino-Tibetan language spoken primarily in the Mizoram state of India*), (iii) Khasi (*an Austroasiatic language spoken in Meghalaya, India*), (iv) Manipuri/Meiteilon (*a Sino-Tibetan language and the official language of Manipur, India*), (v) Nyishi (*a Sino-Tibetan language of Arunachal Pradesh, India*), (vi) Bodo (*a Sino-Tibetan language of Assam*) and (vii) Kokborok language (*a Sino-Tibetan language spoken primarily by the Tripuri people*).

With the dawn of Neural Machine Translation (NMT) (Vaswani et al., 2017) and availability of parallel corpora, translation systems have achieved significant performance for high and medium resource languages. However quality machine translation system for the low-resource or extremely low-resource languages still remains a major challenge as NMT architectures (supervised) based on encoder-decoder framework need large parallel corpora; the more quality parallel data is available, the better NMT system can learn accurate and fluent translations. A large number of Indian languages are individually resource poor. Development of machine translation systems for low resource Indian languages poses sig-

[1] https://censusindia.gov.in/census.website/data/census-tables

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

Proceedings of the Tenth Conference on Machine Translation, pages 1240–1247

nificant challenges due to the scarcity of parallel corpora and limited linguistic resources including tools. Open source IndicTrans2 (Gala et al., 2023) is a pretrained multilingual machine translation model, specifically designed for the translation of 22 officially recognized Indian languages and does not support many low resource Indian languages such as Khasi, Nyishi and Kokborok. These low-resource Indian languages remained under-represented in other popular pretrained language models such as ByT5 (Xue et al., 2022) and Indic BERT (Doddapaneni et al., 2023) and similarly in popular pretrained machine translation models such as NLLB-200 (NLLB Team et al., 2022). However some of the recent development in techniques such as sub-word tokenization (Kudo and Richardson, 2018), use of monolingual data, data-augmentation techniques such as back translation (Kudugunta et al., 2019), transfer learning, PEFT (Hu et al., 2022) over suitably selected multilingual PLM (pre-trained language models) (Xue et al., 2022) and pre-trained multilingual translation models (NLLB Team et al., 2022; Gala et al., 2023) did open up gates for improving translation systems for low-resource language pairs.

This paper presents ANVITA Indic LR machine translation system, submitted to WMT 2025 shared task on Low-Resource Indic Language by ANVITA team. Our team focused on three languages Khasi (kha), Nyishi(njz), and Kokborok (trp) and participated in six translation directions translating both to and from English.

ANVITA's strategy involved several key steps:

- Transfer Learning: Carefully selected existing public pre-trained models (NLLB, ByT5) based on the relevance to the task's language, script, and tokenization.

- Fine-Tuning: Chosen pre-trained models are fine-tuned using dataset provided by the organizers to develop primary systems with the addition of new vocabulary pertaining to these three languages for better representation.

- Data Augmentation and Contrastive systems development: Selectively incorporated related-language corpora provided by the organizer for primary submission. The contrastive submissions include supplementary corpora sourced from the web, generated synthetically (with and without utilizing monolingual corpora), and also drawn from proprietary data.

As part of synthetic parallel corpora generation back translation is carried out using openly available third party translation tool. ANVITA also employed pre-processing with set of language agnostic heuristics and selective post edits using LLM. On the WMT 2025 official test set, ANVITA achieved BLEU score of 2.41-11.59 with 2.2K to 60K corpora and 6.99-19.43 BLEU scores with the augmented corpora. Overall ANVITA ranked first for {Nyishi, Kokborok} ↔ English and second for Khasi ↔ English on the Official test set across all the evaluation metrics used by the organizer.

The rest of the paper is organized as given below. WMT 2025 task set up is described in 2. Section 3 presents Related work. Brief introduction to datasets is given in Section 4. ANVITA Indic LR System is described in Section 5. Section 5.5 describes the Experimental setup. Section 6 reports Evaluation results. Section 7 concludes the paper along with future directions.

2 Task Setup

WMT 2025 shared task on Low-Resource Indic Language comprised of translation of seven diverse, low resource Indian languages with the objective of developing robust MT systems that produce high-quality translations despite the constraints of data availability. The languages are divided into two categories. Category-1 comprised of five languages and 10 translation directions Assamese↔English, Mizo↔English, Khasi↔English, Manipuri↔English and Nyishi↔English with moderate training data and category-2 comprised of two languages, 4 directions Bodo↔English and Kokborok↔English with very limited training data. For each language pair three submissions were allowed. Primary systems with the constraint of using only the official data with additional monolingual resources and public pretrained models and two optional unrestricted Contrastive systems which may use external or additional parallel corpora beyond the organizer provided corpora.

ANVITA team submissions included three languages (Khasi, Nyishi, Kokborok), six translation directions (Khasi↔English, Nyishi↔English,

Kokborok↔English) and 17 systems comprising six Primary and eleven Contrastive systems.

3 Related Work

Building quality machine translation systems for low resource languages remains a critical research area. NLLB (No Language Left Behind) (NLLB Team et al., 2022) is one such research initiative to address limitations of MT for low resource languages which built large-scale multilingual model for 200 languages (including few low-resource Indian languages) on curated multilingual corpora, using transformer-based architecture (Vaswani et al., 2017). NLLB showed strong zero-shot and few-shot translation capabilities. As, fine-tuning such large models remains computationally expensive, our approach involved use of Low-Rank Adaptation (LoRA) (Hu et al., 2022), a method for parameter-efficient fine-tuning of large models to significantly reduce the number of update parameters. NLLB uses sub-word tokenizers such as sentencepiece (Kudo and Richardson, 2018) or BPE. To have a token free approach towards wider inclusion, ByT5 (Xue et al., 2022) introduced a byte-level variant of the T5 model (Raffel et al., 2020), which operates at the byte level rather than the token or sub-word level. Our work builds on these foundations. We leverage the methodologies and insights gained from WMT 2024's Indic MT shared task (Pakray et al., 2024) efforts but extend them to the more challenging low resource languages like Kokborok and Khasi.

4 Dataset

The training dataset provided by the WMT 2025 organizer varies from 60,000 to as low as 2,269 sentences as summarized in Table-1. ANVITA Primary submissions used only parallel corpora provided by the WMT 2025 organizer. Contrastive systems however used additional parallel data curated as part of this work and is summarized in Table-2. For evaluation, the official test set used is summarized in Table-3

5 ANVITA Indic LR Machine Translation System

This section describes the design of ANVITA systems for three low resource languages Nyishi, Khasi and Kokborok.

5.1 Data Preprocessing

As part of data preprocessing, selected noise filtering as described in (Vegi et al., 2021, 2022) are applied on all the datasets as mentioned in Table 1 and 2 with the objective of reducing corpora noise and improve translation quality. Additionally, non-linguistic artifacts such as timestamps, nan/null etc. present in any language are also removed from the bi-text corpora. Finally, text is processed using moses decoder [3] for punctuation normalization of English Text, and unicode normalization on both Indic and English text. Statistics of corpora before and after preprocessing is captured in Table-4.

5.2 Data for Contrastive Systems

Our supplementary data collection strategy for Contrastive systems include harvesting of monolingual and parallel corpora from web. Further to create synthetic parallel corpora for augmentation, the harvested monolingual resources are back translated (Kudugunta et al., 2019) using openly available third party translation tool, where only sentences with up to 20 words length are considered. Additional data is curated for Khasi and Kokborok languages, as detailed in the Table-2.

5.3 Nyishi ↔ English Translation Systems

For Nyishi ↔ English primary systems, ByT5 pretrained language model (Xue et al., 2022) is fine-tuned with the organizer data. ByT5 does not natively support Nyishi language. Unlike typical T5 variants that use subword tokenization like SentencePiece or BPE, ByT5 operates directly at the byte level. For Contrastive systems, data synthesis technique is used to generate additional parallel corpora from the organizer provided data using a custom method, as described in the subsequent subsection.

5.3.1 Byte-level Tokenization

Byte level (Wei et al., 2021) language and script agnostic tokenizer is used for both Nyishi and English text inline with the ByT5 language model. This approach does not require vocabulary building. So no new vocabulary is added. ByT5 is explored for better adaptation and potential noise robustness.

[3] https://github.com/hplt-project/sacremoses

Language Pairs	Language Family	Language Script	Parallel Corpus (P)	Number of words	Number of unique words	Average sentence length (num_words)
Nyishi (njz) - English (en)	Sino-Tibetan	Latin	60,000	(njz) 324,105 (en) 338,278	(njz) 39,074 (en) 13,647	(njz) 5.4 (en) 5.6
Khasi (kha) - English (en)	Austroasiatic	Latin	26,000	(kha) 966,353 (en) 798,291	(kha) 8,123 (en) 12,778	(kha) 37.16 (en) 30.7
Bodo (brx) - English (en)	Sino-Tibetan	Devanagari	15,215	(bodo) 20,4947 (en) 227,408	(bodo) 32,039 (en) 30,546	(bodo) 13.47 (en) 14.94
Kokborok (trp) - English (en)	Sino-Tibetan	Latin	2,269	(trp) 51,261 (en) 55,498	(trp) 6,619 (en) 6,175	(trp) 22.59 (en) 24.45

Table 1: Statistics WMT 2025 parallel corpora provided by organizer for Nyishi, Khasi, Bodo, Kokborok and used for Primary submission

Language Pairs	Number of sentences	Description	System: Contrastive 1	System: Contrastive 2
Nyishi (njz)-English (en)	30000	1. Curated parallel sentences by employing custom data synthesis technique from Nyishi-English parallel corpus provided by the WMT 2025 organizer with pairwise cosine similarity ≥ 0.8	Yes	No
	10000	2. Curated parallel sentences by employing custom data synthesis technique from Nyishi-English parallel corpus provided by the WMT 2025 organizer with pairwise cosine similarity ≥ 0.9	No	Yes
Khasi (kha)-English (en)	10,000	1. English sentences harvested from children stories and back translated to Khasi.	Yes	Yes
	5,50,000	2. Khasi sentences harvested from news portals and translated to English using back translation technique	Yes	No
	50,000	3. English sentences (taken from Nyishi-English parallel corpus provided by the WMT 2025 organizer) and back translated to Khasi	Yes	Yes
	7000	4. Sentences drawn from proprietary parallel corpora	Yes	Yes
Kokborok (trp)-English (en)	22793	1. Parallel sentences harvested from web	No	Yes
	50000	2. English sentences (taken from Nyishi-English parallel corpus provided by the WMT 2025 organizer) back translated to Kokborok	Yes	No

Table 2: Statistics of supplementary parallel corpora used for contrastive systems

5.3.2 Data Augmentation through Synthesis

In the organizer provided Nyishi-English parallel corpus, average length of sentences is 5.4 and 5.6 for Nyishi and English text respectively. So to create synthetic data involving long sentences from the existing data, a custom data synthesis technique is implemented as described below. The goal is to generate longer and coherent sentence pairs by concatenating compatible bilingual text segments, thereby enriching the training dataset.

Given a bilingual corpus of sentence pairs (S_i, T_i), where S_i is a source sentence and T_i is the corresponding target translation, we combine two such pairs (S_i, T_i) and (S_j, T_j) to form:

$$S_{ij} = S_i + joiner + S_j,$$

$$T_{ij} = T_i + joiner + T_j,$$

only if for S_i and S_j sentence-level cosine similarity is $\geq$ Threshold

- For Contrastive-1 system, *threshold* is chosen to be ≥ 0.8 , which resulted in 30000 parallel sentences. These synthetically created parallel sentences are augmented with the given training set.

- For Contrastive-2 system, *threshold* is chosen to be ≥ 0.9, which resulted into 10000 parallel sentences. These are augmented with the given training set.

Here *joiner* is usually sentence end marker followed by one white space.

The performance results of contrastive-1, contrastive 2 systems are presented in Table-7. Primary systems BLEU scores are better than the contrastive systems in case Nyishi-English language pair, indicating little effectiveness of the data synthesis method on the official test set.

Translation direction	Number of sentences in test file	Number of words	Number of unique words	Average sentence length (num_words)
English (en)→Nyishi (njz)	1,000 (en)	11,355	4,244	11.35
Nyishi (njz)→English (en)	1,000 (njz)	13,323	3,347	13.32
English (en)→Khasi (kha)	1,000 (en)	11,355	4,244	11.35
Khasi (kha)→English (en)	1,000 (kha)	22,399	2,186	22.39
English (en)→Kokborok (trp)	1,000 (en)	11,355	4,244	11.35
Kokborok (trp)→English (en)	1,000 (trp)	13,675	3,563	13.67

Table 3: Statistics of WMT 2025 Official Test set for Khasi, Kokborok, Nyishi

Language Pairs	Parallel data (Pb)- Before data processing	Parallel data (Pa)- After data processing
Nyishi (njz)-English (en)	60,000	51,000
Khasi (kha)-English (en)	26,000	25,995
Bodo (bodo) - English (en)	15,215	12,765
Kokborok (trp)-English (en)	2,269	2,266

Table 4: Statistics of preprocessed parallel corpora

Optimizer	AdamW
learning rate	1e-5
learning rate scheduler	linear with warmup
precision	fp16
patience	5
maximum number of epochs	20
metric_for_best_model	CHRF++

Table 5: Training parameters

peft type	LORA
rank	64
lora alpha	128
lora dropout	0.1
target modules	all

Table 6: LoRA configuration

5.4 {Khasi, Kokborok} ↔ English Translation Systems

Our approach for {Khasi, Kokborok} ↔ English Primary systems involve finetuning of pretrained translation model NLLB-200-distilled-600M (NLLB Team et al., 2022) [4] with the organizer provided data. Language specific additional vocabulary for Khasi and Kokborok are also added to the NLLB vocabulary. For contrastive systems similar techniques are employed, but with additional parallel corpora as described in Table-2.

5.4.1 Data Augmentation Through Related Language Data

For augmenting Kokborok →English training data for primary submission, Bodo↔English corpora provided by the organizer is utilized, as Bodo is related to Kokborok. For better transfer, Bodo text is converted from Devanagri script to Latin script using a romanization tool[5].

5.4.2 Data Augmentation Through Harvesting from Web and Synthesis

For contrastive systems, quality monolingual corpora is compiled from children stories, news portals and also used Nyshi↔English data provided by the organizer. Further, these sentences are back translated using openly available 3rd party translation tool. Sentences are also drawn from proprietary corpora as described in Table-2. As low resource languages Khasi and Kokborok do not have sentence tokenizers, pySBD (Sadvilkar and Neumann, 2020) is used for the same.

5.4.3 Sub-word Tokenization

Sub-word tokenization is one of critical component of any NMT system. In our work, specifically for building Khasi and Kokborok MT systems, SentencePiece (Unigram language model) tokenizer is trained on data provided by the organizer. The SentencePiece (Unigram language model) is chosen as it is compatible with NLLB tokenizer.

[4]https://github.com/facebookresearch/fairseq/tree/nllb

[5]https://github.com/anoopkunchukuttan/indic_nlp_library

Lang. Direction	Primary/ Contrastive	BLEU	METEOR	ROUGE-L	chrF	TER	Cos Similarity	Rank
English (en)→Nyishi (njz)	primary	6.21	0.21	0.28	34.01	81.53	-	First
	contrastive	5.92	0.20	0.27	34.08	82.83	-	First
Nyishi (njz)→English (en)	primary	11.59	0.41	0.51	49.85	74.09	0.79	First
	contrastive-1	11.13	0.42	0.51	48.92	74.17	0.80	First
	contrastive-2	11.25	0.40	0.51	49.36	73.79	0.78	-
English (en)→Khasi (kha)	primary	7.34	0.25	0.34	28.34	75.77	-	Second
	contrastive-1	18.83	0.45	0.54	45.48	55.75	-	-
	contrastive-2	19.43	0.46	0.55	45.93	54.41	-	Third
Khasi (kha)→English (en)	primary	1.99	0.11	0.14	20.88	223.26	0.30	Second
	contrastive-1	7.44	0.38	0.42	41.85	102.87	0.74	Second
	contrastive-2	4.39	0.22	0.28	30.65	123.25	0.55	-
English (en)→Kokborok (trp)	primary	1.76	0.11	0.17	18.58	104.04	-	First
	contrastive-1	6.99	0.30	0.37	38.08	76.26	-	First
	contrastive-2	0.55	0.04	0.05	13.38	335.55	-	-
Kokborok (trp)→English (en)	primary	2.41	0.11	0.18	23.55	129.15	0.36	First
	contrastive-1	2.99	0.16	0.22	25.52	117.73	0.49	First
	contrastive-2	0.79	0.05	0.08	16.46	170.6	0.20	-

Table 7: Performance of ANVITA on the WMT 2025 Official Test set

5.4.4 Fusion of Language Vocabulary with Pre-trained Model

To have better representation of language and reduce OOV words, NLLB vocabulary is updated with the sub-word vocabulary of Khasi and Kokborok languages. Initial vocabulary of NLLB tokenizer is 2,56,204, which is augmented with additional 831 Khasi vocabulary and 235 kokborok vocabulary and this took the final tally of NLLB vocabulary to 2,57,272. As NLLB does not support Khasi and Kokborok, hence special language tokens are also added to NLLB. This vocabulary inclusion enabled effective finetuning and better representation of words from low-resource languages.

5.4.5 {Khasi, Kokborok} ↔ English Model Training

The training steps followed for the four directions are as given below:

- Addition of Khasi and Kokborok language codes to NLLB vocabulary.
- Addition of Khasi and Kokborok sub-word vocabulary to NLLB vocabulary.
- For Kokborok primary systems, Bodo text is Romanized and augmented as presented in Table 1 .
- Primary models are trained on the data as presented in Table 1.
- Supplementary training data of contrastive systems include the data collected and back-translated as presented in Table 2.
- For Kokborok→English both contrastive models, Kokborok language token embedding is shared with the pre-trained NLLB Mizo (lus) language token embedding. Thus, the learned embedding is used for Kokborok token generation during inference time.
- Each language direction is separately fine-tuned on NLLB-200-distilled-600M (NLLB Team et al., 2022) model using Low-Rank Adaptation (LoRA) (Hu et al., 2022) method.
- For {Kokborok, Khasi} → English translation, English language token is set as target language and also it is forced as beginning of sentence. Similarly for other direction, corresponding language tokens are forced as beginning of sentence tokens. This is required since NLLB being a multilingual many to many MT model one needs to ensure tokens from the correct target language are generated.
- For Kokborok→English translation in Contrastive-2 submission, DeepSeek-R1 LLM (DeepSeek-AI et al., 2025) is used for post-editing of English translations.

5.5 Model Training and Experiment Details

All the experiments are conducted on NVIDIA DGX machine with 8xA100 80GB GPU cards.

ANVITA Indic LR used huggingface [6] toolkits for training. Training parameters and LORA configuration for all the experiments are shown in Table-5 and Table-6 respectively. For {Nyishi, Khasi} ↔ English, training batch-size is set to 32, whereas for Kokborok ↔ English training batch-size 16 is used. Gradient accumulation is used to avoid out-of-memory issues while training.

6 Evaluation and Result Analysis

Performance evaluation was carried out by the WMT 2025 organizer using BLEU, METEOR, ROUGE-L, chrF, TER and Cosine Similarity metrics on the Official test set (Table-3) for both Primary and Contrastive systems submitted. The results published by the organizers are shown in the Table-7.

Primary Systems: Nyishi→English with 60K parallel corpora attained BLEU score of 11.59, English→Khasi with 26K parallel corpora attained BLEU score of 7.34 and Kokborok→English with only 2.2K parallel corpora attained BLEU score of 2.41.

Contrastive Systems: With augmented corpora, overall English→Khasi achieved BLEU score of 19.43 and English→Kokborok 6.99.

Scores of our system for {Nyishi, Kokborok}→English directions are relatively better than that of English→{Nyishi, Kokborok} direction; English→ Khasi performance scores are better than Khasi→English direction.

In terms of ranks, ANVITA on the WMT 2025 official test set, secured First rank for {Nyishi, Kokborok} ↔ English and second for Khasi ↔ English across evaluation metrics including, BLEU, METEOR, ROUGE-L, chrF, TER and Cosine Similarity.

7 Conclusion

WMT 2025 shared task on Low-Resource Indic Language Translation posed significant challenging problem of building robust MT system for extremely low-resource Indian languages, many of which have little presence in digital domain. ANVITA team utilizing multi-pronged approach with transfer learning strategy achieved BLEU score of 2.41-11.59 for 2.2K to 60K corpora and secured top ranks. Overall with augmented data, the team achieved BLEU score of 6.99-19.43 which also fell short of robust MT system threshold indicating need for data augmentation in terms of both quality corpora and synthetic data and innovative techniques for data generation, suitable architectural enhancement and better learning objectives.

[6] https://huggingface.co/

Acknowledgments

The authors would like to thank Director, CAIR for his constant encouragement and supports. The authors would also like to thank Prasanna Kumar KR and Chitra Viswanathan for their guidance and enablement.

References

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian

Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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, Toronto, Canada. Association for Computational Linguistics.

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*.

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*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Partha Pakray, Santanu Pal, Advaitha 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. ANVITA-African: A multilingual neural machine translation system for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Chitra Viswanathan, and Prasanna Kumar K R. 2021. ANVITA machine translation system for WAT 2021 MultiIndicMT shared task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 244–249, Online. Association for Computational Linguistics.

Junqiu Wei, Qun Liu, Yinpeng Guo, and Xin Jiang. 2021. Training multilingual pre-trained language model with byte-level subwords.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.