# DoDS-IITPKD:Submissions to the WMT25 Low-Resource Indic Language Translation Task

**Ontiwell Khongthaw*** **G.L. John Salvin** **Shrikant Tryambak Budde**

**Abigairl Nyasha Chigwededza** **Dhruvadeep Malkar** **Swapnil Hingmire**

Mehta Family School of Data Science and Artificial Intelligence,
Department of Data Science, Indian Institute of Technology (IIT) Palakkad, Kerala, India
{142503002,142402010,142402015,142201026}@smail.iitpkd.ac.in,
okhongthaw@gmail.com, swapnilh@iitpkd.ac.in

## Abstract

Low-resource translation for Indic languages poses significant challenges due to limited parallel corpora and linguistic diversity. In this work, we describe our participation in the WMT25 shared task for four Indic languages-Khasi, Mizo, Assamese, which is categorized into Category 1 and Bodo in Category 2. For our PRIMARY submission, we fine-tuned the distilled NLLB-200(600M) model on bidirectional English↔Khasi and English↔Mizo data, and employed the IndicTrans2 model family for Assamese and Bodo translation. Our CONTRASTIVE submission augments training with external corpora from PMINDIA, Google SMOL and GATITOS to further enrich low-resource data coverage. Both systems leverage Low-Rank Adaptation (LoRA) within a parameter-efficient fine-tuning framework, enabling lightweight adapter training atop frozen pretrained weights. The translation pipeline was developed using the Hugging Face Transformers and PEFT libraries, augmented with bespoke preprocessing modules that append both language and domain identifiers to each instance. We evaluated our approach on parallel corpora spanning multiple domains: article based, newswire, scientific, and biblical texts as provided by the WMT25 dataset, under conditions of severe data scarcity. Fine-tuning lightweight LoRA adapters on targeted parallel corpora yields marked improvements in evaluation metrics, confirming their effectiveness for cross-domain adaptation in low-resource Indic languages.

## 1 Introduction

Low-resource language translation remains one of the most persistent challenges in machine translation (MT), particularly for linguistically diverse regions such as India. We observed that in WMT25 the provided corpora spanned biblical, scientific, news, and article-based domain, introducing significant domain shifts that demanded robust adaptation strategies (Pakray et al., 2025). To address these challenges, we developed two primary systems. The first leveraged IndicTrans2, a transformer-based multilingual model optimized for Indic languages, and the second utilized NLLB-200(600M), a distilled multilingual model trained on over 200 languages. Both systems were fine-tuned using Low-Rank Adaptation (LoRA), enabling efficient domain adaptation without retraining the full model. For our contrastive submission, we augmented the training data with external corpora from sources such as PMINDIA (Haddow and Kirefu, 2020), GATITOS (Jones et al., 2023), and Google SMOL (Caswell et al., 2025), allowing us to explore the impact of data diversity on translation quality. This paper presents our system architecture, training methodology, and evaluation results, with a particular focus on how domain-specific corpora and external augmentation influence performance across four low-resource Indic languages: Khasi, Mizo, Assamese, and Bodo. Our approach employs parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) on a pre-trained MT model, enabling a detailed empirical analysis of how large-scale architectures can be effectively adapted for low-resource languages under severe data constraints. The findings contribute to the growing body of research on scalable and adaptable MT systems for underrepresented languages.

## 2 Related Work

Translation quality in low-resource scenarios has been significantly advanced by large-scale multilingual models and lexical augmentation techniques. Fan et al. (2022) introduced No Language Left Behind (NLLB) which demonstrates effective multilingual MT at scale using a Sparsely Gated Mixture of Expert models trained with data that is mined specifically for underrepresented languages.

*Work done at IIT Palakkad.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1010
November 8–9, 2025. 

Their approach achieved substantial BLEU improvements and incorporated safety evaluations using FLORES-200 (Fan et al., 2022). Also, Jones et al. (2023) explored bilingual lexica as a lightweight data augmentation method, showing that collected lexical resources such as GATITOS can significantly enhance performance in unsupervised translation settings.

Toolkits like the HuggingFace Datasets library (Lhoest et al., 2021) also made efforts to support data development and reproducibility, which standardizes access to hundreds of multilingual corpora used in MT research.

For evaluation, several automatic metrics have been proposed to correlate better with human judgments. Lin (2004) developed ROUGE, widely used in summarization but also adopted in MT, which computes n-gram overlap and has influenced newer evaluation benchmarks. Banerjee and Lavie (2005) introduced METEOR, which matches unigrams using surface forms, stems and synonyms, incorporating both precision and recall as well as word order. Snover et al. (2006) proposed Translation Edit Rate (TER), also called Translation Error Rate, which measures the number of edits required to change a system output into one of the references. Popović (2015) proposed chrF, a character n-gram F-score metric that outperforms word-level metrics in many segment-level evaluations.

## 3 Dataset

For our primary submission, we utilized the Indic Machine Translation corpus from the WMT25 Shared Task. This benchmark comprises parallel data for four low-resource Indian languages, stratified into two categories based on training data volume. Category 1 encompasses language pairs with moderate-sized corpora, whereas Category 2 contains the severely data-starved corpora.

The language pairs are delineated as follows:

**Category 1**: en-as (English ↔ Assamese), en-lus (English ↔ Mizo), en-kha (English ↔ Khasi)

**Category 2**: en-bodo (English ↔ Bodo)

The parallel corpora supplied by the WMT25 IndicMT shared task[1] were employed for all model development. Each language pair's dataset was randomly divided into training (70 %), validation (20 %), and internal test (10 %) subsets, as detailed in Table 1. In addition, the task organizers released held-out monolingual test sets containing 1,000 sentences per translation direction for each language pair; these sets were used exclusively for final evaluation.

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

| Language | Total Sentences | Train (70%) | Valid (20%) | Test (10%) |
|---|---|---|---|---|
| Assamese | 54,000 | 37,800 | 10,800 | 5,400 |
| Khasi | 26,000 | 18,200 | 5,200 | 2,600 |
| Mizo | 50,000 | 35,000 | 10,000 | 5,000 |
| Bodo | 15,215 | 10,651 | 3,043 | 1,521 |

Table 1: Summary of Parallel Training Data from the WMT25 Indic MT Dataset.

### 3.1 Contrastive System Dataset

For a comparative analysis of data augmentation, we constructed a contrastive system by supplementing the WMT25 training dataset with additional publicly available parallel corpora. Our goal was to assess the resulting impact on translation performance across low-resource language pairs.

We incorporated data from four primary sources: the PMINDIA corpus (Haddow and Kirefu, 2020), high-quality parallel corpora for multiple Indian languages, sourced from government websites, official publications, and other public domain materials, covering legal, administrative, and general-purpose domains.; the GATITOS dataset (Jones et al., 2023), which provides lexically-augmented data for multilingual translation; the SMOL dataset (Caswell et al., 2025), containing professionally translated sentences for under-represented languages; and the Tatoeba corpus (Tiedemann, 2020), a large, community-sourced collection of multilingual sentence pairs.

The total volume of parallel data for each language after augmentation is detailed in Table 2. This table delineates the contribution of each external corpus alongside the original WMT data.

| Corpus | Assamese (asm) | Bodo (brx) | Khasi (kha) | Mizo (lus) |
|---|---|---|---|---|
| WMT | 54,000 | 15,216 | 26,000 | 50,000 |
| GATITOS | 3,975 | 3,994 | 4,000 | 3,998 |
| Smol Sent | 0 | 863 | 0 | 863 |
| PMINDIA | 9,732 | 0 | 0 | 0 |
| Tatoeba | 0 | 0 | 1,426 | 0 |
| **Total** | 67,707 | 20,073 | 31,426 | 54,861 |

Table 2: Parallel Corpus Statistics for the Contrastive System, detailing the original WMT25 data and supplementary corpora.

## 4 Methodology

Our methodology is focused on fine-tuning state-of-the-art, pre-trained multilingual translation models that excel in low-resource settings. We chose NLLB-200(600M) (Fan et al., 2022) and IndicTrans2 (Gala et al., 2023) as our core architectures. NLLB-200(600M) , developed under the No Language Left Behind initiative, delivers extensive typological coverage and consistently high translation quality across diverse languages (Fan et al., 2022). IndicTrans2, by contrast, incorporates script-aware tokenization and subword segmentation tailored specifically to Indian languages, yielding superior performance on Indic↔English pairs (Gala et al., 2023).

By fine-tuning these complementary models on the WMT25 IndicMT parallel corpora and on the augmented corpus for our contrastive system, we established a strong performance baseline and systematically quantified the gains afforded by data augmentation.

### 4.1 Preprocessing

We employed a three-step preprocessing pipeline to ensure data consistency and compatibility with our models:

1. **Text Normalization:** English segments were processed using the `MosesPunctNormalizer` (Koehn et al., 2007), while a custom function (`preproc()`) performed Unicode NFKC normalization and non-printable character removal for Khasi and Mizo.

2. **Language Tagging:** Each sentence was prepended with a language-specific tag (e.g., <eng_Latn>, <kha_Latn>) to guide the multilingual model during fine-tuning.

3. **Dataset Structuring:** The processed sentence pairs were structured into a Hugging Face `DatasetDict` (Lhoest et al., 2021), enabling efficient batching, shuffling, and training via the `Trainer` API (Wolf et al., 2020).

### 4.2 System Description

#### 4.2.1 Primary Submission

Our primary systems are based on fine-tuning two state-of-the-art multilingual models—NLLB-200(600M) and IndicTrans2—selected for their complementary strengths on low-resource and Indic-script translations.

**NLLB-200(600M) for Khasi and Mizo:** We adopted the `facebook/nllb-200-distilled-600M` checkpoint (Fan et al., 2022) for Khasi and Mizo tasks.

**Model & Tokenizer:** The standard `NLLBTokenizer` handles Mizo without modification; for Khasi we registered a new language token (<kha_Latn>) at token ID 256204 to correctly signal the source and target language.

**LoRA Fine-Tuning:** We applied Low-Rank Adaptation (LoRA) to all linear layers, updating only adapter weights. This approach enables efficient domain adaptation with fewer trainable parameters compared to full fine-tuning. Training ran for 30 epochs under Adafactor (learning rate $1\times10^{-5}$, batch size 32) with early stopping after 10 evaluations. Evaluation metrics were BLEU, METEOR, ROUGE-L, chrF and TER. Detailed LoRA hyperparameters appear in Table 4.

**IndicTrans2 for Bodo and Assamese:** For Bodo and Assamese, we used the `ai4bharat/indictrans2-indic-en-dist-200M` model (Gala et al., 2023), which employs an `IndicProcessor` to prepend language tokens such as <brx_Deva> and <asm_Beng>.

**LoRA Fine-Tuning:** We mirrored the NLLB-200(600M) setup (Adafactor, $1 \times 10^{-5}$ learning rate, 32-sentence batch, 30 epochs, early stopping) and applied identical LoRA settings (see Table 4). The resulting adapter checkpoints are saved as lightweight artifacts.

#### 4.2.2 Contrastive Submission

To quantify the effect of data augmentation, we retrained the same base models on extended parallel corpora. The tokenization and training pipeline remained identical, with two key LoRA adjustments to accommodate the increased data volume.

**Model Setup:** We reused NLLB-200(600M) for Khasi/Mizo (Fan et al., 2022) and IndicTrans2 for Bodo/Assamese (Gala et al., 2023). All supplementary bilingual data underwent the preprocessing and language-tagging workflow described in Section 3.

**LoRA Adaptation:** We increased the LoRA rank to 64 and $\alpha$ to 128 to provide greater adaptation capacity for the contrastive data, while retaining LoRA's parameter efficiency. Training was reduced to 15 epochs (Adafactor, $1 \times 10^{-5}$ learning rate,

| Direction | BLEU | | METEOR | | ROUGE-L | | chrF | | TER | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | C | P | C | P | C | P | C | P | C |
| as-en | 21.40 | 21.75 | 0.695 | 0.690 | 0.701 | 0.703 | 66.14 | 65.77 | 54.90 | 53.77 |
| en-as | 17.54 | 17.64 | 0.422 | 0.422 | 0.007 | 0.007 | 57.75 | 57.71 | 71.17 | 74.81 |
| kha-en | 4.31 | 5.52 | 0.239 | 0.289 | 0.293 | 0.349 | 31.33 | 34.85 | 131.86 | 113.30 |
| en-kha | 14.20 | 20.08 | 0.370 | 0.452 | 0.431 | 0.534 | 39.95 | 47.36 | 87.50 | 59.98 |
| lus-en | 10.38 | 11.81 | 0.537 | 0.544 | 0.576 | 0.581 | 55.09 | 55.17 | 86.84 | 74.39 |
| en-lus | 14.26 | 14.72 | 0.415 | 0.407 | 0.515 | 0.506 | 48.51 | 48.55 | 72.22 | 69.49 |
| bodo-en | 21.68 | 22.11 | 0.627 | 0.629 | 0.679 | 0.688 | 62.95 | 63.55 | 54.29 | 52.84 |
| en-bodo | 24.45 | 24.97 | 0.513 | 0.519 | 0.168 | 0.169 | 67.71 | 67.81 | 51.84 | 51.50 |

Table 3: Results for all language pairs: Primary Submission Results (P) vs Contrastive Submission Results (C).

batch size 32), as detailed in Table 4. Performance comparisons against the primary systems isolate gains attributable to data augmentation.

| Parameter | Primary Submission | Contrastive Submission |
|---|---|---|
| Optimizer | Adafactor | |
| Learning rate | $1 \times 10^{-5}$ | |
| Epochs | 30 | 15 |
| Precision | bf16 | |
| PEFT type | LoRA | |
| Rank ($r$) | 16 | 64 |
| Alpha ($\alpha$) | 32 | 128 |
| Dropout | 0.05 | |
| Target modules | all linear layers | |

Table 4: LoRA Configuration for Primary and Contrastive Submissions.

## 5 Results

We evaluate our system submissions on the WMT IndicMT shared task for four low-resource Indian languages: Assamese, Khasi, Mizo, and Bodo. Tables 3 presents the comprehensive results for our primary and contrastive submissions respectively across all bidirectional translation pairs. All systems are evaluated using standard automatic metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), chrF (Popović, 2015), and TER (Snover et al., 2006).

The results demonstrate that our contrastive submissions generally achieved better or comparable performance across most language pairs and metrics compared to the primary submissions.

## 6 Conclusion

In this paper, we described the DoDS-IITPKD submissions to the WMT25 Low-Resource Indic Language Translation Task. Our systems were designed for multiple Indic-English and English-Indic translation directions, focusing particularly on Category-I languages of NorthEast India. We explored a combination of pre-trained multilingual models (IndicTrans, NLLB-200(600M)),fine-tuning strategies and LoRA-based efficient adaptation. Future work will focus on more domain-robust adaptation and incorporating quality estimation for improved translation reliability.

## 7 Acknowledgments

We thank the organizers of the WMT25 Shared Task on Low-Resource Indic Language Translation for providing the datasets and evaluation framework. We also gratefully acknowledge the Department of Data Science, IIT Palakkad, for computing resources and infrastructure support.

## 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, Ann Arbor, Michigan. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. Smol: Professionally translated parallel data for 115 under-represented languages.

Angela Fan, Shruti Bhosale, Holger Schwenk, and et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jay Gala, Sahil Choudhary, Ajitesh Sharma, Vinay Nair, Anoop Kunchukuttan, Pratik Patel, Anirudh Srinivasan, and Anupam Singh. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume, Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova Villanova, Leandro von Werra, Victor Sanh, Lewis Debut, Julien Chaumond, Mariama Drame, Lewis Tunstall, Eduardo del Moral, Javier Soriano, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha 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 under Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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: Technical Papers*, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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.