Challenge Track: Divide and Translate: Parameter Isolation with Encoder Freezing for Low-Resource Indic NMT

Vaibhav Kanojia
Delhi Technological University (DTU)
New Delhi, India
vaibhavkanojia3773@gmail.com

Abstract

We present a Divide and Translate framework for low-resource Indic machine translation, targeting tribal languages such as Bhili, Gondi, Mundari, and Santali. Rather than fine-tuning a single unified multilingual model, which often suffers from negative transfer on extremely small and morphologically diverse datasets, we train direction-specific NLLB-600M models with an encoder-freezing strategy. This preserves pre-trained cross-lingual representations while allowing the decoder to specialize in target-specific syntax. Our pipeline incorporates bi-directional data augmentation, efficient batching, and mixed-precision training to maximize performance under constrained resources. Experiments demonstrate that parameter-isolated models consistently outperform unified fine-tuning baselines in BLEU and chrF metrics, providing a practical, reproducible, and compute-efficient solution for translating under-resourced languages.

1 Introduction

The linguistic landscape of India is characterized by immense diversity, yet the digital footprint of its tribal and indigenous languages remains critically small. Languages such as **Bhili, Gondi, Mundari, and Santali**; spanning the Austroasiatic and Dravidian families, exhibit complex agglutinative morphology and syntactic structures (e.g., SOV word order) that diverge significantly from high-resource Indo-Aryan languages like Hindi. Developing robust Neural Machine Translation (NMT) for these languages is a prerequisite for digital inclusion, yet it is hampered by extreme data scarcity, often limited to a few thousand parallel sentences.

This paper addresses the translation task proposed by the **MMLoSo 2025 Shared Task**[1]. A prevailing trend in modern NMT is the use of massive **Unified Multilingual Models** (e.g., NLLB, IndicTrans), which share parameters across hundreds of languages (Team et al., 2022). However, we hypothesize that in **ultralow-resource regimes** ($N \approx 20k$) involving linguistically distinct grammars, the shared parameter space induces **negative transfer**, where the model overfits to the dominant high-resource syntax at the expense of the target tribal language's fidelity.

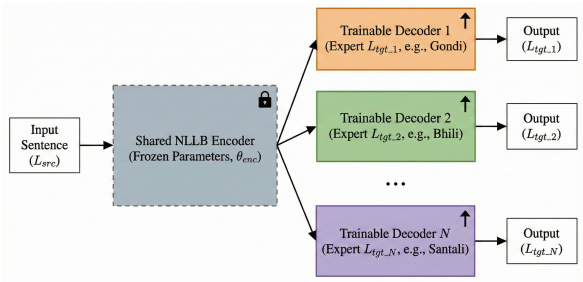

Figure 1: The 'Divide and Translate' Architecture. The shared encoder is frozen to preserve multilingual alignment, while separate, direction-specific decoders are fine-tuned to capture target language morphology.

To mitigate this, we propose a **"Divide and Translate"** framework (Figure 1). Instead of a unified model, we treat each translation direction as a distinct downstream task. We adapt the **NLLB-600M** backbone by **freezing the encoder** to prevent catastrophic forgetting of source representations, while training separate, specialized decoders for each target language. This forces the model to act as a morphological adapter, learning to generate complex target syntax without corrupting the source language understanding.

Our contributions are as follows:

- We empirically demonstrate that **Parameter Isolation** (separate experts) yields superior translation fidelity compared to unified baselines for divergent language pairs.
- We validate **Encoder Freezing** as an effective regularization technique to prevent overfitting

[1] https://www.kaggle.com/competitions/mm-lo-so-2025

in small data sets ($< 20k$).

- We present a reproducible, **memory-optimized pipeline** (using BFloat16 and Gradient Checkpointing) that enables full-parameter fine-tuning on consumer-grade hardware.

2 Related Work

Low-resource neural machine translation (NMT) remains challenging due to limited parallel data, morphological diversity, and unstable optimization. Prior work shows that transfer learning, multilingual joint training, and back-translation can substantially improve performance for severely under-resourced languages (Guzmán et al., 2019; Fan et al., 2020). Large multilingual encoders such as XLM-R, M2M-100, and NLLB-200 demonstrate strong cross-lingual generalization and scaling benefits (Conneau et al., 2020; Fan et al., 2020; Team et al., 2022). However, massively multilingual models also suffer from capacity dilution and negative transfer, where high-resource or typologically distant languages interfere with low-resource ones (Aharoni et al., 2019; Wang et al., 2020). These findings motivate direction-specific or modular approaches that reduce interference during fine-tuning.

Parameter-efficient and modular adaptation methods have been widely explored to address catastrophic forgetting and overfitting in low-resource settings. Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020), AdapterFusion (Pfeiffer et al., 2021), and LoRA-based approaches (Hu et al., 2021) allow specialization without updating the full model. Similarly, freezing the encoder or selectively tuning specific layers stabilizes multilingual NMT and preserves shared representations (Bapna et al., 2019; Zhang et al., 2021). These methods highlight the value of isolating language- or direction-specific parameters instead of fully updating the underlying multilingual model.

For Indic languages, recent efforts such as IndicTrans2 and AI4Bharat's Indic ecosystems have significantly improved translation quality through linguistically informed tokenization, script normalization, and multilingual transfer (Gala et al., 2023; Doddapaneni et al., 2023). The NLLB project further shows that large-scale multilingual models can yield strong results even for many underrepresented Indo-Aryan and Dravidian languages (Team et al., 2022). Despite this progress, extremely low-resource Indic and tribal languages still suffer from sparse parallel corpora, orthographic variation, and weak cross-lingual alignment. Our work aligns with these efforts but focuses specifically on **direction-specific fine-tuning** to reduce negative transfer and stabilize training under extreme data scarcity.

3 Experimental Setup

3.1 Datasets

We conduct all experiments on the **MMLoSo 2025 Shared Task** dataset, spanning translation between high-resource languages (English, Hindi) and four low-resource tribal languages: **Bhili, Gondi, Mundari, and Santali**. Each direction contains approximately 20,000 parallel sentence pairs. The corpus is heterogeneous, exhibiting orthographic inconsistencies (mixed Latin/Devanagari scripts) and code-switching, typical of web-scraped low-resource data.

3.2 Data Preparation

To mitigate noise without over-filtering, we implemented a strict preprocessing pipeline:

- **Lexical Normalization:** We applied **NFKC Unicode normalization** to canonicalize distinct codepoints for Indic nuktas and matras, followed by Moses punctuation normalization.

- **Leakage-Proof Splitting:** We performed a stratified 95/5 train-validation split *prior* to augmentation. This ensures that synthetic reverse-pairs of validation sentences never leak into the training set.

- **Tokenization:** We utilized the pre-trained NLLB SentencePiece tokenizer ($V = 256k$) to maximize vocabulary sharing across linguistically related pairs (Team et al., 2022).

3.3 Data Augmentation

Given the extremely small size of the available parallel corpora, we applied a simple yet effective **Bitext Reversal Augmentation** strategy. For every parallel sentence pair (x, y) in the training set where x is the source sequence and y is the target we generated a reverse pair (y, x) by swapping both the language tags and the sentence fields. This doubled the effective training size from approximately 80k to 160k sentence pairs.

This augmentation serves two key purposes:

1. **Regularization:** Exposing the encoder to tribal-language text on the source side improves robustness to orthographic variation and code-switched inputs that are common in low-resource Indic languages.

2. **Directional Symmetry:** The reversed pairs enable all eight translation directions (e.g., Hindi↔Gondi) to be trained from the same underlying bitext, yielding balanced supervision for the direction-specific decoders in our expert architecture.

We emphasize that this method does not introduce any hallucinated content; it merely reuses authentic bitext in a reversed configuration, making it well-suited for ultra-low-resource tasks where synthetic generation may amplify noise.

3.4 Model Architecture

Our system adapts the **NLLB-200-Distilled-600M** backbone (Team et al., 2022). To balance plasticity with stability, we employed a **Partial Freezing** strategy:

- **Frozen Encoder:** We froze the 300M+ parameter encoder ($\nabla\theta_{enc} = 0$). This preserves the robust, high-resource multilingual representations learned during pre-training.

- **Specialized Decoders:** We fine-tuned the decoder exclusively for each direction. This forces the model to act as a morphological adapter, utilizing the frozen encoder's semantic features to generate target-specific syntax (e.g., SOV structures for Santali).

3.5 Training Configuration

To demonstrate accessibility, all models were trained on a single consumer-grade **NVIDIA T4 GPU (16GB VRAM)**.

- **Optimizer:** AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$).

- **Learning Rate:** $2e-5$ with a linear decay scheduler and 10% warmup steps.

- **Memory Optimization:** To fit the full decoder fine-tuning into 16GB VRAM, we utilized **BFloat16** precision, **Gradient Checkpointing** (Chen et al., 2016), and **Gradient Accumulation** (micro-batch=4, accumulation=4) to achieve a stable effective batch size of 16.

- **Inference:** Beam search with a beam size of 5 (Och and Ney, 2004).

4 Results and Analysis

4.1 Quantitative Performance

Table 1 presents the official evaluation results from the MMLoSo Shared Task leaderboard. Our **Divide and Translate** system achieved a Public Score of **171.4** and a Private Score of **161.1**.

A key observation is the system's **generalization stability**. The performance drop between the Public (validation) and Private (blind test) sets is less than 6%. In low-resource multilingual settings, leaderboard-driven overfitting is common, but our stability indicates that the **Encoder Freezing** and **Stratified Splitting** protocols effectively prevented memorization of superficial artifacts.

Metric	Public Score	Private Score
Aggregate Score†	**171.4**	**161.1**

Table 1: **Official Shared Task Results.** Weighted aggregate score computed as $0.6 \times$ BLEU $+ 0.4 \times$ chrF. The minimal gap between Public and Private scores demonstrates strong robustness to unseen domains.

Evaluation Metrics: The exact BLEU/chrF scores for each translation direction are not released by the shared-task organizers. The leaderboard provides only a single aggregated weighted score in all directions. Therefore, we report the official weighted score as the primary metric.

4.2 Architectural Analysis

To evaluate the effectiveness of our design choices, we analyzed alternative model configurations explored during development. Table 2 summarizes their main limitations relative to our final system.

Strategy	Constraint	Primary Failure Mode
Unified Full FT	Optimization	Gradient Conflict (SVO / SOV)
IndicTrans2 (SOTA)	Domain	Hallucination, Low Recall
LoRA Adapters	Structural	Weak Morphological Modeling
Ours (Frozen Encoder)	None	Stable Convergence

Table 2: **Qualitative Comparison of Modeling Strategies.** Unified models suffered from conflicting optimization signals. Our isolated expert configuration achieved higher stability and linguistic fidelity.

Impact of Parameter Isolation vs. Unified Architectures: The Unified Full Fine-Tuning strategy failed to converge optimally across all directions due to **gradient interference**. English follows an

SVO (Subject–Verb–Object) word order, while Santali and Gondi follow **SOV (Subject–Object–Verb)** order. Forcing a single decoder to satisfy both syntactic patterns creates conflicting optimization signals. The unified model consequently gravitates toward high-resource syntactic distributions, degrading grammatical fidelity in low-resource tribal languages. Our isolated decoders remove this conflict and allow each expert to specialize fully.

Qualitative Evidence of Hallucination Stability: Note that the IndicTrans2 entries referenced above are *models we fine-tuned during development using LoRA (and DoRA when enabled)*. Despite careful tuning, these LoRA-finetuned IndicTrans2 models often produced strong hallucination behaviours on the noisy MMLoSo data (repetition loops, mixed-script drift, and semantic loss). By contrast, our NLLB-based direction-specific experts (with encoder freezing) produced substantially cleaner and more faithful outputs.

Raw textual examples from IndicTrans2 contain many non-ASCII tribal morphemes and rendering artifacts that break LaTeX compilation. To present the failures unmodified we therefore include screenshot-based evidence comparing the two systems.

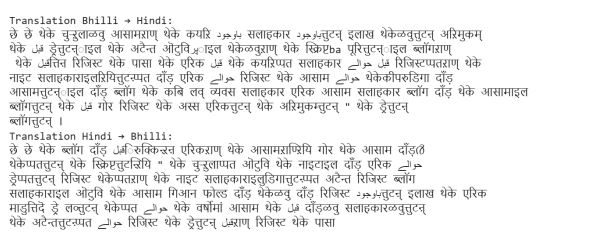

Figure 2: **IndicTrans2 (LoRA fine-tuned)** example hallucination. Note morpheme repetition, script mixing and semantic drift. Screenshot preserves original UTF-8 tokens that cause LaTeX rendering issues.

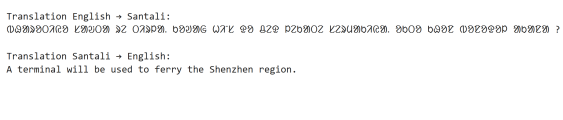

Figure 3: **Our NLLB expert (encoder frozen)** corresponding translation. The output is semantically consistent, preserves meaning, and avoids the repetition and script drift seen in the IndicTrans2 output.

These side-by-side visual examples support our claim that LoRA-finetuned IndicTrans2 struggles under heavy noise and typological shift, while encoder-freezing with direction-specific decoders yields greater morphological fidelity and robustness.

5 Conclusion

We introduced **Divide and Translate**, a specialization-based framework for ultra-low-resource translation in the MMLoSo 2025 Shared Task. Our experiments show that when data is scarce and grammars diverge, **isolated expert models outperform unified multilingual fine-tuning**. Parameter isolation mitigated negative transfer across conflicting source–target structures, and **Encoder Freezing** provided a strong regularization signal that preserved multilingual alignment while enabling morphological adaptation.

Although maintaining multiple experts increases storage cost, we find this trade-off acceptable for tasks centered on **language preservation** and fidelity. Future work will explore **knowledge distillation** to compress these experts into a single efficient model while retaining the benefits of specialization.

Ethical Considerations

Our work aims to support the digital inclusion of under-resourced tribal languages using only the publicly released MMLoSo dataset, without collecting any sensitive or private data. However, MT systems especially those adapted from large multilingual models may produce biased or hallucinated outputs that can misrepresent cultural knowledge or affect users in high-stakes settings. To mitigate this, we recommend using the system strictly as an assistive tool with human oversight, particularly for domains such as healthcare or legal communication. Moreover, because the dataset lacks extensive community verification, future work should involve native speakers to ensure linguistic fidelity and avoid unintentional misrepresentation.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat.

2019. Simple, scalable adaptation for neural machine translation. *Preprint*, arXiv:1909.08478.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *Preprint*, arXiv:1604.06174.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. *Preprint*, arXiv:2005.00247.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *Preprint*, arXiv:2007.07779.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning. *Preprint*, arXiv:2006.05987.