

MTG-LIRA: Scaling Neural Machine Translation for Low-Resource Indian Languages through Multilingual Learning and Synthetic Data

Vedarth Hede
C-DAC, Pune

Neha Gupta
C-DAC, Pune

Harish Bapat
C-DAC, Pune

Harsh Ekbote
C-DAC, Pune

Abstract

Low-resource Indian languages continue to face significant challenges in neural machine translation due to the scarcity of high-quality parallel corpora and the diversity of scripts and linguistic characteristics. This paper presents MTG-LIRA’s submission to the WMT 2026 Low-Resource Indic Language Translation Shared Task, focusing on scalable multilingual translation for several low-resource Indian language pairs. Our approach is built upon a unified multilingual Transformer architecture that enables parameter sharing and knowledge transfer across related languages. To further improve translation quality, we investigate multiple adaptation strategies, including multilingual transfer learning, synthetic data augmentation, task-specific fine-tuning, checkpoint averaging, and weighted model averaging, selecting the most effective approach for each language pair. We submitted systems for low resource language pairs like English↔Assamese, English→Mizo, English→Manipuri (Bengali), English↔Manipuri (Meitei Mayek), and English→Bodo with both primary and contrastive submissions where applicable. In addition, we conducted exploratory experiments on several other low-resource Indian languages to assess the generalizability of our multilingual framework. The submitted systems demonstrate that combining a shared multilingual foundation with language-specific adaptation strategies provides an effective and scalable solution for neural machine translation in low-resource Indian languages.

1 Introduction

Neural Machine Translation (NMT) has achieved remarkable success in recent years, primarily due to the Transformer architecture (Vaswani et al., 2017) and the emergence of large-scale multilingual models such as No Language Left Behind (NLLB) (Costa-jussà et al., 2022) and

IndicTrans2 (Gala et al., 2023). Despite these advances, many Indic languages continue to suffer from limited translation quality because of the scarcity of parallel corpora, diverse writing systems, and rich linguistic characteristics. These challenges make low-resource Indic machine translation an active area of research.

In this paper, we present MTG-LIRA’s submission to the WMT 2026 Low-Resource Indic Machine Translation Shared Task. Our approach is based on a unified multilingual Transformer architecture that enables knowledge sharing across related languages. Building upon this common framework, we explore several enhancement strategies, including multilingual transfer learning, synthetic data augmentation, task-specific fine-tuning, checkpoint averaging, and weighted model averaging, selecting the most suitable approach for each submitted language pair.

Our official submissions include systems for English→Assamese, English→Mizo, English→Manipuri (Bengali), English→Manipuri (Meitei Mayek), English→Bodo, Assamese→English, and Manipuri (Meitei Mayek)→English, with both primary and contrastive submissions where applicable. We also conducted exploratory experiments on additional low-resource Indic languages to evaluate the generalizability of our multilingual framework

2 Data

The performance of multilingual Neural Machine Translation (NMT) systems is largely influenced by the quality, diversity, and organization of the training data. To develop scalable translation systems for low-resource Indic languages, we

combined the official WMT 2026 shared task datasets with multilingual parallel corpora and synthetic parallel data generated for extremely low-resource languages. The multilingual training corpora was organized by linguistic similarity to facilitate effective cross-lingual knowledge transfer, while synthetic data was incorporated to alleviate severe data scarcity for selected language pairs. Before model training, all datasets underwent a unified preprocessing pipeline to remove noise, normalize language-specific text, and ensure consistent data quality across all multilingual corpora.

2.1 Training Data

The official datasets released as part of the WMT 2026 Low-Resource Indic Machine Translation Shared Task served as the primary benchmark for system development and evaluation. To improve multilingual representation learning, these datasets were complemented with the BPCC multilingual parallel corpus (Gala et al., 2023), where related Indic languages were jointly grouped into shared multilingual models. Separate multilingual models were constructed for the Indo-Aryan and Tibeto-Burman language (except Manipuri and Bodo) families to promote effective parameter sharing while preserving linguistic similarity among the participating languages. Table 1 and 2 summarizes the multilingual training corpora used in this work.

Language	Sentence Pairs
Assamese	54,000
Bodo	15,215
Mizo	50,000
Manipuri (Bengali)	23,867
Manipuri (Meitei Mayek)	17,000

Table 1: Official WMT 2026 Indic language Dataset Statistics

Language	Sentence Pairs
Assamese	1,609,595
Bodo	305,285
Manipuri (Bengali)	39,275
Manipuri(Meitei Mayek)	187,143

Table 2: BPCC Dataset Statistics

2.2 Synthetic Parallel Data Generation

To address the severe scarcity of parallel data for Mizo, we generated a large synthetic English–Mizo corpus using Meta's No Language Left Behind (NLLB) model. The official WMT 2026 training data contained approximately 50K English–Mizo sentence pairs, whereas approximately 5M English sentences were selected from the English side of the BPCC English–Hindi corpus for synthetic augmentation. These sentences were translated from English into Mizo using NLLB-200-instruct-3.3B Model, resulting in a synthetic corpus approximately 100 times larger than the available official Mizo parallel data. The generated sentence pairs were then processed using the same normalization, quality filtering, alignment validation, and duplicate-removal pipeline applied to the other training corpora. The resulting synthetic data was combined with the available real English–Mizo parallel data and incorporated into the multilingual Tibeto-Burman training corpus. This augmented model was used as the primary system, while a further fine-tuning stage on the official WMT English–Mizo data was used to construct the contrastive system.

2.3 Data Preprocessing

The parallel corpora for all language pairs collected from multiple sources exhibited considerable variations in formatting, NFKC encoding, sentence quality, and alignment. Therefore, all datasets, including official WMT 2026 released training data and synthetically

generated data, were processed through a unified preprocessing pipeline before being used for model training.

Language-specific Normalization: Each language was normalized using language-specific text normalization techniques to ensure orthographic consistency across the multilingual corpus. Unicode normalization, script-specific character normalization, punctuation standardization, and whitespace correction were applied where appropriate. These steps helped reduce inconsistencies arising from multiple writing conventions and improved vocabulary sharing during multilingual training.

Quality Filtering: The parallel corpora were filtered to remove noisy sentence pairs that could negatively impact model performance. Empty sentences, excessively long sentence pairs, misaligned translations, and malformed text were discarded. Sentence length constraints were also applied to eliminate extreme outliers while preserving the overall diversity of the training data.

Duplicate Removal and Corpus Validation: Duplicate sentence pairs were removed from both the official and synthetic datasets to minimize redundancy and reduce potential bias during training. Following deduplication, the corpora were validated to ensure correct source-target alignment, consistent sentence counts, and the absence of corrupted or incomplete records before proceeding to multilingual training.

3 Methodology

Our multilingual translation systems are based on a shared Transformer architecture trained using multilingual joint learning. While all submitted systems employ the same underlying model architecture and subword tokenization framework, different training strategies were adopted depending on the characteristics and resource availability of each language pair. This section describes the common model

architecture, the tokenization scheme, and the language-specific training strategies used for the official WMT submissions.

3.1 Transformer Architecture

All multilingual translation systems were developed using the OpenNMT-py framework (Klein et al., 2017), which provides a flexible implementation of the Transformer-based Neural Machine Translation architecture. A common Transformer encoder-decoder architecture (Vaswani et. al., 2017) was employed across all submitted systems to enable multilingual parameter sharing while maintaining a consistent experimental setup. The same model configuration was used for both the forward (English→Indic) and reverse (Indic→English) multilingual models, with only the training data and language-specific adaptation strategies differing between individual systems. The complete architectural configuration and hyperparameters are summarized in Table 3.

Parameter	Value
Encoder layers	6
Decoder layers	6
Hidden Size	512
Word Embedding Size	512
Attention Heads	8
Self Attention	Scaled Dot-Product
Positional Encoding	Enabled
Feed Forward Activation	GELU
Dropout	0.1
Attention Dropout	0.1
Optimizer	Adam
Learning Rate	0.0001

Table 3: Transformer Model Configuration

3.2 Subword Tokenization

To effectively handle rare and out-of-vocabulary words in low-resource multilingual settings, we employed Byte Pair Encoding (BPE) (Sennrich et al., 2016) for subword tokenization. Separate subword vocabularies were learned for the two multilingual language groups to better capture their linguistic characteristics. For the Indo-Aryan multilingual models, a shared BPE vocabulary with 32,000 merge operations was trained over the combined multilingual corpus. In contrast, a common BPE vocabulary with 16,000 merge operations was learned for the Tibeto-Burman multilingual models. Since both English and the participating Tibeto-Burman languages primarily use the Latin script, a shared vocabulary facilitates greater subword overlap and improves parameter sharing across languages while maintaining a compact vocabulary suitable for the comparatively smaller training corpus.

3.3 Language-wise Training Strategy

For language pairs with multiple submitted systems, the primary system was designated as the main system representing our proposed multilingual training strategy, while contrastive systems were used to evaluate additional language-specific adaptation or model combination strategies like fine-tuning the multilingual Tibeto-Burman base model on the official WMT 2026 training set for English to Mizo language pair. Thus, primary and contrastive designations reflect the experimental role of each system rather than the relative performance on the official blind test set.

English → Assamese (En → As): For the primary submission, we selected the multilingual Indo-Aryan base model as the main multilingual system. The base model was trained on BPCC data for Assamese, Bodo, Manipuri (Bengali), Manipuri (Meitei Mayek) along with other low resource Indo-Aryan languages of India. This strategy facilitates cross-lingual vocabulary transfer and multilingual representation learning,

helping mitigate data scarcity for the participating languages. Although Manipuri and Bodo languages are part of Tibeto-Burman language family, they were included in this ‘Indo-Aryan’ grouping owing to shared script with other Indic languages. The contrastive system was obtained by further fine-tuning this model using the official WMT 2026 English–Assamese training data to examine the effect of task-specific adaptation. This strategy enabled us to compare the performance of the generalized multilingual model against a task-specific adapted model.

Assamese → English (As → En): The English translation system for Assamese was developed using the reverse multilingual Indo-Aryan base model with a similar strategy as explained above. The model was trained through multilingual joint learning on the reverse multilingual corpus without any additional fine-tuning or model averaging, allowing the multilingual representations learned during training to be directly evaluated for the reverse translation direction.

English → Mizo (En → Lus): Owing to the limited availability of parallel English–Mizo data including the official WMT 2026 released data, the multilingual Tibeto-Burman base model was augmented using a large synthetic parallel corpus generated with the NLLB-200-distilled-3.3B model using the English side of the English-Hindi BPCC Dataset. This base model served as our primary submission, as this system represents our main strategy for addressing severe parallel-data scarcity. The contrastive system was obtained by further fine-tuning this model using the official WMT 2026 English–Mizo training data to assess the additional effect of task-specific adaptation.

English → Manipuri (Bengali Script) (En → Mni_Beng): The English–Manipuri (Bengali script) translation system was directly obtained from the multilingual Indo-Aryan base model trained using joint multilingual learning. Manipuri (Bengali Script) language translations

were helped by overlapping Vocabulary with the Assamese Language which also uses Bengali Script.

English $\rightarrow$ Manipuri (Meitei Mayek)(En $\rightarrow$ Mni_Mtei): For the English–Manipuri (Meitei Mayek) translation task, the final system was obtained through checkpoint averaging of the multilingual Indo-Aryan base model. Five of the best-performing checkpoints were averaged to improve model stability and generalization, and the resulting averaged model was submitted as the primary system.

Manipuri (Meitei Mayek) $\rightarrow$ English (Mni_Mtei $\rightarrow$ En): The reverse translation system for Manipuri (Meitei Mayek) was developed using the multilingual Indic-to-English base model using a similar strategy as Assamese $\rightarrow$ English translations.. Similar to the forward direction, the final primary submission was generated by averaging the five best-performing checkpoints of the reverse multilingual model, resulting in a more stable and robust translation system.

English $\rightarrow$ Bodo (En $\rightarrow$ Brx): For the primary submission, we selected the weighted checkpoint-averaged model as the main system. A weighted averaging ratio of 3:1 was adopted to emphasize the strongest-performing checkpoint while retaining the benefits of model averaging. Furthermore, the original multilingual base model was submitted as the contrastive system to isolate the effect of weighted checkpoint averaging.

4 Results

4.1 Evaluation Metrics

The submitted systems were evaluated using three widely adopted automatic machine translation metrics: BLEU, chrF++, and COMET. BLEU (Papineni et al., 2002) measures translation quality based on modified n-gram precision with a brevity penalty and remains one of the standard evaluation metrics for machine

translation. chrF++ (Popović, 2017) computes character n-gram F-scores while incorporating word n-grams, making it particularly effective for morphologically rich and low-resource languages. In addition to lexical overlap metrics, we report COMET (Rei et al., 2022), a neural evaluation metric that estimates translation quality by considering the source sentence, machine translation output, and reference translation. Together, these metrics provide complementary perspectives on translation quality.

4.2 Results and Analysis

Table 4 summarizes the performance of the official MTG-LIRA submissions on the WMT 2026 Low-Resource Indic Machine Translation Shared Task. Overall, the results demonstrate that multilingual joint learning provides a strong foundation for low-resource Indic machine translation, while language-specific enhancements such as fine-tuning, synthetic data augmentation, checkpoint averaging, and weighted model averaging further improve translation quality. These findings highlight the importance of selecting appropriate adaptation strategies rather than adopting a uniform training methodology across all language pairs. All scores in Table 4 are from the official WMT 2026 blind test-set evaluation.

Lang. Pairs	Model Type	BLE U	CHR F++	COMET
En $\rightarrow$ As	P	12.17	45.05	78.57
En $\rightarrow$ As	C	14.44	47.04	79.44
As $\rightarrow$ En	P	15.04	43.31	78.77
En $\rightarrow$ Lus	P	16.2	42.06	61.21
En $\rightarrow$ Lus	C	22.72	51.1	65
En $\rightarrow$ Mni_Beng	P	10.61	37.24	67
En $\rightarrow$ Mni_Mtei	P	12.63	46	90.85
Mni_Mtei $\rightarrow$ En	P	24.43	53.66	71.73

En→Brx	P	15.91	48.32	61.75
En→Brx	C	15.87	48.11	61.72

P: Primary system; C: Contrastive system

Table 4: Evaluation Scores

For **English→Assamese translation**, The multilingual Indo-Aryan base model established a strong baseline. Further fine-tuning on the official WMT training data improved translation quality, indicating that multilingual pre-training provides effective cross-lingual representations while task-specific adaptation helps capture domain-specific lexical and contextual patterns.

Similarly, The reverse multilingual model for **Assamese→English** demonstrated that multilingual joint training effectively transfers knowledge from multiple Indic languages for translation into English. The results suggest that the shared multilingual representations learned during training are sufficiently robust to generalize well without requiring additional language-specific adaptation.

For low-resource language pairs like **English→Mizo**, synthetic data generation proved to be an effective strategy for mitigating data scarcity. The addition of large-scale synthetic parallel data improved multilingual representation learning, while subsequent fine-tuning on the official WMT corpus further adapted the model to the target domain, resulting in a substantial improvement in overall translation quality.

The multilingual Indo-Aryan base model produced competitive translations for **English→Manipuri (Bengali script)** language pair. This demonstrates the effectiveness of multilingual transfer learning for closely related scripts of Indic languages, where cross-lingual parameter sharing enables the model to generalize well despite limited parallel training data.

Furthermore, the Checkpoint averaging Method improved the stability and generalization

capability of the multilingual model for the **Meitei Mayek** script. The averaged model produced more consistent translations, which indicates that combining multiple high-performing checkpoints reduces sensitivity to individual checkpoint selection and can improve model stability and generalization..

A similar trend was observed for the reverse translation direction from **Meitei Mayek to English**. The results demonstrate that model averaging is equally beneficial for Indic-to-English translation, improving both translation consistency and overall model generalization.

For **Bodo** language translations, Weighted checkpoint averaging produced consistent improvements over the multilingual Indo-Aryan base model, demonstrating that assigning greater importance to the strongest-performing checkpoint while incorporating complementary information from additional checkpoints leads to a more robust translation system. Although the gains were modest, they validate the effectiveness of weighted model averaging as a lightweight enhancement strategy for multilingual NMT.

5 Conclusion

In this paper, We presented MTG-LIRA’s systems for the WMT 2026 Low-Resource Indic Machine Translation Shared Task and investigated multilingual joint training together with language-specific adaptation strategies for several low-resource Indic language pairs. Our experiments show that the most effective adaptation strategy varies across languages: fine-tuning benefits Assamese and Mizo, synthetic augmentation is particularly useful for severely data-scarce Mizo, while checkpoint and weighted model averaging provide lightweight improvements for other selected language pairs.

Despite these gains, the study has several limitations. In particular, the large synthetic corpus used for Mizo is generated automatically

and may contain translation errors or artifacts inherited from the NLLB model; increasing synthetic data volume therefore does not necessarily imply a proportional increase in data quality. In addition, automatic evaluation metrics may behave differently across scripts and extremely low-resource languages, and neural metrics such as COMET should therefore be interpreted cautiously when evaluating languages with limited representation in their underlying evaluation models. Future work will investigate stronger synthetic-data quality estimation, data-selection and filtering methods, multilingual sampling strategies, and human evaluation for these languages. Extending the framework to the remaining low-resource languages of the shared task is another important direction toward more robust and inclusive Indic machine translation.

References

- Bhagyashree Wagh, Harish Bapat, Neha Gupta, and Saurabh Salunkhe. (2025). Transformers: Leveraging OpenNMT and Transfer Learning for Low-Resource Indian Language Translation. In Proceedings of the Tenth Conference on Machine Translation, pages 1227–1232, Suzhou, China. Association for Computational Linguistics.
- Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., et al. 2022. *No Language Left Behind: Scaling Human-Centered Machine Translation*.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.
- Jay Gala et al. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. Transactions of Machine Learning Research. (For the complete list of all 16 authors, please refer to the referenced document).
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. [Findings of wmt 2024 shared task on low-resource indic languages translation], In Proceedings of the Ninth Conference on Machine Translation, pp. 654-668. 2024.
- 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, “Findings of WMT 2026 shared task on Low-resource Indic Languages Translation”, In Proceedings of the Eleventh Workshop on Machine Translation (WMT) under EMNLP 2026, 28-29 October, 2026, Budapest, Hungary.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318.
- Popović, M. 2017. *chrF++: Words Helping Character n-grams*. Proceedings of the Second Conference on Machine Translation (WMT), pp. 612–618.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. 2022. *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task*. Proceedings of the Seventh Conference on Machine Translation (WMT).
- 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.

Sennrich, R., Haddow, B., and Birch, A. 2016. *Neural Machine Translation of Rare Words with Subword Units*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1715–1725.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS 30).