

PANINI: Improving Low-Resource English $\leftrightarrow$ Manipuri Translation in Two Scripts via rsLoRA and Three-Stream Self-Augmentation

Rudali Huidrom^{1†}, Vikas Kumar^{2†}, Hoomexsun Pangsatabam³, Pinaki Das^{4,5},
K Kabi Khanganba^{6,7}, Nongmaithem Zeno⁸, Nicholas Khumukcham², Mangalton Okram⁹,
Justice Konjengbam², Sai Varun Jamalpoor¹, Alex D Nelson Khangembam¹⁰,
Anand Konjengbam², Vikram Goyal¹

¹IIT Delhi ²IIT Dharwad ³NIT Manipur ⁴IIT Bombay ⁵Monash University

⁶Wadhvani AI ⁷BHASHINI ⁸JNU ⁹IIT Jodhpur ¹⁰Independent Researcher

Correspondence: huidromrudali@gmail.com

Abstract

We describe team PANINI’s submission to the WMT 2026 Shared Task on Low-Resource Indic Language Translation. We submit primary systems for English $\leftrightarrow$ Manipuri (Meiteilon) in *both* the Bengali (mni_Beng) and Meitei Mayek (mni_Mtei) scripts, covering all four directions. Rather than one shared multilingual adapter, we train four independent rank-stabilized LoRA (rsLoRA) adapters over IndicTrans2-1B, one per direction. Each adapter is trained on the authentic bitext plus three filtered synthetic streams the model generates for itself: back-translation, forward translation, and iterative pseudo-labelling, with subword regularisation and a script-aware processing pipeline. Our system obtains a COMET score of 92.31 and a BERTScore of 98.85 on English $\rightarrow$ Manipuri (Meitei Mayek), the best in the direction. On the into-English directions, it reaches a COMET of 82.90 for mni_Beng $\rightarrow$ en and 80.20 for mni_Mtei $\rightarrow$ en. These results show that per-direction adapters combined with self-generated synthetic data are effective even when authentic bitext is limited to a few thousand sentence pairs, and that a script-aware processing pipeline is necessary to produce well-formed output in both Manipuri scripts. The code and resources are released on GitHub.⁰

1 Introduction

Neural machine translation has improved rapidly for high-resource language pairs. Most of the world’s languages have not shared in this progress, because large parallel corpora do not exist for them. The Workshop on Machine Translation (WMT) Low-Resource Indic Language Translation series exists to close this gap for scheduled and under-documented languages of India.

[†] The first two authors made equal contributions to all aspects of the main paper; both are designated as first authors.

⁰<https://github.com/RHuidrom96/WMT-2026-Team-PANINI.git>

Manipuri (Meiteilon) is a Tibeto-Burman language of Northeast India, a scheduled language of the Indian constitution, and one of the more difficult targets in the WMT Low-Resource Indic Language Translation series (Pal et al., 2023; Pakray et al., 2024, 2025). It presents a complication that most of the other languages in the task do not: it is written in *two* scripts. The Bengali-Manipuri script (mni_Beng) is the historically dominant orthography of the modern printed corpus, while Meitei Mayek (mni_Mtei), the indigenous Brahmic abugida, has been officially revived and is increasingly used in education and government. The WMT26 Indic MT task (Pakray et al., 2026) treats the two scripts as separate language pairs, giving four English $\leftrightarrow$ Manipuri directions.

Manipuri has also been a weak point in past WMT editions of this shared task. In 2025, a fine-tuned NLLB-200 (Costa-Jussà et al., 2022) system scored BLEU 4.15 on English $\rightarrow$ Manipuri and only 3.06 on Manipuri $\rightarrow$ English (Joshi et al., 2025), well below that team’s Assamese results in the same submission. A separate method took the opposite approach, injecting word-level noise from Bengali into the source side. It improved Assamese $\rightarrow$ English translation but degraded on Manipuri $\rightarrow$ English, reaching only BLEU 5.74 even with authentic Manipuri data included (Hamada et al., 2025). The authors trace this gap to Manipuri’s greater lexical and grammatical distance from its donor language.

This paper describes team PANINI’s primary submissions to all four directions. Our contributions are:

1. A **per-direction rsLoRA** adaptation of IndicTrans2-1B (Gala et al., 2023), with four independent adapters rather than a single multilingual one, plus a script-aware deterministic pre/post-processing pipeline (§3).
2. A **three-stream self-augmentation** recipe

(back-translation (BT), forward translation (FT), iterative pseudo-labelling (Iter)) where the model generates all its own synthetic data, filtered for quality, mixed at fixed ratios, with subword regularisation.

2 Related Work

Most Indic MT systems in this shared task fine-tune a pretrained Transformer (Pakray et al., 2025). Earlier work on English–Manipuri MT also explored Transformer-based NMT using OpenNMT, demonstrating its effectiveness for low-resource Manipuri translation (Singh et al., 2023b, 2024). Multilingual pretraining is the standard way to compensate for scarce bitext. mBART (Liu et al., 2020) and mBART-50 (Tang et al., 2020) pretrain broadly across many languages. IndicBART (Dabre et al., 2022), IndicTrans (Ramesh et al., 2022), and IndicTrans2 (Gala et al., 2023) pretrain only on Indic languages instead. NLLB-200 (Costa-Jussà et al., 2022) also covers Manipuri, but at a much larger scale and with less language-family focus. We adapt IndicTrans2-1B instead, since it covers only 22 Indic languages and should devote more capacity to Manipuri’s script and morphology.

Augmentation can add noise to existing pairs or generate new pairs entirely; we take the second route. Back-translation (Sennrich et al., 2016a; Hoang et al., 2018) translates target text back into the source with a reverse model. Transfer methods borrow signal from a related high-resource pair instead (Singh et al., 2023a; Zoph and Knight, 2016; Nguyen and Chiang, 2017), and zero-shot multilingual training gets a similar effect through shared parameters (Johnson et al., 2017). Our streams use none of these; all three come from the shared task’s own bitext, generated by the model already being trained on it. For segmentation, byte-pair encoding (Sennrich et al., 2016b) and the SentencePiece unigram model (Kudo and Richardson, 2018) are the usual choices, and BPE was reported as more stable on this task last year (Joshi et al., 2025). We keep IndicTrans2’s vocabulary fixed instead and apply subword regularisation (Kudo, 2018) during training. For adaptation, LoRA (Hu et al., 2022) and rsLoRA (Kalajdzievski, 2023) make fine-tuning a 1B-parameter model practical under shared-task compute limits. We use rsLoRA, since it converged more reliably than plain LoRA at rank 16 in our development runs.

3 Task and Dataset

The organizers released parallel corpora, which are in Table 1 for English↔Manipuri in both scripts, together with development and blind test sets (Pakray et al., 2026). Manipuri appears in Category 1 of the task, alongside Assamese, Mizo and Khasi. The shared task aims to develop high-quality machine translation systems under limited parallel data conditions. We used only the officially provided bitext plus data generated by our own model from it; we used no external monolingual or parallel corpora, so all synthetic data is *self*-augmentation in the strict sense. The raw dataset for English–Manipuri (Bengali script) contains 24,687 sentence pairs; we used 1,000 for development, 1,000 for test and the remaining 22,687 for training. Similarly, the English-to-Meitei raw datasets are around 18k, but we used 16k for training, 1k for dev and 1k for test. Blind test sets of 1,000 sentence pairs were also released by the organizers.

Table 1: Statistics of the parallel corpora used for the English–Manipuri shared task

Lang Pair	Script	Dataset	Parallel sents
Eng–mni	Bengali	Training	22687
		Validation	1000
		Test	1000
Eng–Meitei	Meitei	Training	16000
		Validation	1000
		Test	1000

Prior to model training, we conducted an extensive data analysis and implemented a script-aware processing pipeline to enhance corpus quality. This procedure encompassed the elimination of redundant sentence pairs, the filtration of empty or malformed occurrences, the eradication of erroneous translations, and the assurance of consistent source-target alignment. Upon completion of processing, the sanitized corpora comprise 15,842 training and 1,000 development sentence pairs for the Meitei Mayek script, alongside 21,841 training and 976 development sentence pairs for the Bengali script. The resultant datasets maintain similar lengths for source and target sentences while significantly diminishing noise, thereby offering a dependable basis for model training and assessment. Table 2 presents the statistics of the cleaned datasets used in all the experiments.

Table 2: Statistics of the cleaned English–Meitei parallel corpora used for training and development. Sentence length statistics are reported in tokens. Vocabulary sizes correspond to the number of unique tokens in each split.

Corpus	Split	Pairs	Tokens		Avg. Length		Max. Length (En/Mni)	Vocabulary	
			English	Manipuri	English	Manipuri		English	Manipuri
Meitei Mayek	Train	15,842	287,515	285,061	18.15	17.99	77/89	44,769	54,103
	Dev	1,000	23,905	18,257	23.91	18.26	50/44	6,011	6,993
Bengali	Train	21,841	397,172	335,694	18.18	15.37	126/87	23,716	44,169
	Dev	976	18,514	15,447	18.97	15.83	77/71	4,646	5,984

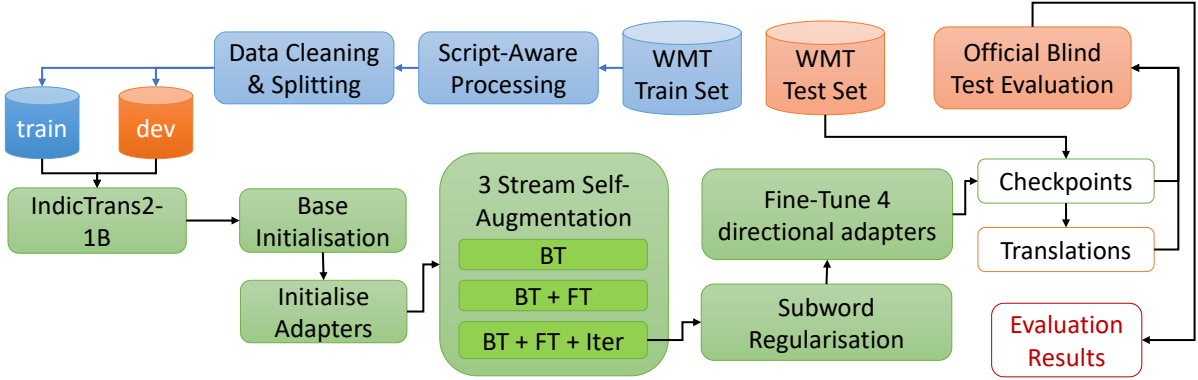


Figure 1: Overview of the PANINI end-to-end translation and self-augmentation pipeline.

3.1 Script-Aware Preprocessing

Column assignment is verified using Unicode block ratios (Meitei Mayek U+AAE0–U+AAF6/U+ABC0–U+ABED; Bengali U+0980–U+09FF), auto-swapping source/target if the detected script contradicts the pipeline’s declared direction. Text is NFC-normalized, stripped of reference markers and dash-joined digit ranges, and passed through script-specific cleanup: native-digit conversion, Devanagari-leak removal, and Latin transliteration for Meitei columns; leaked-script removal and de-duplication for English columns. Pairs that are empty, identical, or outside 2–200/1–200 source/target tokens are filtered, with an explicit train/dev/test overlap check before saving.

3.2 Postprocessing

Decoded output is detokenized via IndicProcessor, and a regular expression removes any leaked leading language tag (e.g. eng_Latn mni_Mtei). For the fine-tuned model, English-target decoding uses additional generation constraints, followed by a direction-specific normalisation pass on the tag-stripped output. References and hypotheses are re-normalised to NFC before scoring to avoid decomposed-Unicode BLEU/chrF deflation, and BLEU, chrF++, TER, METEOR, BERTScore and COMET are computed on the resulting text.

4 System Description

Figure 1 provides a comprehensive overview of our end-to-end system architecture and data pipeline. Starting from the official WMT parallel corpora, our workflow integrates script-aware data sanitization, a three-stream self-augmentation regimen generated entirely by the base model, and the training of four independent, per-direction rsLoRA adapters on top of IndicTrans2-1B.

4.1 Base model and rsLoRA adaptation

We build on IndicTrans2-1B, using the en-indic checkpoint for the two English→Manipuri directions and the indic-en checkpoint for the two Manipuri→English directions. Rather than full fine-tuning, we attach rank-stabilized LoRA adapters (Hu et al., 2022; Kalajdzievski, 2023) via PEFT (Mangrulkar et al., 2022). rsLoRA rescales the adapter by $\alpha/\sqrt{r}$ rather than α/r , which keeps the update magnitude stable as rank grows and, in our development experiments, gave more reliable convergence at $r = 16$ than vanilla LoRA on data of this size.

Adapters are placed on q_proj, k_proj, v_proj, out_proj, fc1 and fc2 with $r = 16$, $\alpha = 32$, dropout 0.05. Training uses AdamW (Loshchilov and Hutter, 2019) with a cosine schedule, warmup ratio 0.05, weight decay 0.01, gradi-

Direction	Pool	chrF++ $\uparrow$	BLEU $\uparrow$	METEOR $\uparrow$	TER $\downarrow$	BERTScore $\uparrow$	COMET $\uparrow$
En $\rightarrow$ Mni (Beng)	14	40.03 (4)	7.48 (11)	20.02 (4)	83.00 (4)	84.48 (7)	67.71 (5)
Mni (Beng) $\rightarrow$ En	14	66.08 (3)	35.79 (5)	71.98 (3)	49.72 (4)	95.74 (3)	82.90 (2)
En $\rightarrow$ Mni (Mtei)	9	45.87 (3)	9.60 (3)	27.55 (3)	81.18 (3)	98.85 (1)	92.31 (1)
Mni (Mtei) $\rightarrow$ En	9	65.12 (3)	36.56 (3)	70.29 (3)	49.12 (2)	96.02 (2)	80.20 (2)

Table 3: PANINI official scores, with rank among primary systems in parentheses. “Pool” is the number of primary systems in the direction.

ent clipping at 1.0, label smoothing 0.1, maximum sequence length 256, and an effective batch size of 32 (per-device 8, gradient accumulation 4), seed 16. The initial adapter (R1) is trained on the authentic bitext for 5 epochs at 2×10^{-4} ; every subsequent augmentation stage is initialised from R1 and fine-tuned for 3 epochs at 1×10^{-4} . Training and inference use bfloat16 and FlashAttention-2 (Dao, 2023) where supported.

Crucially, we train four independent adapters, one per direction. Sharing an adapter between the two scripts is tempting, they encode the same language, but it forces the decoder to disambiguate the output script from the language tag alone, which we found to be the dominant failure mode in early experiments and, as Section A.1 shows, in several other teams’ submissions.

4.2 Three-stream self-augmentation

The submitted system for every direction is the cumulative configuration combining three synthetic streams, all generated by IndicTrans2 itself:

Back-translation (BT). Target-language sentences from the authentic bitext are translated back into the source language with the reverse-direction model, and the synthetic source is paired with the *authentic* target (Sennrich et al., 2016a; Edunov et al., 2018). This preserves a gold target, which matters when the target is the low-resource side.

Forward translation (FT). Source sentences are first synonym-paraphrased (replacement probability 0.15, one variant per sentence) and then translated into the target language (Zhang and Zong, 2016). This diversifies the source side while keeping the source-target relation model-consistent.

Iterative pseudo-labelling (Iter). Source sentences are translated into pseudo-targets by the current adapter and re-added to training (He et al., 2020), with pseudo-labels identical to the gold reference discarded.

BT and FT pairs are filtered by LaBSE cosine similarity (Feng et al., 2022) at a threshold of 0.75;

pseudo-labels are filtered by self-training confidence at 0.78. The surviving streams are mixed with the authentic data at sampling ratios 0.40 (BT), 0.30 (FT) and 0.25 (Iter).

4.3 Subword regularisation

On top of the mixed corpus, we apply subword regularisation (Kudo, 2018) at training time via a custom collator, sampling SentencePiece segmentations (Kudo and Richardson, 2018) with $\alpha = 0.1$. No new data files are produced; the regularisation acts purely as a stochastic re-segmentation of each batch. For an agglutinative target such as Manipuri, this is the cheapest available defence against the model memorising a single segmentation of a morphologically complex word.

4.4 Script-aware processing pipeline

Source side. Unicode NFC normalisation; stripping of reference markers ([1], (1), <word N>); mapping of numeric-range en- and em-dashes to ASCII hyphens (they are otherwise out-of-vocabulary and a reliable source of language-tag artifacts in the output).

Target side. For Meitei Mayek output: script-specific digit and Latin-letter normalisation, and stripping of leaked Devanagari letters (excluding the danda, which is shared punctuation). For English output: acronym-preserving capitalisation, duplicate-word collapsing, and removal of stray Brahmic fragments. Repetitive or over-short outputs are detected and either regenerated greedily or truncated. Every intervention is flagged, so the fraction of sentences touched is measurable.

4.5 Decoding

Beam search with beam size 5, length penalty 1.0, up to 256 new tokens, and a soft repetition penalty of 1.3. The hard n -gram block is *disabled* during the main decode, in an agglutinative, morphologically repetitive target, it suppresses legitimate output and is enabled only on the degeneration-retry path.

4.6 Evaluation

In the official evaluation of WMT, the organizers evaluate the quality of automatic translation using the blind test sets. The word and character level accuracy is calculated using the BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), and chrF++ (Popović, 2017). In order to measure semantic similarity apart from just matching references, the evaluation also incorporates neural metrics such as BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2022).

5 Official Results & Analysis

Our submitted system to the Shared Task is PANINI. Table 3 reports our official scores across chrF++, BLEU, METEOR, TER, BERTScore and COMET for all four directions, along with our rank among primary systems. For English→Manipuri (Bengali), our system reaches a chrF++ of 40.03, a METEOR of 20.02, a TER of 83.00, and a COMET of 67.71, placing fourth on most of these metrics but only eleventh on BLEU (7.48). On Manipuri (Bengali)→English, the system performs more strongly, with a chrF++ of 66.08, a BLEU of 35.79, a METEOR of 71.98, a BERTScore of 95.74, and a COMET of 82.90, ranking second or third across the board. The English→Manipuri (Meitei Mayek) direction gives our best result overall: a COMET of 92.31 and a BERTScore of 98.85, the highest of any system in the direction, even though BLEU (9.60) and chrF++ (45.87) place us only third. For Manipuri (Meitei Mayek)→English, we obtain a chrF++ of 65.12, a BLEU of 36.56, a TER of 49.12, and a COMET of 80.20, again ranking second on COMET.

Two patterns stand out across the four directions. First, our COMET and BERTScore scores are consistently strong even where BLEU is not, most visibly on English→Manipuri (Meitei Mayek), where we obtain a COMET of 92.31 and a BERTScore of 98.85 while BLEU remains at 9.60. This gap suggests that our outputs are semantically faithful and well-formed even when they do not match the reference word-for-word. Second, our BLEU scores stay in a moderate, stable range across all four directions (7.48–36.56), with no direction collapsing to a near-zero score, which points to the robustness of training separate adapters for each script and direction rather than sharing parameters across them. Taken together, these patterns

indicate that per-direction adapters, combined with self-generated synthetic data and script-aware post-processing, produce translations that remain consistent in quality across both scripts and both translation directions, even in cases where surface-level overlap with the reference is limited.

6 Conclusion & Future Work

We described PANINI’s submission to the WMT 2026 Shared Task on Low-Resource Indic Language Translation, covering all four English↔Manipuri directions in both the Bengali and Meitei Mayek scripts. Our approach combined per-direction rsLoRA adapters over IndicTrans2-1B with three self-generated synthetic streams, back-translation, forward translation, and iterative pseudo-labelling, alongside subword regularisation and a script-aware processing pipeline. This setup achieved a COMET of 92.31 and a BERTScore of 98.85 on English→Manipuri (Meitei Mayek), and COMET scores of 82.90 and 80.20 on the two into-English directions. Our analysis also showed that BLEU should be treated with caution on English→Manipuri (Bengali), where it disagreed with every other automatic metric, and that training a dedicated adapter for each script is important, since script confusion was a common failure mode in our early experiments.

In future work, we plan to extend the script-aware pipeline with a learned quality-filtering model in place of fixed LaBSE and confidence thresholds, and to explore combining IndicTrans2 with NLLB-200 outputs through ensembling. We also intend to investigate curriculum strategies that introduce the synthetic streams gradually rather than in fixed sampling ratios, which may further improve stability on the Bengali-script direction. Additionally, we will perform ablation studies of the model used on the low-resource language pair in our future study.

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.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard,

- and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863.
- Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.
- Jay Gala, Pranjal A. Chitale, A. K. Raghavan, 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. *Transactions on Machine Learning Research*.
- Shoki Hamada, Tomoyosi Akiba, and Hajime Tsukada. 2025. Akibanlp-tut: Injecting language-specific word-level noise for low-resource language translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1259–1264.
- Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations (ICLR)*.
- Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd workshop on neural machine translation and generation*, pages 18–24.
- 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 (ICLR)*.
- Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, and 1 others. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Nisheeth Joshi, Palak Arora, Anju Krishnia, Riya Lonchenpa, and Mhasilenuo Vizo. 2025. Bvslp: Machine translation using linguistic embellishments for indicmt shared task 2025. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1265–1270.
- Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 66–75.
- 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.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>.
- Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: short papers)*, pages 296–301.
- Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Kumar 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 (WMT)*, pages 532–553.

- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of the WMT 2024 shared task on low-resource Indic language translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 654–668.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 2026. Findings of WMT 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Conference on Machine Translation (WMT)*, Budapest, Hungary.
- 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 (WMT)*, pages 682–694.
- 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 (ACL)*, pages 311–318.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 612–618.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 186–191.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, and 1 others. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.
- Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLP AI).
- Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.
- Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.
- Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1535–1545.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.
- Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34.

Team	chrF++ rank	BLEU rank	COMET rank	Worst BLEU
gators	2.25	3.00	3.25	8.04
PANINI	3.25	5.50	2.50	7.48
Winterpeg	5.75	5.50	5.75	0.00
EROL	6.50	5.75	7.50	3.96
BVSLP	8.25	6.25	7.25	3.82
RozarNLP-cr7	8.75	8.75	9.00	0.00

Table 4: Mean rank over the four Manipuri directions, for the six teams entering all four. “Worst BLEU” is the team’s lowest BLEU across them.

A Comparative Analysis

A.1 Script robustness separated the field

Of the 24 participating teams, only six submitted primary systems in all four Manipuri directions: gators, PANINI, EROL, BVSLP, Winterpeg and RozarNLP-cr7. Two of the six collapsed:

- Winterpeg *wins* En→Mni(Beng) outright (BLEU 11.95, COMET 70.36, first on both) and places third on Mni(Beng)→En, but scores **BLEU 0.00 in both Meitei Mayek directions** (COMET 77.39 and 40.19, last in both).
- RozarNLP-cr7 is competitive on Mni(Beng)→En (BLEU 33.52) but degenerates to **BLEU 0.00 on En→Mni(Beng)** (chrF++ 0.06, BERTScore 3.56, an empty-output failure) and BLEU 0.13 on Mni(Mtei)→En.

This is the clearest system-level finding in the Manipuri track. A team can hold first place in one script and produce nothing at all in the other. Our lowest BLEU across the four directions is 7.48; for both Winterpeg and RozarNLP-cr7 it is 0.00. Table 4 summarises the six all-four teams.

PANINI has the **best mean COMET rank (2.50) of any team covering all four directions**, narrowly ahead of gators (3.25). We state this carefully: the margin comes almost entirely from gators’ comparatively weak En→Mni(Beng) COMET (rank 9/14), and gators beats us on every surface metric on average. The defensible claim is that gators is the strongest overall system in the track, and that PANINI is the most *consistent* system on semantic adequacy across both scripts and both directions.

We attribute this consistency to two design choices rather than to model capacity, which we did not have: four independent per-direction adapters,

Direction	<i>n</i>	MET.	TER	chrF++	COMET
En→Mni(Beng)	12	-0.27	+0.05	-0.33	-0.20
Mni(Beng)→En	11	+0.92	-0.91	+0.96	+0.91
En→Mni(Mtei)	8	+0.93	-0.95	+0.74	+0.86
Mni(Mtei)→En	7	+1.00	-0.71	+1.00	+0.93

Table 5: Spearman ρ between BLEU and each other official metric, across systems within a direction. TER is inverted, so negative ρ there means agreement.

and a deterministic script-aware post-processing pass that guarantees the output is well-formed in the intended script.

A.2 On En→Mni(Beng), BLEU disagrees with everything else

Our weakest official rank is BLEU 11/14 on En→Mni(Beng), against ranks 4-5 on METEOR, TER, chrF++ and COMET in the same direction. This is not noise. Table 5 reports the Spearman rank correlation between BLEU and each other metric, computed across systems within each direction (degenerate systems with BLEU < 1 excluded).

In three of the four directions, BLEU is near-perfectly rank-correlated with every other metric ($\rho = +0.74$ to $+1.00$). In En→Mni(Beng) it is mildly *anti*-correlated with all of them. The pattern is visible in the leaderboard itself: the three systems with the best METEOR / chrF++ / TER / COMET after Winterpeg, DELAB-IIITM (BLEU rank 12, COMET rank 2), JH_NLP_APAShreya2 (BLEU rank 10, COMET rank 2) and PANINI (BLEU rank 11, COMET rank 5), all sit in the bottom four on BLEU, while MTG-LIRA is second on BLEU and eleventh on COMET, and NLP-MT_Amrita_IndicTrans2 is third on BLEU and eleventh on METEOR.

Why BLEU is the wrong instrument here.

BLEU scores exact matches of *word-level* n -grams against a *single* reference. Four properties of the target defeat it. (i) *The word is the wrong unit*. Manipuri is agglutinative: an orthographic word packs a root with a chain of bound morphemes marking aspect, mood, case and number, so a BLEU token is not a lexical choice but a whole morphological derivation. A correct root carrying an equally acceptable affix is penalised exactly as heavily as an unrelated word. Character n -gram metrics exist to recover precisely this lost partial credit (Popović, 2017), and our chrF++ rank (4) against our BLEU rank (11) is that credit made visible. (ii) *The penalty is multiplicative*. A reference

of L tokens contains $L - 3$ four-grams, and one divergent token invalidates up to four of them. Agglutination compresses content into fewer, longer words, so L is small: in a 12-token sentence a single morphological variant destroys some 44% of the 4-gram precision mass, and BLEU’s geometric mean over orders amplifies the resulting near-zero counts. (iii) *One reference cannot cover the admissible space*. Manipuri is verb-final with relatively free constituent order and optional case marking; several orderings are equally grammatical, and all but one earn no credit. (iv) *Bengali-script orthography is not fully standardised*. Spelling, conjunct formation and clitic spacing vary across sources, so strings that are phonologically and semantically identical differ as word tokens while remaining near-identical as character sequences. Point (iv) also explains why the effect is confined to this direction: Meitei Mayek shares Manipuri’s morphology exactly, so morphology alone cannot account for Table 5; what differs is orthographic uniformity, with Meitei Mayek being a recently re-standardised script.

What this means for our system. Our output here is not less accurate, it is less *reference-identical*. Self-augmentation drifts the output distribution towards the model’s own morphological realisations, and subword regularisation deliberately pulls the decoder off memorised reference token sequences; the product is well-formed, morphologically consistent and semantically faithful output that no longer reproduces the reference’s particular spellings and orderings. Every metric able to see past those choices, METEOR, TER, chrF++ (all rank 4) and COMET (rank 5), places us in the top third of a fourteen-system field; only the one that cannot does not. The systems ranked immediately above us on BLEU sit near the bottom on COMET (MTG-LIRA: BLEU 2, COMET 11), a trade we would not make. This is consistent with the broader finding that neural metrics track human judgement far better than BLEU (Freitag et al., 2022).

Implication for checkpoint selection. The analysis applies to development-set model selection as much as to the leaderboard. Our internal selection used sacreBLEU (Post, 2018) with 13a tokenisation uniformly across the four pipelines, a defensible default, and the right one for the three directions in which BLEU tracks the other metrics. For En→Mni(Beng), however, a BLEU-based

criterion is selecting *against* the very property the augmentation is meant to induce, and we expect a chrF++- or COMET-based criterion to recover further headroom in that direction at no cost elsewhere. We regard this, rather than the model or the data, as the most promising single change for the next iteration.

A.3 Direction and script asymmetry

The best-in-track BLEU is 45.86 into English from Bengali-script Manipuri, but only 11.95 into Bengali-script Manipuri and 16.24 into Meitei Mayek. Every team, ours included, is roughly three to four times better translating *out of* Manipuri than *into* it, the standard low-resource morphology asymmetry, but unusually severe here. The into-English directions are also where other teams failed hardest on the *source* side: on Mni(Beng)→En, Akiba-NLP scores BLEU 0.50 with TER 292.36, NLP-MT_Amrita_IndicTrans2 scores 0.28, and IndicMT scores 0.10. Feeding Bengali-script Manipuri to an unadapted Indic→English model appears to fail catastrophically, presumably through confusion with Bengali proper. A per-direction adapter over the indic-en checkpoint sidesteps this entirely.