

RozarNLP-cr7 at WMT 2026: LoRA-Fine-Tuned NLLB for Low-Resource Machine Translation in Northeastern Indian Languages

Suyash Yadav
Sitare University, India
yadavsuyash723@gmail.com

Utkarsh Chaudhary
IIITDM Kurnool, India
123cs0012@iiitk.ac.in

Alapan Kuila
IIITDM Kurnool, India
alapan.cse@iiitk.ac.in

Abstract

We describe the RozarNLP-cr7 constrained submissions to the WMT 2026 Indic Machine Translation Shared Task. We adapt the permitted NLLB-200 Distilled 1.3B checkpoint with two bidirectional multilingual LoRA adapters trained exclusively on the official parallel datasets provided by the WMT 2026 organizers. We report official direction-level results and supplement them with WMT25 cross-benchmark validation and an output-side analysis in the appendix. The adapted systems outperformed the submitted NLLB baseline in both Mizo directions and IndicTrans2 for Bengali-script Manipuri-to-English, but performance was uneven. In particular, the primary system performed strongly for Bengali-script Manipuri-to-English but failed to produce satisfactory reverse-direction translations; the cause remains unresolved. Overall, the findings underscore the need to validate parameter-efficient adaptation separately across languages, scripts, and translation directions.

1 Introduction

The WMT 2026 Indic Machine Translation Shared Task evaluates translation between English and low-resource languages of Northeastern India.¹ We participated in the constrained setting for Assamese, Manipuri in Bengali and Meitei Mayek scripts, Mizo, and Khasi. These directions combine limited parallel data with Bengali, Latin, and Meitei Mayek writing systems. They therefore test whether a shared parameter-efficient model can support translation across differences in script, parallel-data availability, and pretrained language coverage. Joint adaptation may enable transfer, but shared parameters can also couple languages with different

¹Northeastern India comprises the states of Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, and Tripura; the regional label provides geographic context and does not imply that each language is spoken only within these states.

needs; we assess performance separately by language, script, and direction.

Our systems adapt NLLB-200 Distilled 1.3B (NLLB Team et al., 2022) with LoRA (Hu et al., 2022). We refer to the adapted systems as NLLB+LoRA and the unadapted checkpoint as NLLB. Model A is shared across Assamese, Mizo, and both Manipuri scripts, while Model B covers Khasi and Nyishi. Both adapters are bidirectional and use language identifiers for directional control; pretrained IndicTrans2 (Gala et al., 2023) and unadapted NLLB provide the available comparisons.

Because WMT26 test references were unavailable, we supplement the official scores with an output-side inspection (Appendix B) and a separate cross-benchmark evaluation on the WMT25 gold test set (WMT25 Indic MT Organizers, 2025). Neither analysis substitutes for reference-based WMT26 evaluation.

Our contributions are:

- multilingual LoRA systems with documented data construction, language identifiers, tokenizer extension, training, and inference;
- direction-wise WMT26 comparisons with submitted baselines and the strongest primary submissions; and
- output-side error inspection and reference-based WMT25 validation.

2 Task Description

Constrained setting. We participate under the *constrained* track. We used the permitted pretrained NLLB-200 and IndicTrans2 checkpoints and adapted only NLLB, using official parallel corpora; no monolingual data were used.²

²Training code and configurations: <https://github.com/Iamdevsuyash/WMT26>.

Language	Pairs	Bidirectional
Assamese	54,000	108,000
Manipuri (Bengali)	23,687	47,374
Manipuri (Meitei Mayek)	17,000	34,000
Mizo	18,964	37,928
Khasi	24,700	49,400
Nyishi	56,000	112,000
Total	194,351	388,702

Table 1: Official training pairs with English and directional instances after augmentation. Nyishi was used in training before withdrawal from official evaluation.

Language pairs. We cover both English translation directions for the five language–script conditions above. Nyishi was withdrawn after the training-data release; it remains in our Khasi adapter and WMT25 evaluation, but has no official WMT26 results.

For consistency, direction labels use mni for Bengali-script Manipuri, mni-Mtei for Manipuri in Meitei Mayek script, and njz for Nyishi. Full tokenizer identifiers retain their FLORES-200-style script suffixes, such as mni_Mtei and njz_Latn (NLLB Team et al., 2022). We use only the FLORES-200 language-code naming convention; no FLORES-200 evaluation sentences are used for training, validation, or testing.

Metrics. Official test-set evaluation reports BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), chrF++ (Popović, 2017), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2022). Higher values are better for every reported metric except TER. We emphasize BLEU to compare submissions on the same shared-task test set under the same scoring conditions. The other metrics provide complementary evidence, including semantic quality through COMET.

3 Dataset

The shared task provided official parallel training data but no official development or validation set. Table 1 gives the released training sizes and the sizes after bidirectional augmentation.

We normalize language identifiers and remove records with unknown codes or empty source/target strings. The training and development splits already exist separately before bidirectional augmentation. Within each split, each remaining parallel pair (x, y) yields two directional instances, (x, y)

and (y, x) , with source and target identifiers; both directions remain in the same split. The resulting records are shuffled within each split with seed 42. We use the development split for internal validation and assess translation quality separately using official WMT26 results and WMT25 references; internal scores appear in Table 3.

4 Related Work

Multilingual Transformers (Vaswani et al., 2017; Johnson et al., 2017) share representations across directions; mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and NLLB-200 (NLLB Team et al., 2022) extend this through pretraining. LoRA learns low-rank updates to a frozen backbone (Hu et al., 2022), trading adaptation capacity against efficiency (Xu et al., 2023).

IndicTrans2 covers scheduled Indian languages (Gala et al., 2023), while WMT Indic tasks benchmark low-resource translation across multilingual, back-translation, SMT, and parameter-efficient approaches (Pal et al., 2023; Pakray et al., 2024, 2025). English–Nyishi translation has also been studied directly (Kakum et al., 2023). We focus on shared multilingual LoRA adaptation with pretrained systems as comparisons.

5 System Description

Model selection. NLLB-200 is a translation-specific multilingual encoder–decoder with broad low-resource coverage, including native support for Assamese, Bengali-script Manipuri, and Mizo (NLLB Team et al., 2022). It provides a consistent architecture and pretrained initialization for our LoRA experiments, though not equivalent coverage of all tested scripts. Comparing adaptation across other backbones, such as mBART-50, M2M-100, MADLAD-400, or IndicTrans2, is outside our scope; IndicTrans2 is evaluated only as a pretrained comparison system.

5.1 Multilingual Architecture and Adapters

We train two independent multilingual LoRA adapters for NLLB-200 Distilled 1.3B in separate runs, freezing the pretrained backbone. Model A covers English–Assamese, English–Mizo, and English–Manipuri in Bengali and Meitei Mayek scripts; Model B covers English–Khasi and English–Nyishi. Each pair uses its assigned adapter in both directions, jointly learned with source and target language identifiers. The adapters differ in

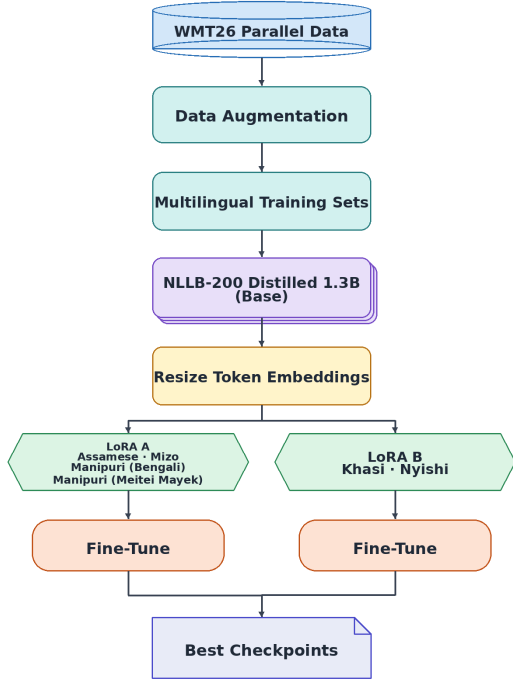


Figure 1: Training pipeline. Model A covers Assamese, Mizo, and Manipuri in both scripts; Model B covers Khasi and Nyishi. All pairs are bidirectional with English.

training-language composition and learned parameters, not merely initialization. The heuristic grouping combines three languages across three scripts in Model A and two Latin-script languages in Model B. Without grouping ablations, we cannot isolate script-sharing effects. Figure 1 shows the workflow; data construction, LoRA, training, and decoding settings appear in Sections 3, 5.4, 6.1, and 6.4.

5.2 Vocabulary Extension

NLLB-200 natively supplies identifiers for Assamese, Mizo, and Bengali-script Manipuri. We add identifiers for the other language-script conditions listed in Table 2; these tokens provide directional control but not lexical or script knowledge. Tokenizer and embedding implementation details appear in Appendix A.

5.3 Meitei-Mayek Tokenization

We added the previously absent `mni_Mtei` special token, but kept the original shared subword vocabulary for source segmentation. Across 11 inspected Meitei-Mayek inputs, 285 of 578 tokens (49.3%) were `<unk>`; every input contained unknown tokens. Sentence-level counts and the exact inputs appear in Appendix C. This limited script coverage is a plausible adaptation constraint: language identifiers and

Language	Code	Status
English	<code>eng_Latn</code>	Native
Assamese	<code>asm_Beng</code>	Native
Manipuri (Bengali)	<code>mni_Beng</code>	Native
Manipuri (Meitei Mayek)	<code>mni_Mtei</code>	Added
Mizo	<code>lus_Latn</code>	Native
Khasi	<code>kha_Latn</code>	Added
Nyishi	<code>njz_Latn</code>	Added

Table 2: Language identifiers used by the submitted configurations.

attention-only LoRA do not add lexical units. However, these observations do not establish the cause of generation errors.

5.4 LoRA Parameterization

We freeze the NLLB backbone and train LoRA updates with rank $r = 8$, scaling factor $\alpha = 16$, dropout 0.05, and no bias updates on `q_proj`, `k_proj`, `v_proj`, and `out_proj`. Following PEFT defaults, the low-rank matrices use Kaiming-uniform initialization for A and zero initialization for B , giving an initially zero adapter update. Sharing adapters limits memory use but couples languages with different pretrained coverage.

6 Experiments

6.1 Training Configuration

Both adapters are trained for three epochs with AdamW (Hugging Face defaults), a learning rate of 1×10^{-4} , a cosine schedule, and a 0.10 warmup ratio. The per-device batch size is 4, with 8 gradient accumulation steps (effective batch size 32). Source and target sequences are limited to 256 tokens with dynamic padding and truncation. We use BF16 when supported, otherwise FP16, with TF32 and gradient checkpointing enabled. With seed 42, we evaluate every 250 steps and select the checkpoint with the lowest internal-validation loss (Section 3). Hardware and implementation details appear in Appendix A.

6.2 Internal Validation

With WMT26 test references unavailable, we used the held-out development split for internal model selection and diagnostics. Table 3 reports corpus-level BLEU and chrF++ for the trained NLLB+LoRA model, evaluated separately by translation direction. These are internal diagnostics, not official WMT26 test-set results.

Direction	BLEU	chrF++
as→en	33.65	53.05
en→as	17.01	43.02
mni→en	34.87	60.70
en→mni	15.92	55.02
lus→en	37.51	55.57
en→lus	22.03	46.72

Table 3: NLLB+LoRA scores on the 5% internal validation split (Section 3). These non-independent diagnostics are not official WMT26 results.

6.3 Manipuri (Meitei Mayek) Adaptation

During preliminary experiments, NLLB-200 produced empty, repetitive, or malformed outputs for English→Manipuri (Meitei Mayek). We therefore selected pretrained IndicTrans2, which explicitly supports Manipuri in Meitei Mayek script (Gala et al., 2023), as the primary system in both directions. Model A was retained as the official mni→en contrastive submission; its result is discussed in Section 7.2.

6.4 Inference

At inference time, we force the requested target-language identifier and use batched 3-beam search (Freitag and Al-Onaizan, 2017) with `early_stopping=True`. Inputs are truncated to 256 tokens, and `max_new_tokens=256` limits generation. Target control uses `forced_bos_token_id`; special tokens and language prefixes are removed using `strip_lang_prefix`.

6.5 Submitted Systems

NLLB+LoRA is primary for Assamese, Bengali-script Manipuri, Mizo, and Khasi. Contrastive submissions use IndicTrans2 for the first two and NLLB for Mizo; no Khasi baseline was submitted. The Meitei Mayek submissions are described above. All system assignments appear with their scores in Table 4.

7 Official WMT26 Results

7.1 Official Scores

Table 4 reports all official metrics for our primary and contrastive submissions (Pakray et al., 2026). Directions are ordered by descending primary-system BLEU, with each available contrastive system listed immediately below its primary counterpart.

7.2 Direction-wise Comparison

Mizo and Assamese. NLLB+LoRA improves over NLLB in both Mizo directions: by 3.81 BLEU into English and 2.67 out of English. For Assamese, it trails IndicTrans2 by 1.69 BLEU into English and 0.75 in the reverse direction.

Manipuri (Bengali). Our highest primary score, 33.52 BLEU into English, exceeds IndicTrans2 by 10.88 BLEU and 4.74 chrF++. In contrast, English-to-Manipuri scores 0.00 BLEU against 10.18 for IndicTrans2, despite reaching 15.92 on internal validation (Table 3) and 4.64 on WMT25 (Table 5). This discrepancy suggests instability across evaluation conditions, not a general inability to translate into Bengali-script Manipuri. The available records cannot distinguish adaptation, decoding, or submission-time processing as the cause.

Manipuri (Meitei Mayek). The NLLB+LoRA contrastive system scores 33.75 BLEU into English, compared with 0.13 for the IndicTrans2 primary; IndicTrans2 reaches 3.46 in the reverse direction. These results do not constitute a bidirectional NLLB comparison. Transfer through the shared backbone and English decoder may contribute to the strong English-target result, but this hypothesis was not tested. Generating English avoids Meitei-Mayek output tokens without resolving the source-side coverage limitation (Section 5.3).

Khasi. NLLB+LoRA scores 9.04 BLEU from English and 4.54 into English, with TER reaching 189.51 in the latter direction. No Khasi baseline was submitted, so these results cannot isolate the effect of adaptation.

7.3 Comparison with Leading Primary Systems

Our primary submissions trail the strongest primary systems in the official results (Pakray et al., 2026). The BLEU gaps are 12.34 for Bengali-script Manipuri-to-English, 3.76 and 4.14 for Assamese into and out of English, and 2.06 and 1.78 for Mizo, respectively. For Khasi, the strongest primary scores are 29.30 from English and 26.28 into English, well above our 9.04 and 4.54.

8 Cross-Benchmark Validation on WMT25

We compare base and adapted checkpoints on the WMT25 gold test set (WMT25 Indic MT Organizers, 2025) (Table 5). This is a separate benchmark,

Direction	RozarNLP-cr7 system	BLEU↑	METEOR↑	TER↓	chrF++↑	BERTScore↑	COMET↑
mni→en	NLLB+LoRA (primary)	33.52	70.51	50.03	63.81	95.38	81.69
	IndicTrans2 (contrastive)	22.64	61.99	71.33	59.07	92.69	72.13
lus→en	NLLB+LoRA (primary)	31.40	65.18	56.15	60.08	95.35	79.94
	NLLB (contrastive)	27.59	60.89	61.34	56.87	94.81	77.54
as→en	NLLB+LoRA (primary)	23.50	57.95	67.20	54.08	94.56	85.49
	IndicTrans2 (contrastive)	25.19	58.89	67.97	55.47	94.72	85.58
en→lus	NLLB+LoRA (primary)	19.33	46.08	67.11	47.93	81.24	64.68
	NLLB (contrastive)	16.66	41.85	70.50	45.39	80.05	64.48
en→as	NLLB+LoRA (primary)	15.24	36.91	72.40	49.53	86.10	82.15
	IndicTrans2 (contrastive)	15.99	40.84	64.43	50.49	86.43	82.54
en→kha	NLLB+LoRA (primary)	9.04	32.34	144.02	29.25	72.90	55.24
kha→en	NLLB+LoRA (primary)	4.54	25.05	189.51	25.75	87.84	51.38
en→mni-Mtei	IndicTrans2 (primary)	3.46	17.98	119.20	38.60	96.67	84.57
mni-Mtei→en	IndicTrans2 (primary)	0.13	6.92	137.67	13.42	84.65	42.06
	NLLB+LoRA (contrastive)	33.75	67.20	53.67	62.93	95.45	78.44
en→mni	NLLB+LoRA (primary)	0.00	0.17	100.00	0.06	3.56	31.03
	IndicTrans2 (contrastive)	10.18	17.81	86.81	39.03	84.26	67.60

Table 4: Official WMT26 scores for the RozarNLP-cr7 submissions. Higher is better for BLEU, METEOR, chrF++, BERTScore, and COMET; lower is better for TER. Metric values are reproduced on the organizers’ displayed scale from the uploaded official final-results file; BERTScore and COMET are displayed on a 0–100 scale. Model A is defined in Section 5.1.

not a second shared-task submission; dataset and domain differences prevent direct score comparisons with WMT26.

Direction	NLLB	NLLB+LoRA	Δ
en→as	18.00	17.69	−0.31
as→en	20.55	22.97	+2.42
en→lus	20.38	24.01	+3.63
lus→en	16.79	17.61	+0.82
en→mni	7.67	4.64	−3.03
mni→en	11.37	10.47	−0.90
en→kha	0.00	7.19	+7.19
kha→en	0.00	2.25	+2.25
en→njz	0.00	1.49	+1.49
njz→en	0.00	8.98	+8.98

Table 5: Cross-benchmark BLEU validation using WMT25 gold references. Positive Δ indicates improvement over the corresponding base checkpoint.

Mizo improves in both directions, and Assamese-to-English gains 2.42 BLEU. Bengali-script Manipuri declines in both directions relative to NLLB; its WMT26 gain was against a different comparator, IndicTrans2. Khasi and Nyishi improve over zero-scoring base models, but absolute scores remain low.

9 Limitations

Resource constraints prevented controlled ablations of backbone size, LoRA rank, corpus size, tokenizer extension, language grouping, and decoding. We therefore cannot attribute gains or failures to individual design choices. Comparisons also use different baselines across languages, and Khasi has no submitted baseline; they do not uniformly isolate the effect of LoRA. WMT25 provides a separate check, but dataset and domain differences limit generalization to WMT26.

Unavailable WMT26 references and limited source-language expertise restrict our manual inspection to output form, not translation adequacy. The English-to-Manipuri (Bengali) failure could not be reproduced, leaving its cause unresolved. The 11-input Meitei-Mayek tokenization analysis is illustrative, not a corpus-wide coverage estimate.

10 Conclusion

We presented two bidirectional multilingual LoRA adapters for constrained low-resource translation with NLLB-200 Distilled 1.3B. Mizo shows the most consistent gains over the unadapted checkpoint across WMT26 and WMT25. Bengali-script Manipuri improves over IndicTrans2 into English on WMT26, but its reverse-direction failure rules

out a claim of uniformly reliable adaptation. Limited Meitei-Mayek vocabulary coverage further highlights the need to distinguish language control from script support.

The central lesson is to validate adaptation separately by language, script, and direction. Future work should prioritize script-aware tokenization, controlled adapter-grouping comparisons, and reproducible checks of decoding and submission pipelines.

References

- Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic MT Evaluation*, 2005.
- Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3207.
- 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. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages, 2023. URL <https://arxiv.org/abs/2305.16307>.
- Edward J. Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.
- Melvin Johnson et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. Neural machine translation for limited-resource English–Nyishi pair. *Sādhana*, 2023.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No Language Left Behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. doi: 10.48550/arXiv.2207.04672.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Sarah Lyngdoh, and Riyanka Manna. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, 2024.
- 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. Findings of WMT 2025 shared task on low-resource Indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 532–553, 2025.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetri-mayum 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 Conference on Machine Translation*, Budapest, Hungary, October 2026. Forthcoming, October 28–29, 2026; author list and pagination to be updated from the final proceedings.
- Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 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, 2023.
- Kishore Papineni et al. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- Maja Popović. chrF++: words helping character n-grams. In *WMT*, 2017.
- Ricardo Rei et al. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *WMT*, 2022.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 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*, pages 223–231, Cambridge, Massachusetts, USA, 2006. Association for Machine Translation in the Americas.
- Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- WMT 2025 Low-Resource Indic Languages Translation Shared Task Organizers. Test Data Gold WMT 2025. Google Drive dataset archive, 2025. URL <https://drive.google.com/file/d/1C2slIg368xamg9-AnuWN5rFpVeB01ct/view>. Gold reference test data for the WMT 2025 Low-Resource Indic Languages Translation Shared Task; accessed August 5, 2026.
- Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023. URL <https://arxiv.org/abs/2312.12148>.
- Linting Xue et al. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.
- Tianyi Zhang et al. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

A Reproducibility Details

Table 6: Implementation and execution environment.

Category	Configuration
Model	facebook/nllb-200-distilled-1.3B; encoder-decoder Transformer; multilingual sequence-to-sequence
Tokenizer	Native NLLB tokenizer where available; added identifiers from Table 2; resized input/output embeddings; new rows initialized to the pretrained-embedding mean
Hardware	NVIDIA RTX 2000 Ada, 17 GB VRAM; CUDA 12.8; Ubuntu/Linux
Software/time	Hugging Face Transformers and PEFT (exact versions not recorded); training time not recorded

B Qualitative Analysis of English Outputs

We manually inspected the English translations generated for the Indic→English directions for visible surface-level phenomena, including fluency, repetition, formatting, serialization artifacts, and unusual register. This inspection illustrates observable generation behavior only. Because source-language adequacy and reference-based correctness were not assessed, the examples should not be interpreted as quantitative estimates of translation error or as bilingual human evaluation.

Manual inspection. We assigned descriptive, non-exclusive labels to visible English-output errors; outputs without evident defects were marked “no clear surface error.” Representative examples appear in Appendix ??, Table ?. They document observed output patterns, not the frequency of individual error types.

Assamese. The inspected outputs included fluent and well-formed English translations as well as outputs containing spacing and punctuation errors. For example, one output contains missing whitespace around an apostrophe and a quoted expression, as in *major’re d flag’for*. Other inspected examples did not exhibit an obvious surface-level error.

Manipuri (Bengali). The inspected outputs contained several distinct forms of malformed generation, including literal \u09BC sequences, repeated possessive suffixes, citation-like index sequences, repeated quotation marks, and malformed or implausible phrases such as repeated “2019 BC.” Some outputs began with fluent English before containing Unicode escape sequences, repeated quotation marks, or citation-like sequences such as [1] [2] [3] [4].... The examples document these observable output patterns; their underlying causes cannot be determined from the generated outputs alone.

Mizo. The inspected outputs included fluent and well-formed English translations. The representative examples did not exhibit an obvious surface-level error. This observation concerns the form of the generated English and does not establish adequacy with respect to the source sentences.

Khasi. The inspected outputs included repetition and unusual register. For example, one output repeatedly generates the phrase *of the gate*. Another combines contemporary place names such as Assam and Manipur with archaic diction, including *thy* and *thee*. These examples illustrate the types of repetition and register inconsistencies observed in the generated English. They do not establish the relative frequency of these phenomena across languages or constitute bilingual human evaluation.

C Meitei-Mayek Tokenization Examples

Table 8 lists the 11 representative Meitei-Mayek input sentences used for the tokenizer analysis in Section 5.3. The sentence IDs correspond to those reported in Table 7.

Table 7: NLLB tokenization counts for inputs M1–M11. Unknown tokens (<unk>) account for 49.3% of all tokens.

ID	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	Total
Tokens	47	55	42	42	59	46	48	68	48	64	59	578
<unk>	23	27	21	21	28	23	24	34	23	32	29	285

Table 8: Representative Meitei-Mayek input sentences used for tokenizer analysis.

[illegible]