Towards High-Quality Machine Translation for Kokborok: A Low-Resource Tibeto-Burman Language of Northeast India

Badal Nyalang * and Biman Debbarma

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2322.v1

Keywords: machine translation; low-resource NLP; endangered languages; data augmentation; human evaluation; South Asian languages; kokborok; kokborokMT

 doi:10.20944/preprints202603.2322.v1

Article

Towards High-Quality Machine Translation for Kokborok: A Low-Resource Tibeto-Burman Language of Northeast India

Badal Nyalang [1,*] and Biman Debbarma [2]

[1] MWire Labs, Shillong, Meghalaya, India
[2] Department of Kokborok, Tripura University, Agartala, Tripura, India
* Correspondence: badal@mwirelabs.com

Abstract

We present KokborokMT, a high-quality neural machine translation (NMT) system for Kokborok (ISO 639-3: `trp`), a Tibeto-Burman language spoken primarily in Tripura, India with approximately 1.5 million speakers. Despite its status as an official language of Tripura, Kokborok has remained severely under-resourced in the NLP community, with prior machine translation attempts limited to systems trained on small Bible-derived corpora achieving BLEU scores below 7. We fine-tune the NLLB-200-distilled-600M model on a multi-source parallel corpus comprising 36,052 sentence pairs: 9,284 professionally translated sentences from the SMOL dataset, 1,769 Bible-domain sentences from WMT shared task data, and 24,999 synthetic back-translated pairs generated via Gemini Flash from Tatoeba English source sentences. We introduce `trp_Latn` as a new language token for Kokborok in the NLLB framework. Our best system achieves BLEU scores of 17.30 (en→trp) and 38.56 (trp→en) on held-out test sets, representing substantial improvements over prior published results. Human evaluation by three annotators yields mean adequacy of 3.74/5 and fluency of 3.70/5, with substantial agreement between trained evaluators ($\kappa = 0.67$). We will release the model, data, and code publicly under CC-BY-4.0 upon acceptance.

Keywords: machine translation; low-resource NLP; endangered languages; data augmentation; human evaluation; South Asian languages; kokborok; kokborokMT

1. Introduction

Kokborok is one of the indigenous languages spoken by the Tiprasa people of Tripura, Northeastern India. The name itself is a compound of *kok* (language) and *borok* (people), literally meaning "language of the people." With approximately 1.5 million speakers across Tripura, the Chittagong Hill Tracts of Bangladesh, and other parts of Northeast India, Kokborok holds official language status in Tripura alongside Bengali. It belongs to the Bodo-Garo branch of the Tibeto-Burman language family within the larger Sino-Tibetan family, and is characterised by SOV word order, postpositions, and tonal phonology.

Despite its official status and substantial speaker population, Kokborok remains severely under-resourced in natural language processing. Prior computational work has been largely limited to morphological analysis [1], POS tagging, and rule-based named entity recognition. Machine translation for Kokborok has received minimal attention: the WMT Low-Resource Indic Language Translation shared tasks [2–4] have included Kokborok since 2023, providing the only published NMT baselines. The best WMT 2025 submission [5] achieved BLEU scores of 6.99 (en→trp) and 2.99 (trp→en). These low scores reflect the extreme data scarcity and domain restriction of available training data, rather than any inherent untranslatability of the language.

In this paper, we address this gap with the following contributions:

 doi:10.20944/preprints202603.2322.v1

- We develop KokborokMT, a significantly improved NMT system for Kokborok, by fine-tuning NLLB-200 [6] with a novel `trp_Latn` language token.
- We construct a 36,052-sentence parallel corpus combining professional translations from SMOL [7], WMT Bible data, and synthetic back-translations generated from Tatoeba English sentences using Gemini Flash.
- We demonstrate that synthetic data augmentation via LLM back-translation yields consistent improvements across all evaluation metrics and test conditions.
- We provide comprehensive ablation studies comparing zero-shot NLLB, fine-tuning without synthetic data (System 1), and fine-tuning with synthetic data (System 2).
- We investigate LaBSE-based quality filtering for synthetic data and report that filtering scores are unreliable for Kokborok due to the language's absence from LaBSE's training data, an important negative finding for the community.
- We conduct human evaluation with three annotators including a linguistic expert and a domain specialist, yielding mean scores of 3.74/5 (adequacy) and 3.70/5 (fluency).
- We release the model and evaluation scripts publicly to facilitate further research on Kokborok NLP.

2. Background and Related Work

2.1. Kokborok: Language and Script

Kokborok has a native script called Koloma, used during the reign of the Tripura kings, which is currently undergoing a revival effort. However, contemporary digital use and NLP research overwhelmingly employs the Roman script, which we adopt throughout this work. The language has nine dialects named after tribal communities (Debbarma, Reang, Jamatia, Noatia, etc.). Kokborok is an SOV language, uses suffixes as tense markers (*-o* for present, *-kha* for past, *-nai* for future), and employs two phonemic tones (level and high). Adjectives follow the nouns they modify, and plural markers appear sentence-finally on nouns.

2.2. Prior NLP Work on Kokborok

Computational work on Kokborok has been sparse. [1] developed a morphological analyser achieving approximately 80% accuracy. POS taggers using CRF and SVM approaches have been reported at around 84% accuracy. A rule-based NER system achieved 83% F-score [8]. Vowel recognition using LPCC features has also been explored. For machine translation, the WMT Low-Resource Indic Language Translation shared task has included Kokborok since 2023 [2–4], providing the only published NMT baselines. The best WMT 2025 submission [5] achieved BLEU of 6.99 (en→trp) and 2.99 (trp→en). We also note that OPUS and HuggingFace contain no publicly available Kokborok parallel data beyond the WMT shared task releases, confirming the extreme scarcity of resources for this language.

2.3. Low-Resource MT and Back-Translation

Back-translation [9] is a well-established technique for augmenting low-resource MT training data by translating monolingual target-side text into the source language. Recent work has demonstrated that LLMs can serve as effective generators of synthetic parallel data for low-resource settings. The NLLB-200 model [6] has become a standard backbone for low-resource MT fine-tuning, covering 200 languages with strong multilingual representations that transfer well to unseen languages through continued training. Adding new language tokens to multilingual models and fine-tuning on target language data is an established approach for extending coverage beyond the original training set. Fine-tuning NLLB on domain-specific or language-specific data consistently improves over zero-shot performance for languages at the margins of its coverage.

 doi:10.20944/preprints202603.2322.v1

3. Data

3.1. Parallel Corpus Construction

Our training corpus combines three sources, totalling 36,052 sentence pairs after preprocessing and deduplication.

3.1.1. SMOL (9,284 Sentences)

The SMOL dataset [7] provides professionally human-translated parallel data for 123 low-resource languages. For Kokborok, SMOL comprises two sub-datasets: SMOLDOC (6,016 sentences), consisting of LLM-generated English documents on diverse topics professionally translated into Kokborok; and GATITOS (4,211 sentences), a token-level resource. An additional 57 SMOLSENT sentences were found to have reversed source and target columns during preprocessing and were corrected prior to use. SMOL represents the highest-quality component of our training data, covering diverse domains including health, education, culture, technology, and everyday conversation.

3.1.2. WMT Bible Corpus (1,769 Sentences)

The WMT Low-Resource Indic Language Translation shared task [4] provides 2,269 Kokborok-English parallel sentence pairs derived from Bible translations. We reserve 500 sentences as a held-out test set and use the remaining 1,769 for training. While this corpus is domain-restricted, it provides additional training signal and enables direct comparison with prior WMT baselines on the same test distribution.

3.1.3. Synthetic Back-Translation (24,999 Sentences)

We generate synthetic en→trp parallel data using English source sentences from the Tatoeba project (`agentlans/tatoeba-english-translations` on HuggingFace) [10], a well-known multilingual sentence collection used widely in MT research. We apply a length filter retaining sentences between 5 and 20 words, followed by deduplication using exact string matching, yielding 25,000 unique English sentences. Kokborok translations were generated in batches using the Google Gemini Flash API (`gemini-2.5-flash-preview` model) with the following system instruction: *"You are a professional English to Kokborok translator. Translate each line accurately. Maintain the line order. Output ONLY the translations."* The total API cost was approximately INR 600 (USD $7). Tatoeba sentences cover everyday conversational and factual domains, complementing the more formal register of SMOLDOC.

3.2. Quality Filtering Investigation

Following standard practice in synthetic MT pipelines, we investigated LaBSE-based quality filtering [11] to identify and remove low-quality synthetic pairs. We computed cosine similarity between LaBSE embeddings of English source and Kokborok target sentences across all 24,999 pairs. The resulting score distribution (mean: 0.287, std: 0.216) was substantially lower than typical cross-lingual similarity scores for supported languages. Manual inspection of pairs across all score ranges confirmed that translation quality was consistently acceptable even at very low similarity scores (e.g., 0.04–0.15). We attribute the low scores to Kokborok's absence from LaBSE's training languages, rendering the embeddings unreliable for cross-lingual alignment with Kokborok. We therefore retain all 24,999 synthetic pairs without filtering and report this as a cautionary finding: LaBSE-based quality filtering is not applicable to languages absent from the model's training data.

3.3. Data Splits and Deduplication

We construct evaluation sets from the two highest-quality sources:

- **SMOL Test Set (500 sentences):** Randomly sampled from SMOLDOC sentences only, ensuring domain diversity and professional translation quality.
- **WMT Test Set (499 sentences):** Randomly sampled from the WMT Bible corpus, enabling direct comparison with WMT shared task results.

 doi:10.20944/preprints202603.2322.v1

- **Development Set (500 sentences):** Randomly sampled from remaining SMOL sentences after test extraction.

Zero overlap between training data (including synthetic pairs) and all test sets is verified by exact English-side string matching prior to any model training. The final training set contains 36,052 sentence pairs across all three sources.

3.4. Data Statistics

Table 1 summarises the corpus composition. Figure 1 (right) shows the relative contribution of each source.

Table 1. Corpus statistics for KokborokMT.

Source	Sentences	Type
SMOL (train)	9,284	Professional
WMT Bible (train)	1,769	Professional
Gemini BT (Tatoeba)	24,999	Synthetic
Total Train	**36,052**	
Dev (SMOL)	500	Professional
Test (SMOL)	500	Professional
Test (WMT)	499	Professional

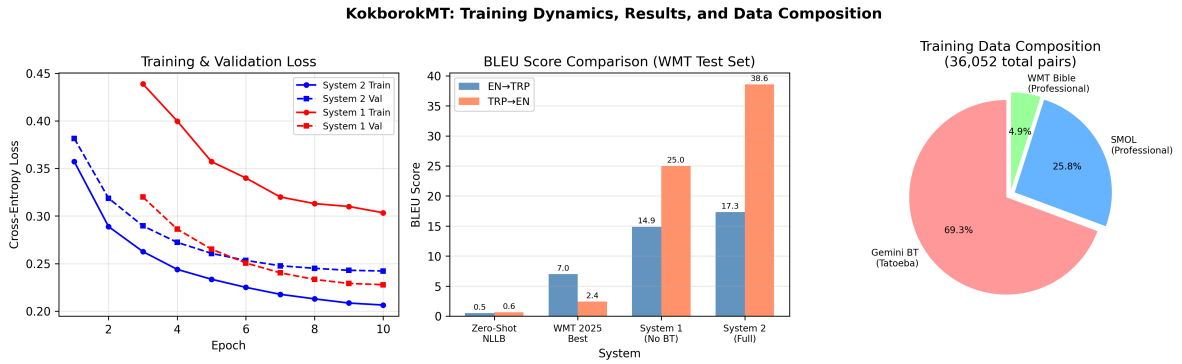

Figure 1. Left: Training and validation loss curves for System 1 (no synthetic data, red) and System 2 (full pipeline, blue). System 1 shows instability at epoch 1 due to smaller dataset size but converges cleanly from epoch 3. Centre: BLEU score comparison across all systems on the WMT test set for both translation directions. Right: Training data composition showing contribution of each source to the 36,052 total sentence pairs.

4. Methodology

4.1. Base Model and Language Token

We fine-tune `facebook/nllb-200-distilled-600M` [6], a 600M parameter sequence-to-sequence transformer covering 200 languages. Kokborok is not among NLLB's supported languages. We add a new special token `trp_Latn` to the tokenizer vocabulary (assigned ID 256204) and resize the model's embedding matrix accordingly. This allows the model to condition generation on Kokborok as a distinct target language while leveraging representations from typologically related Tibeto-Burman languages already present in NLLB such as Burmese and Tibetan.

4.2. Training Setup

We train both translation directions simultaneously by concatenating the original and direction-flipped datasets, yielding 72,096 training pairs. This joint training approach encourages shared representations for both directions.

 doi:10.20944/preprints202603.2322.v1

4.2.1. Hyperparameters

We train for 10 epochs using the AdamW optimiser with learning rate 2e-5, linear warmup over 500 steps, weight decay 0.01, and batch size 32. Mixed precision training (fp16) is used throughout. Maximum sequence length is 128 tokens. Training is conducted on a single A40 GPU, completing in approximately 3.5 hours for System 2 and 1.1 hours for System 1.

4.2.2. Model Selection

Checkpoints are saved at each epoch and the best model is selected based on validation loss on the SMOL development set. For both systems, validation loss continued to decrease through epoch 10, with final checkpoints selected (System 2 val loss: 0.2422; System 1 val loss: 0.2278).

4.3. Experimental Conditions

We evaluate three systems:

- **Zero-Shot NLLB:** The base NLLB-200-distilled-600M model with `trp_Latn` token added but no fine-tuning.
- **System 1 (No BT):** Fine-tuned on SMOL + WMT only (11,053 pairs; 23,098 with both directions). No synthetic data.
- **System 2 (Full):** Fine-tuned on SMOL + WMT + Gemini synthetic data (36,052 pairs; 72,096 with both directions). Primary system.

5. Evaluation

5.1. Automatic Metrics

We evaluate using a comprehensive suite of automatic metrics matching the WMT shared task evaluation protocol [4]:

- **BLEU** [12]: Computed using sacreBLEU [13] with default tokenisation.
- **chrF** [14]: Character n-gram F-score via sacreBLEU.
- **ROUGE-L** [15]: Longest common subsequence F-measure.
- **METEOR** [16]: Alignment-based metric using WordNet.
- **TER** [17]: Translation edit rate (lower is better).
- **Cosine Similarity:** Semantic similarity via LaBSE embeddings [11].
- **COMET** [18]: Neural evaluation using `Unbabel/wmt22-comet-da`.

All systems use beam search with beam size 4. We evaluate both directions on both test sets (four conditions per system).

5.2. Human Evaluation

5.2.1. Data

Human evaluation was conducted on 50 en→trp translations generated by System 2. Source sentences were sampled from the Tatoeba dataset [10] (`agentlans/tatoeba-english-translations`, HuggingFace), separately from all training and automatic evaluation data. From 6,765,220 total rows, we first deduplicated on the English column (1,417,346 unique sentences), then filtered for readability ≥ 2.5 and quality ≥ 0.0, yielding 222,317 candidate sentences. We randomly sampled 50 sentences (random_state=99) covering everyday conversational and factual domains.

5.2.2. Annotators

Three annotators independently rated all 50 translations: (1) a linguistic expert with expertise in Kokborok language structure; (2) a native Kokborok speaker; and (3) a native researcher specialising in Kokborok linguistics. Annotators were not informed of each other's ratings.

 doi:10.20944/preprints202603.2322.v1

5.2.3. Criteria

Each translation was rated on two 1–5 scales: **Adequacy** (does the translation preserve the meaning of the source?) and **Fluency** (is the translation natural and grammatically correct?), following standard MT human evaluation practice.

5.2.4. Results

Table 2 presents individual and aggregate scores. Agreement between the two trained evaluators (linguistic expert and native researcher) was substantial ($\kappa = 0.67$ for both adequacy and fluency), while the untrained native speaker showed lower agreement ($\kappa = 0.13$), consistent with known scale interpretation variability among non-expert annotators in MT human evaluation [19]. Mean scores across all annotators of 3.74/5 (adequacy) and 3.70/5 (fluency) indicate that KokborokMT successfully preserves meaning in most cases and produces largely natural output.

Table 2. Human evaluation results (1–5 scale, $n = 50$). Cohen's κ reported for each annotator pair.

Annotator	Adequacy	Fluency
Linguistic expert	3.76	3.76
Native speaker	3.64	3.54
Native researcher (Kokborok)	3.84	3.80
Mean	**3.74**	**3.70**
κ (expert vs researcher)	0.672	0.677
κ (expert vs native)	0.134	0.109
κ (native vs researcher)	0.004	0.019

5.3. Reproducibility

Full sacreBLEU signatures for System 2 are provided in Appendix A for exact reproducibility.

5.4. Automatic Evaluation Results

Table 3 presents full automatic evaluation results. Figure 1 (centre) visualises BLEU scores across systems.

Table 3. Automatic evaluation results. System 2 (Full) is our primary system. TER is lower-is-better; all other metrics are higher-is-better. WMT 2025 best system results are from Pakray et al. [4] and were trained exclusively on Bible-domain data; our systems additionally use SMOL professional translations, making direct comparison indicative rather than strictly controlled.

System	Direction	BLEU	chrF	ROUGE-L	METEOR	TER↓	Cos Sim	COMET
SMOL Test Set (general domain)								
Zero-Shot NLLB	en→trp	0.50	11.89	0.0261	0.0132	139.51	0.1939	0.2697
Zero-Shot NLLB	trp→en	0.63	17.07	0.0675	0.0526	130.30	0.1872	0.2880
System 1 (No BT)	en→trp	13.35	46.22	0.3873	0.3112	75.95	0.7707	0.6938
System 1 (No BT)	trp→en	32.91	49.41	0.5498	0.5091	55.07	0.7466	0.6604
System 2 (Full)	en→trp	**15.25**	**47.67**	**0.3896**	**0.3138**	**74.26**	0.7596	**0.6958**
System 2 (Full)	trp→en	**38.56**	**53.92**	**0.5919**	**0.5602**	**50.15**	**0.7911**	**0.6926**
WMT Test Set (Bible domain)								
Zero-Shot NLLB	en→trp	0.09	12.76	0.0136	0.0056	121.59	0.3361	0.2545
Zero-Shot NLLB	trp→en	0.32	16.89	0.0560	0.0390	123.30	0.2701	0.2888
System 1 (No BT)	en→trp	14.87	42.34	0.3908	0.3136	86.11	0.7268	0.6718
System 1 (No BT)	trp→en	24.99	45.14	0.5078	0.4306	70.23	0.7889	0.6413
System 2 (Full)	en→trp	**17.30**	**47.11**	**0.4332**	**0.3483**	**76.81**	**0.7479**	**0.7064**
System 2 (Full)	trp→en	**28.03**	**48.18**	**0.5449**	**0.4713**	**66.31**	**0.8171**	**0.6640**
WMT 2025 Shared Task Best Systems [4] (Bible test set only; other metrics not reported)								
WMT Best	en→trp	6.99	38.08	0.367	0.300	76.26	–	–
WMT Best	trp→en	2.99	25.52	0.218	0.163	117.73	0.487	–

 doi:10.20944/preprints202603.2322.v1

6. Analysis

6.1. Impact of Fine-Tuning

Zero-shot NLLB produces near-zero BLEU scores (0.09–0.63) for both directions, confirming that Kokborok lies outside NLLB's effective coverage despite the addition of the `trp_Latn` token. Fine-tuning with even a modest gold corpus (System 1, ∼11k pairs) produces dramatic improvements, reaching BLEU 32.91 (trp→en, SMOL) and 14.87 (en→trp, WMT). This represents a roughly 30× improvement over zero-shot, demonstrating the critical importance of even modest amounts of high-quality parallel data for languages absent from multilingual model coverage.

6.2. Impact of Synthetic Back-Translation

System 2 consistently outperforms System 1 across all four evaluation conditions and all metrics. The gains are most pronounced for trp→en on the SMOL test set (+5.65 BLEU), where the model benefits most from the additional Tatoeba-derived training signal. For en→trp, gains are more modest (+1.90 BLEU on SMOL, +2.43 on WMT), reflecting the asymmetric nature of our synthetic data: Gemini translated English → Kokborok, so the synthetic data primarily strengthens the en→trp direction at training time, yet the bidirectional training setup propagates improvements to both directions. The consistent improvements validate Gemini Flash as a practical source of augmentation data for extremely low-resource languages at minimal cost.

6.3. LaBSE Quality Filtering

We investigated LaBSE-based quality filtering on the 24,999 synthetic pairs. Mean cross-lingual cosine similarity was 0.287 (std: 0.216), far below typical values for supported language pairs. At thresholds of 0.3, 0.4, 0.5, and 0.6, only 44.6%, 32.8%, 20.0%, and 10.0% of pairs would be retained respectively. Manual inspection confirmed translation quality was acceptable across all score ranges, including pairs with similarity scores below 0.1. We conclude that LaBSE similarity scores are not a reliable quality signal for Kokborok due to its absence from LaBSE's training data. This is an important negative finding: quality filtering methods that rely on multilingual embeddings should be validated on a per-language basis before application to truly unseen languages.

6.4. Human Evaluation Analysis

The human evaluation scores of 3.74/5 (adequacy) and 3.70/5 (fluency) are consistent with the automatic metric performance. The native researcher in Kokborok linguistics gave the highest scores (3.84/3.80), while the untrained native speaker gave the lowest (3.64/3.54), a pattern commonly observed in MT human evaluation where expert annotators tend to apply scales more consistently [19]. The strong agreement between the two trained evaluators ($\kappa = 0.67$) provides confidence in the reliability of the human assessment. Overall, the scores indicate the system is practically useful — preserving most meaning and producing largely natural output — while acknowledging room for improvement, particularly in fluency.

6.5. Domain Generalisation

Both systems achieve competitive scores on both the general-domain SMOL test set and the Bible-domain WMT test set, demonstrating reasonable cross-domain generalisation despite training on data from both domains. The higher trp→en BLEU scores on SMOL (38.56) versus WMT (28.03) reflect domain match with the SMOL training component.

6.6. Translation Direction Asymmetry

trp→en consistently outperforms en→trp across all systems and test sets. System 2 achieves BLEU 38.56 (trp→en) versus 15.25 (en→trp) on the SMOL test set. This asymmetry is expected: translating into English benefits from NLLB's strong English generation capabilities, while generating Kokborok requires producing a low-resource language with limited representation in the base model. Closing

 doi:10.20944/preprints202603.2322.v1

this gap likely requires more gold Kokborok-side training data and potentially monolingual continued pretraining.

6.7. Comparison with Prior Work

Our System 2 achieves BLEU 17.30 (en→trp) versus the WMT 2025 best of 6.99, and BLEU 38.56 (trp→en) versus 2.99. We note that this comparison is not strictly controlled: WMT systems trained exclusively on Bible-domain data, while our systems additionally use SMOL professional translations covering diverse domains. The performance gap reflects both the richer training data and the effectiveness of our fine-tuning approach. Even on the WMT Bible test set alone, our system substantially outperforms WMT baselines, suggesting that the improvements are not solely attributable to domain match.

7. Limitations

This work has several limitations. First, our synthetic back-translation data was generated from English source sentences only (en→trp direction via Gemini), limiting Kokborok-source diversity in synthetic data. Future work could generate trp→en synthetic data from monolingual Kokborok text. Second, automatic metrics have known limitations for low-resource languages where reference translations may have limited lexical overlap with model outputs; our human evaluation partially addresses this concern. Third, we evaluate only on Roman-script Kokborok; the Bengali-script variant used in some official contexts is not addressed. Fourth, we do not attempt tokenizer adaptation or monolingual continued pretraining for Kokborok, which may further improve generation quality. Fifth, our comparison with WMT 2025 systems is indicative rather than strictly controlled due to differences in training data. Sixth, human evaluation was conducted only for the en→trp direction; trp→en human evaluation is left for future work.

8. Conclusion

We have presented KokborokMT, a significantly improved neural machine translation system for Kokborok, achieving BLEU 17.30 (en→trp) and 38.56 (trp→en) and substantially surpassing prior published results. Human evaluation by three annotators confirms practical translation quality with mean adequacy 3.74/5 and fluency 3.70/5. Our work demonstrates that a combination of professionally translated data from SMOL, domain-specific parallel data, and LLM-generated synthetic back-translations can bootstrap effective MT for an extremely low-resource Tibeto-Burman language. We additionally report that LaBSE-based quality filtering is unreliable for languages absent from the model's training data, a cautionary finding for the community. We will release our model and evaluation pipeline publicly under CC-BY-4.0 to support further research on Kokborok and other under-resourced languages of Northeast India.

Data Availability Statement: All datasets used in this work are publicly available under open licenses: SMOL (CC-BY), Tatoeba (CC-BY), and WMT Bible data (public domain). Synthetic back-translations were generated using the Google Gemini Flash API in compliance with its terms of service for research use. Human evaluators participated voluntarily and are anonymised in this paper. The model will be released under CC-BY-4.0 upon acceptance. We acknowledge that MT systems for low-resource languages may produce errors that could mislead users; we recommend human review for critical applications.

Acknowledgments: We thank the SMOL team at Google for releasing high-quality Kokborok translations, and the WMT shared task organisers for maintaining the Low-Resource Indic Language Translation benchmark. We are grateful to our human evaluators for their careful assessments.

Appendix A. sacreBLEU Signatures

The following sacreBLEU signatures are provided for exact reproducibility of System 2 results:

 doi:10.20944/preprints202603.2322.v1

- EN→TRP (SMOL): `BLEU = 15.25 47.9/20.4/10.2/5.5 (BP=1.000 ratio=1.009 hyp_len=8142 ref_len=8068)`
- EN→TRP (WMT): `BLEU = 17.30 46.9/22.6/12.0/7.4 (BP=0.988 ratio=0.988 hyp_len=12733 ref_len=12884)`
- TRP→EN (SMOL): `BLEU = 38.56 65.5/43.1/32.2/24.7 (BP=0.997 ratio=0.997 hyp_len=8595 ref_len=8620)`
- TRP→EN (WMT): `BLEU = 28.03 58.2/33.8/21.7/14.5 (BP=1.000 ratio=1.010 hyp_len=14381 ref_len=14243)`

COMET scores computed using `Unbabel/wmt22-comet-da` via `unbabel-comet`. LaBSE embeddings via `sentence-transformers` library.

Appendix B. Training Loss Curves

Table A1. Training and validation loss curves for both fine-tuned systems.

Epoch	Train Loss	Val Loss
System 2 (Full, With BT)		
1	0.3573	0.3816
2	0.2889	0.3187
3	0.2626	0.2896
4	0.2439	0.2723
5	0.2335	0.2607
6	0.2250	0.2533
7	0.2177	0.2477
8	0.2129	0.2450
9	0.2085	0.2429
10	0.2064	0.2422
System 1 (No BT)		
1	8.4454	1.4459
2	1.8205	0.3779
3	0.4388	0.3199
4	0.3998	0.2862
5	0.3571	0.2651
6	0.3399	0.2506
7	0.3200	0.2403
8	0.3130	0.2333
9	0.3100	0.2291
10	0.3034	0.2278

Appendix C. Human Evaluation Data Pipeline

Human evaluation sentences were sampled as follows: source dataset `agentlans/tatoeba-english-translations` (HuggingFace) [10], 6,765,220 total rows; after deduplication on the English column: 1,417,346 unique sentences; after filtering (readability ≥ 2.5, quality ≥ 0.0): 222,317 sentences; random sample of 50 sentences (`random_state=99`). Domain: general/everyday (conversation, facts, opinions, geography, daily life). These sentences are entirely separate from all training, development, and automatic evaluation splits.

References

1. Debbarma, K.; Patra, B.G.; Das, D.; Bandyopadhyay, S. Morphological Analyzer for Kokborok. In Proceedings of the Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing. The COLING 2012 Organizing Committee, 2012, pp. 41–52.
2. Pal, S.; Pakray, P.; Laskar, S.R.; Laitonjam, L.; Khenglawt, V.; Warjri, S.; Dadure, P.K.; Dash, S.K. Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation. In Proceedings of the Proceedings of the Eighth Conference on Machine Translation. Association for Computational Linguistics, 2023, pp. 682–694. https://doi.org/10.18653/v1/2023.wmt-1.56.

 doi:10.20944/preprints202603.2322.v1

3. Pakray, P.; Pal, S.; Vetagiri, A.; Krishna, R.; Dash, S.K.; Maji, A.K.; Laitonjam, L.; Lyngdoh, S.; Manna, R. Findings of the WMT 2024 Shared Task on Low-resource Indic Languages Translation. In Proceedings of the Proceedings of the Ninth Conference on Machine Translation. Association for Computational Linguistics, 2024, pp. 654–668. https://doi.org/10.18653/v1/2024.wmt-1.54.
4. Pakray, P.; Krishna, R.M.; Pal, S.; Vetagiri, A.; Dash, S.K.; Maji, A.K.; Lyngdoh, S.A.; Laitonjam, L.; Jamatia, A.; Sambyo, K.; et al. Findings of WMT 2025 Shared Task on Low-resource Indic Languages Translation. In Proceedings of the Proceedings of the Tenth Conference on Machine Translation (WMT). Association for Computational Linguistics, 2025. https://doi.org/10.18653/v1/2025.wmt-1.29.
5. ANVITA Team. ANVITA: A Multi-pronged Approach for Enhancing Machine Translation of Extremely Low-Resource Indian Languages. In Proceedings of the Proceedings of the Tenth Conference on Machine Translation (WMT). Association for Computational Linguistics, 2025, pp. 1240–1247. https://doi.org/10.18653/v1/2025.wmt-1.101.
6. NLLB Team.; Costa-jussà, M.R.; Cross, J.; Çelebi, O.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672* **2022**. https://doi.org/10.48550/arXiv.2207.04672.
7. Caswell, I.; Nielsen, E.; Luo, J.; Cherry, C.; Kovacs, G.; Shemtov, H.; Talukdar, P.; Tewari, D.; et al. SMOL: Professionally Translated Parallel Data for 115 Under-Represented Languages. In Proceedings of the Proceedings of the Tenth Conference on Machine Translation (WMT). Association for Computational Linguistics, 2025. https://doi.org/10.18653/v1/2025.wmt-1.85.
8. Debbarma, A.; Bhattacharya, P.; Purkayastha, B.S. Named Entity Recognition for a Low Resource Language. In Proceedings of the International Journal of Recent Technology and Engineering, 2019, Vol. 8, pp. 587–590. https://doi.org/10.35940/ijrte.B2085.098319.
9. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2016, pp. 86–96. https://doi.org/10.18653/v1/P16-1009.
10. Tiedemann, J. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In Proceedings of the Proceedings of the Fifth Conference on Machine Translation. Association for Computational Linguistics, 2020, pp. 1174–1182. https://doi.org/10.18653/v1/2020.wmt-1.139.
11. Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; Wang, W. Language-agnostic BERT Sentence Embedding. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 878–891. https://doi.org/10.18653/v1/2022.acl-long.62.
12. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135.
13. Post, M. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Proceedings of the Third Conference on Machine Translation: Research Papers. Association for Computational Linguistics, 2018, pp. 186–191. https://doi.org/10.18653/v1/W18-6319.
14. Popović, M. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2015, pp. 392–395. https://doi.org/10.18653/v1/W15-3049.
15. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out. Association for Computational Linguistics, 2004, pp. 74–81.
16. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, 2005, pp. 65–72.
17. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, 2006, pp. 223–231.
18. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A Neural Framework for MT Evaluation. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

 doi:10.20944/preprints202603.2322.v1

Association for Computational Linguistics, 2020, pp. 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213.

19. Graham, Y.; Baldwin, T.; Moffat, A.; Zobel, J. Continuous Measurement Scales in Human Evaluation of Machine Translation. In Proceedings of the Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Association for Computational Linguistics, 2013, pp. 33–41.