Lost in Translation? Not Anymore: Empowering Mizo, Khasi, Tulu, and Gondi with LLMs and Encoder–Decoder Models

Anonymous ACL submission

Abstract

We address machine translation for four extremely low-resource Indian languages: Mizo, Khasi, and Tulu, for which we release the largest publicly available parallel corpora (49k, 22k, and 16k sentence pairs respectively), along with curated monolingual corpora and benchmark evaluation sets created from scratch. For Gondi, we utilize and curate 15k sentence pairs from Microsoft Research India's Project ELLORA dataset. We benchmark three model families spanning different parameter scales — mBART-large-50 (610M), NLLB-200-1.3B (1.3B), and Gemma 3 4B-IT (4B) — under zero-shot, supervised fine-tuning (SFT), and continual pre-training followed by SFT (CPT+SFT). Our fine-tuned systems achieve strong results across all four languages: NLLB-200 reaches 43.72 BLEU for English–Tulu, and Gemma 3 4B-IT achieves 34.17 BLEU for English–Khasi using a single joint adapter for all four languages. All datasets and model adapters are publicly released.[1]

1 Introduction

India's extraordinary linguistic diversity presents a significant challenge for language technology. Of the over 1,369 languages spoken across the country (Census Commissioner, 2022), only 22 are constitutionally recognised (Indian-Constitution, 2022), and fewer still are supported by modern machine translation (MT) systems. This gap has real consequences: speakers of unsupported languages are effectively excluded from the digital ecosystem, from access to government services, education, and information.

Neural MT has advanced rapidly with multilingual models such as NLLB-200 (NLLB Team et al., 2022) and mBART-50 (Liu et al., 2020), and more recently instruction-tuned large language models (LLMs) like Gemma 3 (Team, 2025), with approximately 4 billion parameters. However, these models require substantial parallel data and natively support only languages that were well represented during pre-training. Hundreds of Indian languages, including many spoken by millions of people, remain entirely absent from these systems.

English: The sun is shining in the sky.

Mizo: Ni chu vânah a ên mêk .

Khasi : Ka sngi ka shai thaba ha ka suin bneng.

Tulu: ಸೂರ್ಯೆ ಆಕಾಶೊಡು ಮಿಂಚೊಂದುಂಡು.
Sūrye ākāśoḍu miñconduṇḍu

Gondi: आकासम न्हें पोड्डू मिंजिटातो |
Ākāsama nhēṁ pōḍḍū miñjiṭātō |

Figure 1: Example sentences in Mizo, Khasi, Tulu, and Gondi for the English sentence "The sun is shining in the sky."

We focus on four such languages: Mizo (Tibeto-Burman, Latin script, Mizoram), Khasi (Austroasiatic, Latin script, Meghalaya), Tulu (Dravidian, Kannada script, coastal Karnataka and Kerala), and Gondi (Dravidian, Devanagari script, central India). Example sentences are shown in Figure 1. Three of these four languages are entirely absent from the vocabularies of NLLB-200, mBART-50, and Gemma 3, requiring vocabulary extension before any fine-tuning can take place.

This paper makes two primary contributions. First, we release the **largest publicly available parallel corpora** for these language pairs: 49k sentence pairs for English–Mizo, 22k for English–Khasi, and 16k for English–Tulu. For Hindi–Gondi, we utilize and curate 15k sentence pairs from Microsoft Research India's Project ELLORA dataset. We additionally release curated monolingual corpora for Mizo and Khasi to support future low-

[1] Models: NLLB SFT, mBART SFT, Gemma SFT, Gemma CPT+SFT. Dataset: mi_kh_tulu_dataset.

resource NLP research.

Second, we conduct a systematic benchmark of three model families spanning different parameter scales — mBART-large-50 (610M), NLLB-200-1.3B (1.3B), and Gemma 3 4B-IT (4B) — under three training regimes: zero-shot inference, supervised fine-tuning (SFT), and continual pre-training followed by SFT (CPT+SFT). CPT+SFT is applied only to Gemma 3, using a causal language modelling objective on monolingual corpora to expose the model to target language patterns before fine-tuning on parallel data.

2 Related Work

Mizo. Mizo has received the most NLP attention among our four languages. Early work used phrase-based SMT (Thihlum et al., 2020; Devi et al., 2022; Ralte et al., 2023). Mizo and Khasi were included in the WMT23 shared task on low-resource Indic translation (Pal et al., 2023), which released IndicNE-Corp1.0. Gaikwad et al. (2023) fine-tuned NLLB for NE Indian languages, and Sayed et al. (2024) applied NLLB-3.3B with back-translation for Mizo and Khasi. NLLB-200 (NLLB Team et al., 2022) natively supports Mizo (`lus_Latn`), making it the only language in our set with prior multilingual MT model support.

Khasi. Tham (2020) built the first POS tagger for Khasi. Laskar et al. (2021) released EnKhCorp1.0, the first parallel English–Khasi corpus with NMT baselines. P M et al. (2024) fine-tuned mBART for English–Khasi using LaBSE-based filtering of back-translated sentences, achieving chrF of 0.33. Our dataset (22,358 pairs) is the largest publicly available English–Khasi parallel corpus.

Tulu. Narayanan and Aepli (2024) presented the first parallel English–Tulu dataset built on FLORES-200 (Team, 2022) and trained an MT model using Kannada transfer learning, outperforming Google Translate by 19 BLEU points. Our work is the first to fine-tune large pre-trained multilingual models on Tulu with a substantially larger dataset (16,218 pairs).

Gondi. Microsoft Research's Project ELLORA (Bali et al., 2019) produced a Hindi–Gondi MT system. Mehta et al. (2022) report deployment-driven data collection yielding 60,000+ Hindi–Gondi pairs used for MT training. Our training data derives from this corpus. To our knowledge, our work is the first to apply large pre-trained multilingual models to Gondi MT.

Broader Indic MT. IndicTrans2 (Gala et al., 2023) covers all 22 scheduled Indian languages but excludes our four target languages. Mujadia and Sharma (2024) present BhashaVerse, supporting 36 Indian languages and all 1,296 translation directions. Both systems confirm that languages outside the constitutional schedule remain a frontier for Indian NLP.

Multilingual MT and Low-Resource Adaptation. mBART (Liu et al., 2020) demonstrated that denoising pre-training enables strong low-resource MT. Liu et al. (2021) proposed continual pre-training (CPT) on mBART for unseen languages, achieving +2.9–4 BLEU improvements — the approach we extend to Gemma 3. Bapna et al. (2022) studied adapter-based adaptation for new languages, confirming parameter-efficient methods are viable for low-resource settings.

LLMs for Low-Resource Translation. Hendy et al. (2023) showed that GPT models struggle with low-resource language pairs, and Robinson et al. (2023) found similar results for ChatGPT on African languages. Zhu et al. (2024) confirmed these findings across 102 languages. Our Gemma 3 zero-shot results (2.6–7.1 BLEU) align with this literature; our SFT results are the first for a decoder-only LLM on these four languages. .

3 Dataset Development

3.1 Parallel Corpus Construction

We constructed parallel datasets by having English sentences manually translated into Mizo, Khasi, and Tulu by native speakers. Source sentences were drawn from news articles and Wikipedia pages on tourism and sports. After collection, we applied filtering and quality checks to ensure suitability for translation. Figure 2 illustrates the full annotation pipeline.

To ensure consistency across annotators, we developed translation guidelines covering idioms, named entities, honorifics, and sentence structure.

For Gondi, we utilized the Microsoft ELLORA dataset (Bali et al., 2019), which provides Hindi–Gondi parallel data. Table 1 reports the final corpus sizes after cleaning (Section 4.4). These are the training splits used in all experiments unless otherwise noted.

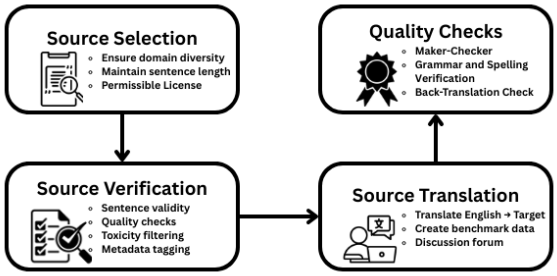

Figure 2: End-to-end annotation workflow for building parallel corpora, including sentence selection, manual translation by native speakers, and quality assurance steps.

Language Pair	Sentences
English–Mizo	49,496
English–Khasi	22,358
English–Tulu	16,218
Hindi–Gondi	15,500

Table 1: Parallel corpus statistics after data cleaning. These sizes are used in all SFT experiments. The larger combined dataset (Mizo 100k, Khasi 50k) incorporating WMT23 data (Pal et al., 2023) is used only in the bidirectional ablation (Appendix A).

3.2 Monolingual Data Collection

For Mizo and Khasi, we crawled permitted news articles from Mizoram and Meghalaya and incorporated additional data from the iHub-Data India corpus (iHub-Data). After preprocessing, we created monolingual datasets for both languages.

For Tulu, native monolingual resources are limited. We instead use Kannada monolingual data from ai4bharat/IndicCorpV2 (Doddapaneni et al., 2023), sampling 300k sentences. Since Tulu is written in Kannada script and shares Dravidian family features with Kannada, this data supports continual pretraining for Tulu.

For Gondi, we use publicly available data from Microsoft Research India's Project ELLORA, supplemented with 300k Hindi sentences sampled from ai4bharat/IndicCorpV2 (Doddapaneni et al., 2023). As our work focuses on Northern Gondi written in Devanagari script, the Hindi monolingual data facilitates target-language pretraining via script and linguistic overlap. Table 2 summarises the monolingual sizes.

3.3 Benchmark Dataset Creation

No publicly available benchmark exists for these languages. We created evaluation benchmarks from scratch. For English–Mizo, English–Khasi, and English–Tulu, we built an N-way parallel benchmark with 1,000 shared English source sentences manually translated by native speakers. Source sentences were drawn from national, regional, and international newspapers and Wikipedia (tourism, sports), and filtered by:

- Sentence length between 7 and 35 words.
- Keyword and sentiment rules to ensure domain diversity and remove inappropriate content.

For Gondi, we applied the same criteria to 1,000 sentence pairs from the ELLORA dataset and retained the corresponding Gondi translations. Table 3 reports benchmark statistics. We use 1000 sentences per direction for evaluation.

Language	Sentences
Mizo	300k
Khasi	250k
Kannada (Tulu proxy)	300k
Hindi (Gondi proxy)	300k

Table 2: Monolingual data used for Gemma CPT. Kannada is used as a proxy for Tulu; Hindi as a proxy for Gondi.

	News Articles	Wiki Pages
Sentences	500	500
Avg. length (words)	24.88	26.30
Domains	11	4

Table 3: Benchmark dataset statistics by subset.

4 Experiments

4.1 Models

We evaluate three multilingual models across different parameter scales and architectures:

- **NLLB-200-1.3B** (NLLB Team et al., 2022): encoder-decoder, 1.3B parameters. Mizo (`lus_Latn`) is natively supported; Khasi, Tulu, and Gondi require vocabulary extension.
- **mBART-large-50** (Liu et al., 2020): encoder-decoder, 610M parameters. All four target languages require vocabulary extension.
- **Gemma 3 4B-IT** (Team, 2025): decoder-only, 4.3B parameters. No vocabulary extension required; target language is specified in the instruction prompt.

4.2 Language ID Interference in Encoder-Decoder Models

Encoder-decoder models such as NLLB and mBART control the output language through a *language ID (LID) token* — a special embedding prepended to the decoder input to force generation in a specific language (Liu et al., 2020; NLLB Team et al., 2022). For languages absent from the original vocabulary, the LID token must be newly added and initialised from a proxy language before fine-tuning. When multiple new language pairs are trained together in a single joint adapter, these embeddings can remain insufficiently differentiated from their proxy initialisations, causing the decoder to generate in the proxy language rather than the intended target.

Table 4 traces our mitigation attempts for NLLB English→Khasi, the most affected direction.

Configuration	Output	BLEU	chrF++
Joint, proxy=`lus_Latn`	Mizo	3.49	22.13
Joint, proxy=`eng_Latn`	English	4.53	21.71
Joint + T=2 + full embedding training	English	4.55	21.70
Joint + multi-proxy initialisation	Mizo	6.56	25.44
Per-pair SFT	**Khasi**	**21.98**	**43.33**

Table 4: LID interference mitigation for NLLB eng→khasi. No joint training strategy produces correct Khasi output; per-pair training is necessary.

No joint strategy resolves the issue. The root cause is gradient imbalance: `kha_Latn` receives decoder updates from only 23k Khasi pairs while the encoder processes English across all 106k joint training instances. Per-pair training eliminates this competition, yielding +15.42 BLEU over the best joint configuration.

For mBART, the `en_XX` proxy for Khasi is more neutral than NLLB's `lus_Latn` (Mizo), which reduces Khasi interference but leaves mBART's newly added Mizo token susceptible to interference from Khasi. Gemma avoids this problem entirely because target language identity is specified in the instruction prompt rather than through a learned LID token. This is a practical advantage for deployment: a single Gemma adapter handles all four languages jointly without any per-language workaround.

4.3 Vocabulary Extension

For NLLB and mBART, we extend the tokenizer with new language tokens for each unsupported language and initialise their embeddings before fine-tuning. A single-proxy initialisation — where the new embedding is copied directly from one related language — caused LID interference (Section 4.2) because the new token remained too close to its proxy throughout joint training. We instead use multi-proxy averaging: the new embedding is initialised as the mean of embeddings from several related languages, plus Gaussian noise (σ=0.05) to break symmetry. This produces a more neutral starting point that is equidistant from all proxies, giving the new token more room to differentiate during fine-tuning. Table 5 lists the proxies used for each language and model.

Language	NLLB	mBART	Rationale
Khasi	`eng,fra_Latn`	`en,fr_XX`	Latin script, no Mizo bias
Tulu	`kan_Knda`	`kn_IN`	Kannada script, Dravidian
Gondi	`hin,mar, npi_Deva`	`hi,mr_IN`	Devanagari, dilutes Hindi
Mizo	native in NLLB	`en,fr_XX`	New token in mBART only

Table 5: Proxy languages used for multi-proxy embedding initialisation of new language ID tokens. Mizo is natively supported by NLLB and requires no initialisation; it is a new token only in mBART.

4.4 Data Cleaning

Before training, all parallel data is filtered using three steps: (1) Unicode normalisation (NFKC), which standardises encoding variants, curly quotes, and em-dashes; (2) script-purity filtering, which discards any sentence pair where an alphabetic character falls outside the expected Unicode block for its language (e.g. Kannada block for Tulu, Devanagari for Gondi, Latin for Mizo and Khasi); (3) minimum length filtering, which removes pairs with fewer than two whitespace-separated tokens on either side. Post-filtering corpus sizes are reported in Table 1.

4.5 Training

We evaluate three training regimes across all models.

Zero-shot inference. Models are evaluated directly on the test sets without any fine-tuning. For NLLB-200, only Mizo is evaluated zero-shot because it is the only natively supported language; Khasi, Tulu, and Gondi are absent from the vo-

cabulary and produce meaningless output. For Gemma 3, we evaluate both zero-shot (no examples) and 3-shot prompting, where three in-context translation examples are drawn from the training set and prepended to the prompt.

Supervised fine-tuning (SFT). We train LoRA adapters (Hu et al., 2022) on all three models using the parallel corpora from Table 1. To balance the training distribution across languages of different sizes, we use temperature-based upsampling (T=2): $p_i \propto n_i^{1/T}$, which smooths the distribution toward uniform and effectively upsamples Khasi ($\times$2) and Gondi ($\times$3) relative to Mizo. For NLLB and mBART, we set `modules_to_save=["embed_tokens","lm_head"]` to train the full embedding matrix in addition to the LoRA parameters, which is necessary for the new language ID tokens to differentiate. Due to LID interference (Section 4.2), we train a dedicated per-pair adapter for NLLB Khasi; Mizo, Tulu, and Gondi use the joint adapter. All four mBART pairs are trained together as a single joint adapter.

For Gemma 3, we fine-tune using the following instruction prompt, which specifies both the source and target language names and scripts. Labels are masked for the instruction prefix so the model learns only to generate the translation token.

Instruction Prompt Template

```
Translate the following
    {src_language} ({src_script})
sentence into {tgt_language}
    ({tgt_script}).
Output only the translated
    sentence.
Do not include explanations.

{source_sentence}<end_of_turn>
```

Continual pre-training followed by SFT (CPT+SFT). CPT+SFT is applied only to Gemma 3. In this regime, we first continue pre-training on the monolingual corpora from Table 2 using a causal language modelling (CLM) objective. This exposes the model to target language vocabulary and patterns before any parallel translation data is introduced, which is particularly important for Mizo and Khasi, where Gemma's base perplexity exceeds 800 (Section 5.3). After CPT converges, the CPT LoRA adapter is merged into the base model weights, and a fresh SFT LoRA is then trained on the parallel data. Merging before SFT ensures the improved language representations are available to the SFT stage rather than being isolated in a separate adapter.

4.6 Hyperparameters

All experiments use AdamW optimiser with cosine learning rate decay, 200 warmup steps, and early stopping with patience 5 evaluated every 500 steps. Effective batch size is 64 across all models (2×A100 40 GB GPUs via `torchrun`). Model-specific hyperparameters are given in Table 6.

Hyperparameter	NLLB	mBART	Gemma
LoRA rank r	64	64	64
LoRA α	128	128	128
LoRA dropout	0.1	0.1	0.05
Learning rate	3e-4	3e-4	2e-4
Max steps	10,000	10,000	10,000
Effective batch	64	64	64
CPT learning rate	—	—	1e-4
CPT max steps	—	—	5,000

Table 6: Training hyperparameters. NLLB and mBART use a higher learning rate than Gemma because they are encoder-decoder models with smaller parameter counts. Gemma CPT uses a lower learning rate than SFT to avoid catastrophic forgetting of general language knowledge.

4.7 Evaluation

All systems are evaluated on **1,000 sentence pairs** per direction from the benchmark datasets described in Section 3.3. We report three automatic metrics:

- **BLEU** (Papineni et al., 2002): corpus-level, computed with `sacrebleu` (`effective_order=True`). BLEU measures n-gram precision against the reference and is widely used for MT evaluation, but can underestimate quality for morphologically rich languages with low word overlap.
- **chrF++** (Popović, 2015): character n-gram F-score with word-order penalty (`word_order=2`). chrF++ is more robust than BLEU for agglutinative languages like Gondi and Tulu, where a single root may appear in many surface forms.
- **COMET** (Rei et al., 2022): neural MT evaluation using `wmt22-comet-da`, which scores translations using a multilingual reference model (XLM-R). COMET correlates better with human judgements than BLEU for high-resource languages, but its reliability is limited for our target

languages due to sparse XLM-R coverage. We report COMET for completeness and treat BLEU and chrF++ as primary metrics.

5 Results

5.1 Tokenizer Fertility

Table 7 reports tokenizer fertility (subword tokens per whitespace word) for all six languages across the three models. Lower fertility indicates better pre-existing vocabulary coverage.

Language	NLLB-1.3B	mBART-50	Gemma-3-4B
English	1.49	1.52	**1.36**
Hindi	1.50	1.56	**1.44**
Mizo	**1.55**	1.89	1.80
Khasi	**1.59**	1.71	1.71
Gondi	**2.01**	2.14	2.00
Tulu	3.32	**3.13**	3.69

Table 7: Tokenizer fertility (tokens per word). Bold = lowest fragmentation per language. Tulu's high fertility across all models reflects the absence of native Kannada-script coverage at training scale.

NLLB-200 has the lowest fertility for Mizo and Khasi, consistent with its native Latin-script NE Indian language support. Gemma achieves the best English and Hindi coverage, reflecting its large-scale general pretraining. Tulu (Kannada script) is highly fragmented across all three tokenizers (3.13–3.69), explaining why CPT on Kannada monolingual data is beneficial for this language. Gondi (Devanagari) is moderately fragmented (2.0–2.1) due to shared script with Hindi, which all models support well.

5.2 Zero-Shot and Few-Shot Baselines

Table 8 reports zero-shot and few-shot baselines. NLLB-200 achieves 20.70 BLEU on Mizo (its only native language) but produces no valid output for the remaining three languages, which are absent from its vocabulary. mBART-50 has no prior support for any of our four languages and is excluded from zero-shot evaluation.

Gemma 3 4B-IT produces non-trivial output across all four languages (2.63–7.06 BLEU) without fine-tuning. Three-shot prompting improves consistently by +0.37–+4.40 BLEU. The high COMET scores relative to BLEU (52–64 vs. 2.6–7.1 BLEU) suggest semantically reasonable output that diverges in surface form from the single reference.

System	Direction	BLEU	chrF++	COMET
NLLB-200-1.3B (zero-shot)				
	eng→mizo	20.70	47.16	68.95
	eng→{khasi,tulu,gondi}	*N/A — unsupported*†		
Gemma 3 4B-IT (zero-shot)				
	eng→mizo	7.06	26.82	55.86
	eng→khasi	3.31	19.80	53.04
	eng→tulu	3.98	23.51	63.59
	hi→gondi	2.63	19.68	51.91
Gemma 3 4B-IT (3-shot)				
	eng→mizo	11.46	35.79	63.35
	eng→khasi	5.21	25.25	55.71
	eng→tulu	4.91	26.60	62.61
	hi→gondi	3.00	20.03	52.77

Table 8: Zero-shot and 3-shot baselines. † NLLB zero-shot inference for unsupported languages generates the proxy initialisation language rather than the target, yielding meaningless scores.

5.3 Supervised Fine-Tuning

Table 9 reports SFT and CPT+SFT results for all three models.

NLLB-200 achieves the best Tulu result (43.72 BLEU, 69.48 chrF++), benefiting from its Kannada-script proximity and single-proxy initialisation for `tcy_Knda`. It also leads on Mizo (34.49 BLEU), leveraging native language support. Its joint adapter fails for Khasi (6.56 BLEU) due to language ID interference; a dedicated per-pair adapter recovers to 21.98 BLEU.

mBART-50 achieves reasonable Khasi performance in joint training (19.69 BLEU), outperforming NLLB joint (6.56). This is because mBART's `en_XX` proxy for Khasi is more neutral than NLLB's `lus_Latn` (Mizo), reducing interference. mBART scores lowest on Mizo (11.71 BLEU), where Mizo is also a new token subject to interference from Khasi during joint training.

Gemma SFT achieves the best Khasi result across all systems in joint training (33.80 BLEU), and strong Mizo (33.06) and Tulu (36.34) scores. As a decoder-only model, Gemma specifies the target language via the instruction prompt rather than a learned language ID token, making it immune to the interference that affects NLLB and mBART.

Gemma CPT+SFT yields marginal improvements on Khasi (+0.37 BLEU) and Tulu (+0.06 chrF++) over Gemma SFT alone, with slight decreases on Mizo (−1.27) and Gondi (−0.43). Table 10 shows that CPT substantially reduces held-

System	Adapter	eng→mizo		eng→khasi		eng→tulu		hi→gondi	
		BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
NLLB-1.3B SFT	Joint	34.49	59.94	6.56*	25.44	**43.72**	**69.48**	8.52	32.34
NLLB-1.3B SFT	Per-pair	—	—	21.98	43.33	—	—	—	—
mBART-50 SFT	Joint	11.71	30.57	19.69	38.93	27.97	54.61	7.68	30.10
Gemma SFT	Joint	33.06	58.20	33.80	58.55	36.34	63.19	7.84	29.79
Gemma CPT+SFT	Joint	31.79	57.26	**34.17**	**58.68**	35.20	63.28	7.41	28.94

Table 9: Fine-tuning results on 1,000-sentence test sets. Bold = best per direction. *NLLB joint adapter for Khasi suffers language ID interference (Section 4.2); per-pair adapter gives 21.98 BLEU. Gondi BLEU underestimates quality across all models — see Section 5.5.

out perplexity for all four languages, confirming genuine language adaptation. The largest reductions are for Mizo (836→28) and Khasi (853→14), where Gemma's base model had near-random language models due to minimal pretraining coverage. Hindi shows the smallest drop (83→28), consistent with Gemma already having reasonable Hindi representations. Despite successful language adaptation, the downstream SFT gains are modest, suggesting that task-specific translation patterns are learned primarily during SFT rather than CPT.

Language	Base PPL	CPT PPL	Δ
Mizo	836.34	28.58	−807.76
Khasi	853.66	14.87	−838.79
Kannada	273.86	30.53	−243.33
Hindi	83.40	28.44	−54.96

Table 10: Held-out perplexity before and after Gemma CPT on monolingual data. Lower is better. Kannada is the proxy for Tulu; Hindi is the proxy for Gondi.

5.4 Language ID Interference

The results confirm that LID interference is the main failure mode for NLLB Khasi in joint training: the joint adapter reaches only 6.56 BLEU, while the dedicated per-pair adapter recovers to 21.98 BLEU (Table 4). Gemma avoids this failure mode because target language control is prompt-based rather than tied to a newly added decoder LID embedding.

5.5 Analysis

Gondi metric gap. Gondi BLEU (7.41–8.52 across systems) underestimates translation quality. Manual inspection confirms genuine Gondi morphology in all system outputs, e.g., (Gondi "door") where the reference uses the Hindi loanword , and correct agglutinative verb forms. chrF++ (28–32) is more informative here.

COMET reliability. COMET (`wmt22-comet-da`) is built on XLM-R, which has limited coverage of our target languages. Scores for English/Hindi→XX directions should be treated as approximate. For XX→English directions (available in the ablation), COMET is more reliable. We report COMET for completeness but treat BLEU and chrF++ as primary metrics.

6 Conclusion

We present parallel corpora, monolingual corpora, and benchmark evaluation sets for four extremely low-resource Indian languages — Mizo, Khasi, Tulu, and Gondi — and benchmark three multilingual models (NLLB-200-1.3B, mBART-large-50, Gemma 3 4B-IT) under zero-shot, SFT, and CPT+SFT training regimes.

Our main findings are: (1) Fine-tuning substantially improves translation quality for all four languages, with NLLB reaching 43.72 BLEU for English–Tulu and Gemma achieving 34.17 BLEU for English–Khasi. (2) Encoder-decoder models trained with newly added language ID tokens suffer from LID interference in joint training, requiring per-language adapters for reliable output. (3) Gemma 3 4B-IT, as a decoder-only model with prompt-based language specification, handles all four languages in a single adapter without interference, making it preferable for deployment when simplicity matters. (4) CPT reduces held-out perplexity dramatically for Mizo (836→28) and Khasi (853→14), confirming genuine language adaptation, though downstream translation gains are modest, suggesting translation mapping is learned primarily during SFT. All datasets and model adapters are publicly released to support future low-resource NLP research for these languages.

Future work. Expanding parallel data for Gondi (currently 15k pairs), collecting multi-reference evaluation sets to better assess morphologically

rich output, and extending to additional Northeast Indian and tribal languages.

Acknowledgments

We thank all contributors and acknowledge the use of AI for grammar checking.

Ethics Statement

This work supports linguistic inclusivity by developing translation resources for Mizo, Khasi, Tulu, and Gondi four extremely low-resource Indian languages. All data was collected in collaboration with native speakers and language experts, with fair compensation for human annotators. Text sources were either publicly available or used with permission, and no personally identifiable information was included.

We ensured culturally respectful and ethical data handling throughout the project. While efforts were made to reduce bias, limitations persist due to the scarcity of digital resources. We encourage future community-driven contributions to strengthen NLP for underrepresented languages.

Limitations

The primary limitation of this work lies in the scale of available parallel data. While our curated datasets enable meaningful fine-tuning and evaluation, the quantity is still insufficient for training production-level translation systems. As a result, the model may struggle with domain generalization and long-range dependencies, especially in complex or out-of-distribution inputs. Furthermore, certain language pairs (e.g., Hindi ↔ Gondi) suffer from extremely sparse alignment, which impacts translation quality.

References

Kalika Bali and 1 others. 2019. ELLORA: Enabling low resource languages with technology. In *Proceedings of the 1st Workshop on Language Technologies for All (LT4All)*.

Ankur Bapna and 1 others. 2022. Building machine translation systems for the next thousand languages.

Census Commissioner. 2022. Census of India 2011 - Language Atlas. [Accessed 05-06-2024].

C. S. Devi, B. S. Purkayastha, and L. S. Meetei. 2022. An empirical study on English–Mizo statistical machine translation with Bible corpus. *International Journal of Electrical and Computer Engineering Systems*, 13(9):759–765.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Pranav Gaikwad, Meet Doshi, Sourabh Deoghare, and Pushpak Bhattacharyya. 2023. Machine translation advancements for low-resource Indian languages in WMT23: CFILT-IITB's effort for bridging the gap. In *Proceedings of the Eighth Conference on Machine Translation*, pages 950–953. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *Transactions on Machine Learning Research*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation.

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*.

iHub-Data. India data commons. `https://ihub-data.iiit.ac.in/`. Accessed: 2025-05-19.

Indian-Constitution. 2022. Languages included in the eighth schedule of the indian constitution. [Accessed: 2 Mar. 2025].

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95, Virtual. Association for Machine Translation in the Americas.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. volume 8, pages 726–742.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718. Association for Computational Linguistics.

Devansh Mehta, Harshita Diddee, Sebastin Santy, Ramaravind K. Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Vishnu Prasad, Anurag Shukla, and Kalika Bali. 2022. Learnings from technological interventions in a low resource language: Enhancing information access in Gondi. *Language Resources and Evaluation*.

Vandan Mujadia and Dipti Misra Sharma. 2024. Bhashaverse: Translation ecosystem for indian subcontinent languages. *arXiv preprint arXiv:2412.04351*.

Manu Narayanan and Noëmi Aepli. 2024. A Tulu resource for machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhinav P M, Ketaki Shetye, and Parameswari Krishnamurthy. 2024. MTNLP-IIITH: Machine translation for low-resource Indic languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 751–755, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

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*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Lalnithanga Ralte and 1 others. 2023. Parts of speech tagged phrase-based statistical machine translation system for English → Mizo language. *SN Computer Science*.

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.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt MT: Competitive for high- (But Not Low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418. Association for Computational Linguistics.

Hamees Sayed, Advait Joglekar, and Srinivasan Umesh. 2024. SPRING Lab IITM's submission to low resource Indic language translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*.

Gemma Team. 2025. Gemma 3.

NLLB Team. 2022. The FLORES-200 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2207.04672*. Part of NLLB.

Medari Tham. 2020. NLP tools for Khasi, a low resource language. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): System Demonstrations*, pages 26–27.

Zaitinthanga Thihlum and 1 others. 2020. Machine translation of English language to Mizo language. In *Proceedings of the International Conference on Emerging Trends in Information Technology and Engineering*. ResearchGate preprint.

Wenhao Zhu, Hongyi Liu, Qiwei Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.

A Data Scale and Bidirectional Ablation

System	en→kha	kha→en	en→mz	mz→en
Joint SFT (4-lang, fwd)	6.56	—	34.49	—
2-lang + full data + bidir	**34.75**	**39.06**	**33.87**	**38.12**

Table 11: BLEU scores for the 2-language bidirectional ablation (NLLB). Full data = Mizo 100k + Khasi 50k, including WMT23 (Pal et al., 2023).

Expanding Khasi to 50k pairs, training bidirectionally, and restricting to two languages together yield +28.19 BLEU for Khasi. Mizo is unchanged (−0.62), confirming the gain is specific to the new language ID token. Due to compute constraints, we do not extend this setting to all four languages.