NE-BERT: A Multilingual Language Model for Nine Northeast Indian Languages

Badal Nyalang
MWire Labs
Shillong, Meghalaya, India
nyalang@mwirelabs.com

Abstract

Large pretrained language models have demonstrated remarkable capabilities across diverse languages, yet critically underrepresented low-resource languages remain marginalized. We present **NE-BERT**, a domain-specific multilingual encoder model trained on approximately 8.3 million sentences spanning 9 Northeast Indian languages and 2 anchor languages (Hindi, English), a linguistically diverse region with minimal representation in existing multilingual models. By employing weighted data sampling and a custom SentencePiece Unigram tokenizer, NE-BERT outperforms IndicBERT-V2 and MuRIL across all 9 Northeast Indian languages, achieving $15.97\times$ and $7.64\times$ lower average perplexity respectively, with $1.50\times$ better tokenization fertility than mBERT. We address critical vocabulary fragmentation issues in extremely low-resource languages such as Pnar (1,002 sentences) and Kokborok (2,463 sentences) through aggressive upsampling strategies. Downstream evaluation on part-of-speech tagging validates practical utility on three Northeast Indian languages. We release NE-BERT, test sets, and training corpus under CC-BY-4.0 to support NLP research and digital inclusion for Northeast Indian communities.

1 Introduction

The performance disparity between high-resource and low-resource languages in modern NLP systems reflects and reinforces existing digital inequities (Joshi et al., 2020). While multilingual models like mBERT (Devlin et al., 2019) provide broad language coverage, they perform poorly on languages with limited web presence and complex morphological structures (Lauscher et al., 2020). This gap is particularly pronounced for the indigenous languages of Northeast India, a region home to over 200 distinct languages (Moseley, 2010) yet largely absent from mainstream NLP research.

Northeast Indian languages present unique challenges: extreme resource scarcity (some with fewer than 1,000 digitized sentences), agglutinative morphology, script diversity (Latin, Bengali-Assamese), and limited standardization. Existing regional efforts like IndicBERT (Kakwani et al., 2020) focus primarily on scheduled Indian languages with substantial corpora, leaving languages such as Khasi, Garo, Pnar, Mizo, and Kokborok critically underserved.

We introduce **NE-BERT**, a ModernBERT-based (Warner et al., 2025) encoder model specifically designed for Northeast Indian languages. Our contributions include:

- A curated multilingual corpus of 8.3M sentences covering 9 indigenous Northeast Indian languages (Assamese, Garo, Khasi, Meitei, Mizo, Naga, Nyishi, Pnar, Kokborok) plus 2 anchor languages (Hindi, English) with strategic weighted sampling (Xue et al., 2021).
- A custom 50,368-token SentencePiece Unigram tokenizer optimized for morphologically rich and agglutinative languages (Kudo and Richardson, 2018), achieving $1.50\times$ better average tokenization efficiency than mBERT.
- Comprehensive evaluation on all 9 Northeast Indian languages demonstrating NE-BERT outperforms IndicBERT-V2 and MuRIL, with particularly strong gains (7–$15\times$) on ultra-low-resource languages like Pnar, Kokborok, and Nyishi.

2 Related Work

2.1 Multilingual Language Models

Early multilingual models like mBERT (Devlin et al., 2019) demonstrated cross-lingual transfer capabilities but suffered from the "curse of multilinguality" (Conneau et al., 2020), performance

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 1–12

degradation as language count increases. XLM-RoBERTa (Conneau et al., 2020) addressed this through larger training corpora (2.5TB) but still exhibited vocabulary fragmentation for low-resource languages. Recent work on language-specific adaptations (Rust et al., 2021) and targeted continued pretraining (Chau and Lin, 2020) shows promising results for bridging this gap.

2.2 Regional Language Models

Several regional initiatives have emerged to address local language needs. IndicBERT (Kakwani et al., 2020) covers 12 scheduled Indian languages with 9B tokens, achieving strong performance on Indo-Aryan and Dravidian languages but with limited coverage of Northeast Indian languages. Similar efforts for African languages (Ogueji et al., 2021; Alabi et al., 2022) demonstrate the viability of region-specific models. However, these approaches typically focus on languages with substantial existing corpora (>1M sentences), leaving ultra-low-resource languages unaddressed.

2.3 Tokenization for Low-Resource Languages

Tokenizer design critically impacts low-resource language performance (Ács, 2021). Byte-Pair Encoding (BPE) (Sennrich et al., 2016), while popular, can fragment morphologically rich words into suboptimal units. SentencePiece Unigram (Kudo and Richardson, 2018) preserves linguistic structures better for agglutinative languages. Weighted sampling during tokenizer training (Lample and Conneau, 2019) helps balance vocabulary allocation across languages with disparate corpus sizes, critical for our extremely imbalanced dataset.

3 Dataset Construction

3.1 Language Selection and Sources

We curate data for 9 Northeast Indian languages plus 2 anchor languages (Table 1). The Northeast Indian languages span three major language families: Sino-Tibetan (Meitei, Mizo, Garo, Kokborok, Nyishi, Naga), Austroasiatic (Khasi, Pnar), and Indo-Aryan (Assamese). We include Hindi and English as anchor languages to facilitate cross-lingual transfer (Artetxe et al., 2020), particularly for tasks requiring code-switching support.

Data sources include:

- **Curated Corpora**: Meitei, Assamese, Mizo, Khasi, Garo, Pnar, and Naga datasets compiled from government documents, news archives, educational materials, and cultural texts.
- **WMT 2025 Shared Task** (WMT 2025 Organizers, 2025): Nyishi and Kokborok parallel corpora from the Workshop on Machine Translation low-resource language track.
- **Public Datasets**: Hindi from verified Hugging Face datasets; English from standard corpora.

3.2 Data Preprocessing

We apply a rigorous cleaning pipeline to ensure data quality:

1. **Length Filtering**: Remove sentences with character length < 20 to eliminate noise, incomplete fragments, and non-linguistic content.
2. **Unicode Normalization**: Apply NFKC normalization (The Unicode Consortium, 2021) to handle script variations, diacritical marks, and ensure consistency across diverse sources.
3. **Whitespace Condensation**: Collapse multiple spaces and normalize line breaks to standardize formatting.

We split data into 99.5% training and 0.5% validation sets (random seed 42). A separate held-out test set is used for final evaluation (Section 7).

Data Availability: We publicly release our training corpus at `Badnyal/ne-multilingual-corpus` and evaluation test sets at `MWirelabs/northeast-languages-test-set` on Hugging Face under CC-BY-4.0 license to support reproducibility and future research on Northeast Indian languages.

3.3 Weighted Sampling Strategy

Following Xue et al. (2021), we implement aggressive weighted sampling to address extreme resource imbalance. Table 1 shows our weighting scheme: ultra-low-resource languages (Pnar with 1,002 sentences, Kokborok with 2,463 sentences) receive $100\times$ upsampling during tokenizer training to ensure adequate vocabulary representation. This prevents vocabulary starvation where rare languages get fragmented into character-level tokens

Language	ISO	Sentences	Tokens	Virtual Count	Weight	Family	Source
Anchor Languages							
Hindi	hin	3,404,007	—	170,200	0.05×	Indo-Aryan	HF Datasets
English	eng	500,000	—	100,000	0.2×	Germanic	HF Datasets
Northeast Indian Languages							
Meitei	mni	1,354,323	42,504,181	1,354,323	1.0×	Sino-Tibetan	Curated
Assamese	asm	1,000,000	38,652,391	1,000,000	1.0×	Indo-Aryan	Curated
Khasi	kha	1,000,000	17,472,606	1,000,000	1.0×	Austroasiatic	Curated
Mizo	lus	1,000,000	26,774,164	1,000,000	1.0×	Sino-Tibetan	Curated
Nyishi	njz	55,870	560,374	1,117,400	20.0×	Sino-Tibetan	WMT 2025
Naga	nag	13,918	508,980	278,360	20.0×	Sino-Tibetan	Curated
Garo	grt	10,817	243,251	216,340	20.0×	Sino-Tibetan	Curated
Kokborok	trp	2,463	89,851	246,300	100.0×	Sino-Tibetan	WMT 2025
Pnar	pbv	1,002	52,144	100,200	100.0×	Austroasiatic	Curated
NE Total		**4,438,393**	**126,857,942**				
Overall Total		**8,342,400**	**—**	**6,583,123**			

Table 1: Corpus statistics showing sentence counts, token counts for NE languages, virtual counts after weighted sampling for tokenizer training, language families, and data sources.

(Rust et al., 2021), which would severely degrade inference efficiency and model performance.

The virtual counts in Table 1 apply only to tokenizer training; actual MLM training uses raw sentence counts to avoid overfitting on limited data. Anchor languages are downweighted (Hindi $0.05\times$, English $0.2\times$) to prioritize Northeast language vocabulary while maintaining cross-lingual transfer capabilities.

4 Tokenization

4.1 Algorithm Selection

We adopt SentencePiece Unigram (Kudo and Richardson, 2018) over the more common Byte-Pair Encoding (BPE) for two primary reasons:

1. **Linguistic Preservation**: Unigram's probabilistic approach reduces harmful subword fragmentation in morphologically rich and agglutinative languages (Kokborok, Garo, Meitei) compared to BPE's greedy merging strategy. This is critical for languages where single words can encode complex grammatical information.

2. **Vocabulary Efficiency**: Unigram naturally balances frequent subword allocation across languages without explicit vocabulary partitioning, allowing our weighted sampling strategy to directly influence token boundaries.

4.2 Tokenizer Configuration

Our tokenizer uses the following configuration:

- **Vocabulary Size**: 50,368 tokens (nearest 128-multiple for efficient Tensor Core execution on modern GPUs)
- **Character Coverage**: 1.0 (full Unicode range to handle all scripts)
- **Maximum Piece Length**: 16 characters
- **Shrinking Factor**: 0.75
- **Sub-iterations**: 2
- **Special Tokens**: `<cls>` (0), `<pad>` (1), `<eos>` (2), `<unk>` (3), `<mask>` (4)

Training on weighted virtual counts (Table 1) ensures that common words in Pnar and Kokborok form single tokens rather than fragmenting into multi-token sequences. This dramatically reduces inference costs and improves semantic coherence for ultra-low-resource languages.

5 Model Architecture

We adopt ModernBERT-base (Warner et al., 2025) as our foundation due to its architectural improvements over classical BERT:

5.1 Architecture Details

- **Encoder Layers**: 22 transformer layers
- **Hidden Dimension**: 768
- **Attention Heads**: 12
- **Total Parameters**: 149M
  - Embedding layer: 38.7M parameters

  - Encoder layers: 110.3M parameters

- **Positional Encoding**: Rotary Position Embeddings (RoPE) with $\theta_{\text{global}} = 160{,}000$, $\theta_{\text{local}} = 10{,}000$ (Su et al., 2024)
- **Attention Mechanism**: Flash Attention 2 (Dao, 2023) for memory efficiency
- **Optimization**: Unpadding enabled for approximately 30% throughput improvement during training

ModernBERT's design enables efficient training on longer contexts while maintaining competitive parameter counts relative to BERT-base (110M) and IndicBERT (66M). The RoPE positional encodings provide better length extrapolation than learned position embeddings, which is beneficial for languages with variable word lengths.

Model Size Justification: We adopt the 149M parameter configuration as an optimal balance between capability and computational efficiency. This size is comparable to mBERT (110M) while being substantially smaller than IndicBERT-V2 (237M) and MuRIL (236M), enabling cost-effective training ($7.31 on a single A40 GPU) and efficient deployment. Our results demonstrate that appropriate tokenization and targeted training data are more critical than raw parameter count for ultra-low-resource language performance.

6 Training

6.1 Training Configuration

We train NE-BERT using masked language modeling (MLM) with 15% masking probability (Devlin et al., 2019). We employ dynamic masking where each epoch sees different masked positions, improving generalization compared to static masking (Liu et al., 2019).

Hyperparameters:

- **Batch size**: 32 per device with 32 gradient accumulation steps (effective batch size 1,024)
- **Learning rate**: 5×10^{-4} with cosine decay schedule
- **Warmup steps**: 1,500
- **Weight decay**: 0.01
- **Training epochs**: 10
- **Precision**: Mixed FP16 with TF32 enabled
- **Optimizer**: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)

6.2 Compute Infrastructure

Training was conducted on a single NVIDIA A40 GPU (48GB VRAM) for approximately 17 hours, with a total compute cost of $7.31. This demonstrates the cost-effectiveness of our approach for resource-constrained research settings. We use PyTorch 2.4+ with Hugging Face Transformers 4.48+ and Flash Attention 2.x.

6.3 Training Dynamics

Training loss decreased smoothly from approximately 10.0 at initialization to 1.62 (training) and 1.64 (validation) at convergence over 10 epochs. The close tracking between training and validation loss indicates no overfitting despite the small corpus size for some languages. This suggests our weighted sampling strategy and data augmentation through dynamic masking effectively prevent memorization.

7 Evaluation

7.1 Evaluation Protocol

We evaluate using perplexity (PPL) on a held-out test set of 500 sentences per language. Perplexity is computed as:

$$\text{PPL} = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{MLM}}(x_i)\right) \quad (1)$$

where $\mathcal{L}_{\text{MLM}}$ is the masked language modeling loss and N is the number of test examples. Lower perplexity indicates better predictive performance.

We also measure tokenization fertility, the average number of subword tokens per word, to assess vocabulary efficiency (Rust et al., 2021). Lower fertility indicates more efficient tokenization, reducing inference costs and improving semantic coherence.

Test Set Construction: We construct held-out test sets of 500 sentences per language through careful deduplication against our training corpus (`Badnyal/ne-multilingual-corpus`). Test sentences are extracted from newer data sources not present in the training set and filtered to ensure minimum length of 15 characters. This provides statistically robust perplexity evaluation while avoiding data leakage. We release test sets publicly at `MWirelabs/northeast-languages-test-set`.

Evaluation Coverage: All 9 Northeast Indian languages included in training are evaluated using the constructed test sets. While Nyishi and Naga test sets are smaller (extracted from WMT 2025 parallel corpora), they provide initial validation of model performance across the complete language coverage.

7.2 Baselines

We compare against three widely-used multilingual models:

- **IndicBERT-V2** (Doddapaneni et al., 2023): 237M parameter encoder trained on 22 Indic languages with 120B tokens. Enhanced version with expanded language coverage and improved architecture.
- **MuRIL** (Khanuja et al., 2021): 236M parameter encoder optimized for Indian languages with 16B tokens. Google's multilingual model for Indic NLP.
- **mBERT** (Devlin et al., 2019): 110M parameter encoder covering 104 languages with large-scale Wikipedia data. Serves as a general-purpose multilingual baseline.

Table 2 compares architectural details across all evaluated models.

Model	Params	Layers	Hidden	Heads
NE-BERT	149M	22	768	12
mBERT	110M	12	768	12
IndicBERT-V2	237M	12	1024	16
MuRIL	236M	24	1024	16

Table 2: Model architecture comparison showing parameter counts, layer depth, hidden dimension, and attention heads.

7.3 Results

Table 3 presents per-language perplexity scores across all 9 Northeast Indian languages plus 2 anchor languages. NE-BERT achieves the lowest average perplexity across Northeast Indian languages (2.21) compared to IndicBERT-V2 (35.29), MuRIL (16.88), and mBERT (2.77). Performance patterns vary by language resource level: NE-BERT achieves superior results on ultra-low-resource languages (Pnar, Kokborok, Garo, Nyishi) where IndicBERT-V2 and MuRIL exhibit catastrophic performance degradation (perplexity >60 for Pnar and Nyishi). On higher-resource languages with extensive Wikipedia coverage (Assamese, Meitei), mBERT maintains competitive performance due to its massive pretraining corpus.

Language	NE-BERT	IB-V2	MuRIL	mBERT
Assamese	**1.76**	9.01	5.62	1.65
Meitei	1.89	3.77	3.22	**1.44**
Khasi	**1.27**	2.40	1.93	1.36
Mizo	**1.90**	8.95	7.00	2.44
Garo	**2.64**	26.37	15.88	3.64
Kokborok	**1.72**	9.15	5.41	2.23
Pnar	**2.92**	66.92	39.65	5.30
Naga	**1.49**	3.83	3.39	1.81
Nyishi	**4.33**	187.20	75.80	6.05
English	**1.55**	14.81	5.57	2.64
Hindi	**1.43**	10.08	5.75	1.79
NE Avg.	**2.21**	35.29	16.88	2.77
Overall	**2.17**	31.14	15.38	2.76

Table 3: Perplexity on 500-sentence test sets. NE-BERT achieves lowest average across all 9 NE languages.

Table 4 shows tokenization fertility scores. NE-BERT's custom tokenizer achieves significantly lower average fertility than all baselines across Northeast Indian languages (1.68 avg. vs. 2.08 for IndicBERT-V2, 2.14 for MuRIL, and 2.51 for mBERT). Table 5 presents bits per character (BPC), an alternative compression metric. NE-BERT achieves 0.347 average BPC compared to 1.497 for IndicBERT-V2, 1.271 for MuRIL, and 0.590 for mBERT, demonstrating superior encoding efficiency.

7.4 Analysis

The performance differences between NE-BERT and baselines reveal several key insights:

Domain-Specific Training Advantage: NE-BERT's consistent superiority over IndicBERT-V2 and MuRIL (average perplexity improvements of $15.97\times$ and $7.64\times$ respectively) validates our hypothesis that domain-specific models with appropriate tokenization outperform general regional models. IndicBERT-V2's corpus focuses heavily on scheduled languages (Hindi, Bengali, Tamil, Telugu) with minimal Northeast representation, resulting in catastrophic failures on ultra-low-resource languages (187.20 PPL on Nyishi, 66.92 on Pnar). MuRIL, despite similar parameter count (236M), shows severe degradation on non-scheduled Northeast languages, highlighting the importance of training data composition over model size alone.

Vocabulary Optimization: The tokenization fertility results (Table 4) and BPC scores (Table 5) demonstrate the effectiveness of weighted Unigram

Language	NE-B	IB-V2	MuRIL	mB
Assamese	1.63	**1.61**	1.72	3.80
Meitei	**1.52**	2.77	3.01	4.17
Khasi	**1.31**	1.77	1.78	1.78
Mizo	**1.36**	1.73	1.79	1.78
Garo	**2.37**	2.84	2.80	2.89
Kokborok	**2.07**	2.07	2.22	2.19
Pnar	**1.61**	1.69	1.69	1.67
Naga	**1.56**	1.93	1.77	1.97
Nyishi	**1.67**	2.34	2.44	2.36
NE Avg.	**1.68**	2.08	2.14	2.51

Table 4: Tokenization fertility (tokens per word) across all 9 NE languages. Lower values indicate more efficient tokenization. NE-BERT achieves lowest average fertility across NE languages.

Language	NE-B	IB-V2	MuR	mB
Assamese	**0.230**	0.880	0.731	0.442
Meitei	**0.220**	0.815	0.778	0.331
Khasi	**0.092**	0.445	0.336	0.158
Mizo	**0.272**	1.163	1.063	0.486
Garo	**0.503**	1.997	1.663	0.800
Kokborok	**0.277**	1.136	0.926	0.433
Pnar	**0.680**	2.795	2.446	1.092
Naga	**0.171**	0.709	0.594	0.321
Nyishi	**0.789**	3.763	3.231	1.306
English	**0.190**	0.862	0.558	0.328
Hindi	**0.192**	0.863	0.659	0.345
NE Avg.	**0.359**	1.545	1.307	0.597
Overall	**0.347**	1.497	1.271	0.590

Table 5: Bits per character (BPC) across all 9 NE languages. Lower is better. NE-BERT achieves lowest average BPC.

sampling. NE-BERT achieves 1.68 average tokens/word on Northeast Indian languages versus IndicBERT-V2's 2.08 and MuRIL's 2.14, representing 19–21% reduction in sequence length. The BPC results provide complementary evidence: NE-BERT's 0.347 average BPC demonstrates superior compression compared to IndicBERT-V2 (1.497) and MuRIL (1.271). This efficiency directly translates to faster inference and reduced computational costs, critical factors for deployment in resource-constrained environments.

Resource-Dependent Performance: The results reveal clear patterns based on language resource levels. On high-resource languages with extensive Wikipedia coverage (Assamese: 1M sentences, Meitei: 1.35M sentences), mBERT's massive pretraining corpus (2.5TB) provides competitive performance (1.65 and 1.44 PPL respectively). However, on ultra-low-resource languages (Pnar: 1,002 sentences, Kokborok: 2,463 sentences, Nyishi: 55,870 sentences) where mBERT has minimal exposure, NE-BERT's targeted training yields substantial gains. The failures of IndicBERT-V2 and MuRIL on these languages (66.92–187.20 PPL) demonstrate that simply scaling model size without adequate language representation is insufficient for ultra-low-resource scenarios.

Script Diversity Handling: The fertility and BPC improvements are particularly striking for non-Latin scripts. Assamese (Bengali-Assamese script) shows substantial efficiency gains in BPC (0.230 vs. 0.442 for mBERT, 0.880 for IndicBERT-V2), while Meitei (also using Bengali-Assamese script) demonstrates similar patterns. For Latin-script languages, NE-BERT achieves competitive or superior efficiency across all metrics. This validates our choice of Unigram tokenization with weighted sampling, which better preserves script-specific morphological boundaries than BPE-based approaches used in baseline models.

Anchor Language Transfer: The strong performance on both anchor languages (English: 1.55 PPL, Hindi: 1.43 PPL) despite being down-weighted during tokenizer training (0.2× and 0.05× respectively) demonstrates effective cross-lingual transfer. NE-BERT outperforms IndicBERT-V2 (14.81 and 10.08 PPL) and MuRIL (5.57 and 5.75 PPL) on both anchor languages, suggesting that our weighted sampling strategy successfully balances vocabulary allocation without sacrificing anchor language performance. This is particularly important for real-world deployment where code-switching between Northeast languages and Hindi/English is common.

8 Downstream Evaluation

To validate practical utility beyond perplexity metrics, we evaluate NE-BERT on part-of-speech (POS) tagging across three Northeast Indian languages using Universal Dependencies annotations (Nivre et al., 2020).

8.1 Experimental Setup

We fine-tune all models on language-specific POS tagging using the following datasets:

- **Khasi**: 519 sentences (414 train, 52 dev, 53 test) (Ghosh et al., 2025)
- **Mizo**: 502 sentences (402 train, 50 dev, 50 test) (Ghosh et al., 2025)

Language	NE-BERT	mBERT	IB-V2	MuRIL
Khasi	**87.7**	82.3	80.1	61.2
Mizo	**73.2**	66.3	55.7	43.1
Nagamese	**86.4**	71.3	41.7	44.8
Average	**82.4**	73.3	59.2	49.7

Table 6: POS tagging accuracy (%) on test sets. NE-BERT outperforms all baselines across all languages.

- **Nagamese**: 214 sentences (170 train, 22 dev, 22 test) (Maiti et al., 2025)

All models are fine-tuned for 5 epochs with learning rate 2×10^{-5}, batch size 16, and standard cross-entropy loss. We compare NE-BERT against mBERT, IndicBERT-V2 (Doddapaneni et al., 2023), and MuRIL (Khanuja et al., 2021). We emphasize that these results are illustrative rather than definitive, given the small size of available annotated datasets.

8.2 POS Tagging Performance

Table 6 presents POS tagging accuracy on test sets. NE-BERT achieves the highest accuracy across all three languages, with an average of 82.4%, outperforming mBERT by 9.1 percentage points and IndicBERT-V2 by 23.2 percentage points.

The results demonstrate that NE-BERT's specialized vocabulary and targeted pretraining translate to improved performance on downstream tasks. The particularly large gap against IndicBERT-V2 and MuRIL on Nagamese (86.4% vs. 41.7–44.8%) highlights the importance of adequate language representation during pretraining.

9 Conclusion

We present NE-BERT, a multilingual encoder model for 9 Northeast Indian languages, demonstrating that domain-specific models with appropriate tokenization can effectively serve ultra-low-resource languages with as few as 1,000 training sentences. Our model outperforms IndicBERT-V2 across all 9 evaluated languages (15.97× average improvement) and achieves competitive or superior performance compared to mBERT (2.21 vs. 2.76 average PPL), with particularly strong gains on ultra-low-resource languages like Pnar (66.92 vs. 2.92 PPL), Kokborok, and Nyishi (187.20 vs. 4.33 PPL). Downstream evaluation on POS tagging validates practical utility, with NE-BERT achieving 82.4% average accuracy across Khasi, Mizo, and Nagamese, outperforming mBERT by 9.1 percentage points.

The key innovations (weighted Unigram tokenization, aggressive upsampling for ultra-low-resource languages, and cost-effective training at $7.31 on a single A40 GPU) provide a practical blueprint for developing language models for underrepresented languages worldwide. Our tokenization efficiency improvements (1.50× better fertility than mBERT, 4.3× better BPC than IndicBERT-V2) demonstrate that careful vocabulary optimization can substantially reduce inference costs while improving model quality.

This work represents a foundation for future NLP research on Northeast Indian languages. We release NE-BERT, tokenizer, training code, and documentation under CC-BY-4.0 to support community-driven improvements and applications.

Limitations

Limited Downstream Evaluation: While we validate NE-BERT on part-of-speech tagging for three languages (Khasi, Mizo, Nagamese), comprehensive evaluation across diverse tasks (named entity recognition, sentiment analysis, machine translation) and all nine trained languages remains future work. The small scale of available POS datasets (214–519 sentences) limits statistical robustness of downstream results.

Encoder-Only Architecture: NE-BERT is limited to representation tasks (classification, NER, embedding generation). Generation tasks (machine translation, summarization, dialogue) require decoder or encoder-decoder architectures.

Ultra-Low-Resource Vulnerability: Languages with fewer than 3,000 sentences (Pnar, Kokborok, Garo, Naga) remain vulnerable to distribution shift. While weighted sampling mitigates vocabulary fragmentation, these models may exhibit unexpected behavior on out-of-distribution inputs.

Future Work

Comprehensive Downstream Evaluation: We are developing benchmark datasets for named entity recognition, sentiment analysis, and additional part-of-speech tagging datasets across all 9 Northeast Indian languages, including Nyishi and Naga. Expanding POS evaluation beyond the current three languages and increasing dataset sizes will provide more robust assessment of NE-BERT's practical

utility.

Decoder Models: Extending our approach to autoregressive architectures would enable generation tasks. We plan to train decoder-only models using the same data curation and tokenization strategies, targeting conversational assistants for Northeast Indian languages.

Data Expansion: Active collaboration with native speaker communities and linguistic experts to expand corpora, particularly for ultra-low-resource languages. Target is 10K+ sentences for Pnar, Kokborok, Garo, and Naga.

Cross-Lingual Transfer Studies: Systematic investigation of zero-shot and few-shot transfer capabilities to related but unrepresented languages (e.g., Bodo, Karbi, Dimasa) to assess generalization beyond training languages.

Deployment Studies: Real-world deployment pilots with government and educational institutions to assess model performance on authentic tasks and gather community feedback.

Ethical Considerations

Bias and Representation

Language models inherit biases present in training data (Blodgett et al., 2020). Our collected corpora, derived from public government, educational, and media sources may contain gender, religious, caste, and other social biases reflecting the perspectives of text authors and publishers. Ultra-low-resource languages face additional risks:

- **Dominance Bias**: High-resource languages (Meitei, Assamese) may dominate model behavior despite weighted sampling, potentially marginalizing ultra-low-resource languages in multilingual contexts.
- **Quality Variance**: Limited data for Pnar, Kokborok, Garo, and Naga increases sensitivity to data quality issues and potential amplification of biases present in small corpora.
- **Hallucination Risk**: Models may generate plausible-sounding but incorrect content when faced with out-of-distribution inputs for ultra-low-resource languages.

We recommend thorough evaluation and community feedback before deploying NE-BERT in sensitive applications such as education, government services, or content moderation.

Linguistic and Cultural Impact

Language technologies can both preserve and threaten linguistic diversity (Bird, 2020). While NE-BERT enables digital inclusion for marginalized languages, potential negative impacts include:

- **Standardization Pressure**: Models may favor formal or written registers over spoken varieties, potentially marginalizing dialectal variation and informal language use.
- **Power Dynamics**: Deployment without community consent or benefit-sharing could reinforce extractive relationships between researchers and language communities.
- **Representation Gaps**: Our dataset primarily reflects government and educational registers, potentially underrepresenting oral traditions, youth language, and non-elite perspectives.

We are committed to:

- Transparent documentation of data sources, model limitations, and intended use cases
- Ongoing collaboration with native speaker communities for feedback and validation
- Benefit-sharing through open-source release and support for community-driven applications
- Respect for community decisions regarding data use and model deployment

Acknowledgements

We thank the reviewers for their valuable feedback. We are grateful to the Workshop on Machine Translation (WMT) 2025 organizers (WMT 2025 Organizers, 2025) for providing the Nyishi and Kokborok parallel corpora. We extend our gratitude to the native speaker communities of Northeast India for their contributions to language preservation efforts. This work was supported by MWire Labs.

References

Judit Ács. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021*, pages 342–349.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Ethan C. Chau and Lucy H. Lin. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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.

Soumyadip Ghosh, Nagaraju Vuppala, Dorothy Marbaniang, and Henry Lalsiam. 2025. Towards resource-rich Mizo and Khasi in NLP: Resource development, synthetic data generation and model building. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 176–185.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent approach to subword tokenization and detokenization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Agniva Maiti, Manya Pandey, and Murari Mandal. 2025. NagaNLP: Bootstrapping NLP for low-resource Nagamese with closed-loop synthetic data. arXiv preprint arXiv:2512.12537.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In

Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4034–4043.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

The Unicode Consortium. 2021. *The Unicode Standard, Version 14.0*. Mountain View, CA.

Benjamin Warner, Pawan Maruf, Marcos Treviso, Alham Fikri Aji, Inigo Jauregi Unanue, Jason Phang, Bin Shao, Jingyu Xu, Chuan Jun Yee, Jiayu Lin, Camille Thorne, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

WMT 2025 Organizers. 2025. Findings of the WMT 2025 shared task on low-resource language translation. In *Proceedings of the Tenth Conference on Machine Translation*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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*, pages 483–498.

Appendix

A Training Loss Curves

Figure 1 shows the complete training and validation loss curves over 10 epochs. The close tracking between training and validation loss throughout training indicates effective generalization without overfitting, despite the small corpus size for ultra-low-resource languages.

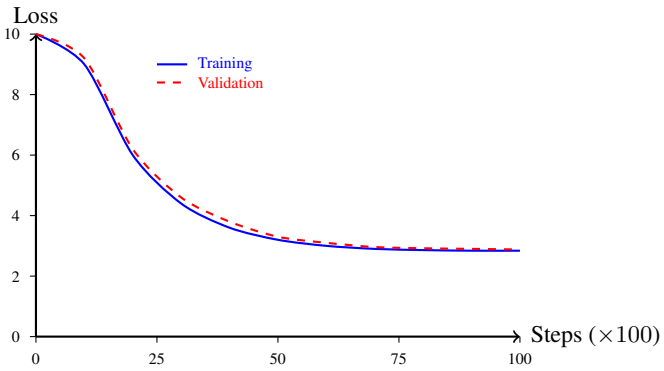

Figure 1: Training and validation loss curves over 10 epochs.

B Language Examples

Table 2 presents representative sentences from each of the 9 Northeast Indian languages in our corpus, demonstrating script diversity and morphological variation across language families.

Language	Example Sentence
Assamese	কোনো লৰাৰ বাৰ্ষা ওপৰত কোকা কোনো নিয়মৰ গাত আচোৰ এটা লাগিলেই তাৰ পিঠিত বেত্রাঘাত পৰিছিল *Kono lorar bara oporot koka kono niymor gat achor eta lagilei tar pithit betraghaat porisil*
Meitei	মসিগী প্রোগ্রাম অসি য়াম্না কান্নবা প্রোগ্রামনি । *Masigi program asi yamna kannaba programni.*
Khasi	Kane kan ïarap ban pynsuk ïa ka leit ka wan jong ki kali Tourist hapdeng ka jylla baroh ar
Mizo	A hnuai ami ang hian a ngai ngaiin I lo dah chhuak dawn teh ang
Garo	ran.gipa, jasenggipa aro balwa jokrurana man.gipa ong.na nanga
Kokborok	Nwng borokni creativity no buji mano da?
Pnar	Lada ym bood wan phi daw man phi kam kitu kiwa tei ia ka iung iong phi hajrong u chyiap.
Naga	aru jisu ekta naw te uthise, itu simon laga naw thakise, aru tai laga naw ke olop pani te loi jabole koise.
Nyishi	Nyilin bo nge jvqtw ngam cengden hvb nyido

Figure 2: Representative sentences from each Northeast Indian language in our corpus. For Bengali-Assamese script languages (Assamese, Meitei), both the original script and Latin transliterations (in italics) are shown.