EACL 2026
RABAT · MOROCCO
Mars · March 24-29, 2026 · مارس

Beyond Multilinguality:

Typological Limitations in Multilingual Models for Meitei Language

Badal Nyalang

MWire Labs, Shillong, India · nyalang@mwirelabs.com

MeiteiRoBERTa

76M words · Bengali script

THE 8TH WORKSHOP ON RESEARCH IN COMPUTATIONAL LINGUISTIC TYPOLOGY AND MULTILINGUAL NLP
SIGTYP 2026 @ EACL · MARCH 29, 2026

MWire

Outline

Motivation: Why Meitei?

THE LANGUAGE

1.8M+

Speakers in Manipur, Northeast India

Tibeto-Burman

Agglutinative morphology, SOV order

Head-final

Phrase structure, complex suffixation

Bengali script

Primary digital writing system

THE GAP: EXISTING MULTILINGUAL MODELS

mBERT 104 languages	Spreads capacity too thin - low-resource langs underserved
MuRIL 17 Indian langs	Tibeto-Burman structure underfitted
IndicBERT 12 Indian langs	Meitei not among 12 languages coverage
XLM-R 100 languages	'Curse of multilinguality' - per-lang performance drops

No publicly available pre-trained LM for Meitei existed before this work.

MeiteiRoBERTa: Architecture & Training

ROBERTA-BASE CONFIGURATION

Architecture base
RoBERTa-base

Total parameters
125 million

Transformer layers
12

Tokenizer
Byte-Pair Encoding (BPE) - 52,000 tokens

Attention heads
12

Script
Bengali script

Hidden dimension
768

Training objective
Masked Language Modeling (MLM) · 15% masking probability

MeiteiRoBERTa: Architecture & Training

2 of 2 — Training Data & Configuration

76 **Million Words**
of curated Meitei text in Bengali script

353,123 blocks · 512 tokens each

DATA SOURCES

IndicCorp v2	Gala et al. (2023) — primary foundation corpus
News archives	Local Meitei news portals and online publications
Government docs	Official government publications and circulars
Literary collections	Digitised literary works and web content

LR 6×10^{-4} · Effective batch size 256 · 3 epochs · NVIDIA A40 · Final loss 4.1855

Evaluation Framework

THREE COMPLEMENTARY METRICS
BASELINES: MBERT (110M, 104 LANGS) AND MURIL (235M, 17 INDIAN LANGS)

01 Perplexity

↓ lower is better

- Held-out validation set: 19,038 Meitei samples
- PPL = exp(average MLM loss)
- Measures how well the model predicts Meitei tokens

02 Tokenization Efficiency

↓ lower is better

- Subword fertility: avg tokens per word
- Measured on 10 diverse Meitei sentences
- Lower fertility → vocabulary better aligned with morphology

03 Semantic Representation

↑ separation is better

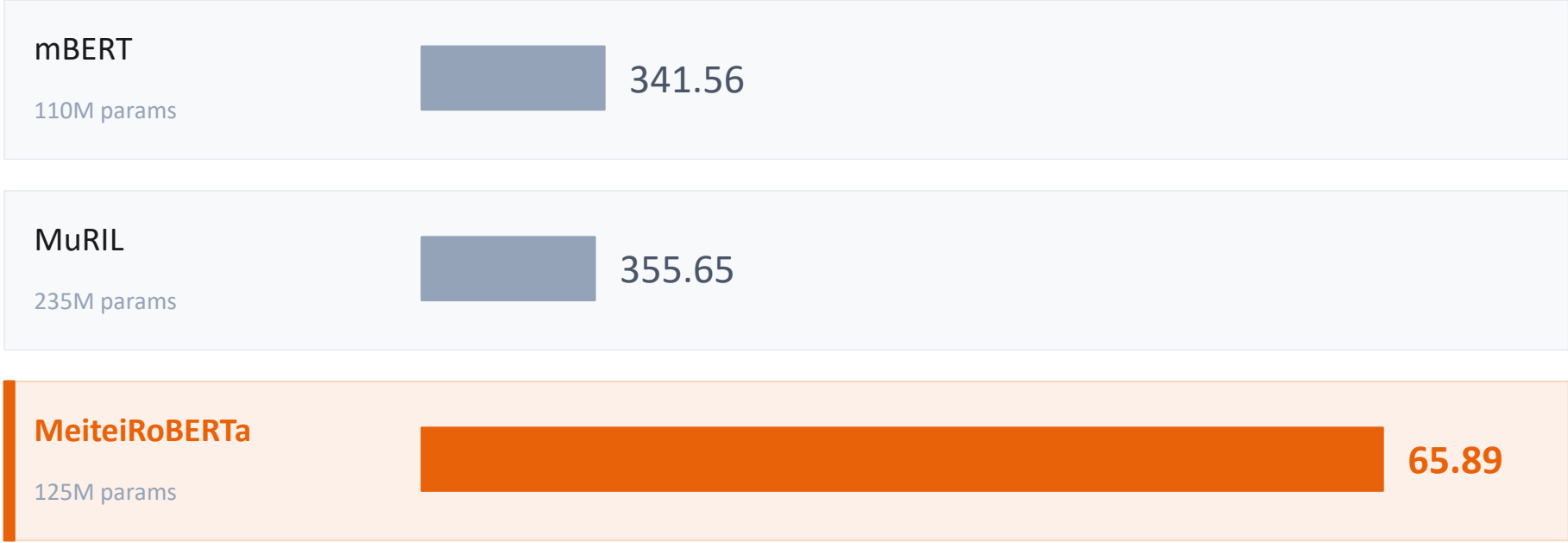
- Cosine similarity of [CLS] embeddings
- 4 curated pairs: 2 similar · 2 dissimilar
- Separation = avg(high) − avg(low similarity)

[CLS] pooling as typological probe: if a model cannot separate Meitei semantics, it has not encoded the language's structure

Results

1 of 3 — Perplexity

LANGUAGE MODELING QUALITY ON 19,038 HELD-OUT MEITEI SAMPLES · LOWER IS BETTER



5.2× better perplexity than mBERT · **5.4×** better than MuRIL — despite fewer parameters than MuRIL (125M vs 235M)

Results

2 of 3 — Semantic Representation Quality

COSINE SIMILARITY OF [CLS] EMBEDDINGS ON 4 CURATED MEITEI SENTENCE PAIRS
HIGHER SEPARATION IS BETTER

Model	High Similarity	Low Similarity	Separation Score
mBERT	0.983	0.948	0.035
MuRIL	0.993	0.993	0.000
MeiteiRoBERTa ★	**0.968**	**0.199**	**0.769**

MuRIL: 0.000 separation

Assigns identical similarity scores to both related and unrelated Meitei sentences - no semantic discrimination.

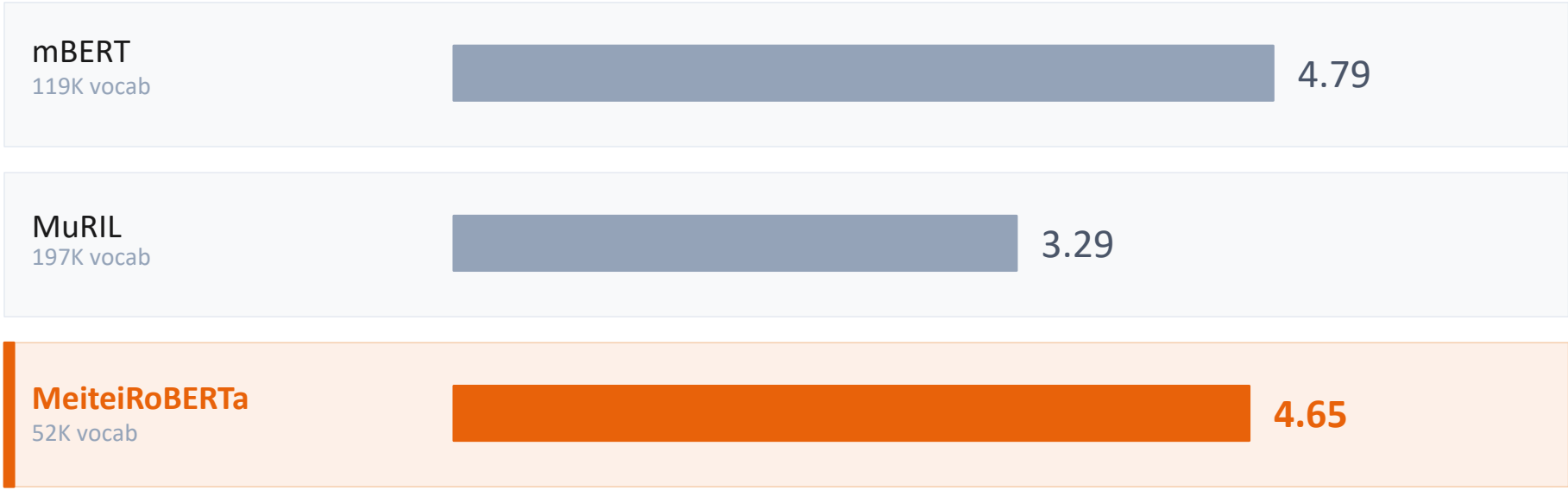
MeiteiRoBERTa: 0.769 separation

Correctly assigns high similarity (0.968) to related pairs and low (0.199) to unrelated ones - genuine semantic understanding.

Results

3 of 3 — Tokenization Efficiency

SUBWORD FERTILITY: AVERAGE TOKENS PER WORD ON 10 DIVERSE MEITEI SENTENCES
LOWER IS BETTER



Key nuance: MuRIL's best fertility (3.29) comes from a 197K vocabulary — yet it scores 0.000 semantic separation. Efficient tokenization alone does not guarantee language understanding.

Discussion: Typological Insights

1 Efficient tokenization ≠ semantic understanding

MuRIL achieves the best subword fertility (3.29) through its large 197K vocabulary, yet produces 0.000 semantic separation on Meitei sentence pairs. Efficient tokenization is a necessary but not sufficient condition for genuine low-resource language support. Dedicated model capacity matters far more.

2 The curse of multilinguality hits typologically distant languages hardest

Meitei's Tibeto-Burman SOV agglutinative structure is typologically distant from the languages that dominate mBERT and MuRIL pretraining. When model capacity is shared across typologically diverse languages, languages at the periphery receive systematically insufficient signal.

Discussion: Typological Insights

3 Targeted data beats raw scale

MeiteiRoBERTa trained on just 76 million words outperforms models trained on orders of magnitude more multilingual data. This demonstrates that language-specific pretraining with an appropriate tokenizer is the dominant variable — not corpus size. This is especially encouraging for endangered languages with limited digital text.

4 Implications extend to 220+ Northeast Indian languages

Northeast India alone hosts over 220 distinct languages, most sharing Meitei's typological profile: agglutinative morphology and head-final syntax. Current multilingual models are structurally disadvantaged for this entire region. Our results suggest that dedicated monolingual models — even on modest data — are the most viable path forward.

Finding: dedicated monolingual pretraining + language-specific tokenization remains the most effective approach for typologically underserved languages

Limitations & Future Work

AREAS FOR IMPROVEMENT IN FUTURE ITERATIONS

Script coverage
Bengali script only - excludes the indigenous Meitei Mayek script still used in traditional and educational contexts. Future work should explore dual-script or script-agnostic representations.

Evaluation scope
Results are intrinsic only - perplexity, tokenization, and semantic probing. Downstream task evaluation (NER, sentiment analysis, machine translation) is needed to confirm practical utility.

Corpus bias
Training data skewed toward formal news and government text. Dialectal variation, informal registers, and oral tradition-derived text are underrepresented.

Community involvement
Model and dataset development occurred without extensive participatory consultation with Meitei language communities. Future efforts should centre community priorities.

Bias audit
Comprehensive evaluation of potential demographic or topical biases in model outputs has not yet been conducted.

Limitations & Future Work

NEXT STEPS FOR MEITEIROBERTA AND FUTURE WORK

→ **Meitei Mayek**	Dual-script model or script-agnostic representations to cover the indigenous script
→ **Downstream tasks**	Fine-tune for named entity recognition, sentiment analysis, and machine translation
→ **Few-shot transfer**	Explore few-shot learning capabilities for related Tibeto-Burman languages
→ **Community-led design**	Participatory development aligned with Meitei speaker community needs and values

OPEN RESOURCES

Model	huggingface.co/MWirelabs/meitei-roberta
Dataset	huggingface.co/datasets/MWirelabs/meitei-monolingual-corpus

Conclusion

MeiteiRoBERTa: Key Takeaways

5.2×	better perplexity than mBERT (65.89 vs 341.56) monolingual pretraining wins
0.769	semantic separation vs 0.035 (mBERT) · 0.000 (MuRIL) genuine language understanding
76M	words sufficient to outperform models on billions targeted data beats raw scale
First	publicly available transformer-based LM for Meitei opening the field for NE Indian NLP

Dedicated monolingual pretraining remains essential for typologically underserved languages

Thank you!

Questions?

MWire

nyalang@mwirelabs.com