Stereotyped by Silence: How LLMs Erase Northeast Indian Languages Through Omission and Orthographic Corruption

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

Abstract

Large language models (LLMs) perpetuate cultural stereotypes not only through biased associations but through systematic omission and orthographic erasure of underrepresented languages. We present empirical evidence of two compounding failure modes affecting Northeast Indian languages: (1) *entity-level invisibility*, where state-of-the-art NER systems score F1 = 0.000 on culturally critical named entities such as Khasi surnames, Garo festivals, and tribal names; and (2) *orthographic corruption*, where LLM tokenizers corrupt semantically meaningful diacritics (ï, ñ) and the Garo morpheme boundary marker (U+00B7, ·) at rates of 18.8–50% across four of five evaluated models. Drawing on NortheastNER (F1 = 0.964, six entity categories, XLM-RoBERTa-base) and a systematic tokenization study across Khasi and Garo, we argue that stereotype-by-omission constitutes a distinct and measurable harm to indigenous language communities. We further show that a custom multilingual tokenizer achieves 26–50% token reduction over five baseline LLMs, demonstrating that culturally grounded infrastructure can partially remediate these failures. Our findings call for cultural representation audits as a standard component of multilingual NLP evaluation.

1 Introduction

Stereotype research in NLP has concentrated on biased associations: models linking gender to occupation, or ethnicity to negative sentiment (Blodgett et al., 2020; Gallegos et al., 2024). This focus, while important, leaves a more fundamental problem unaddressed. When a model cannot recognize the name of a tribal community, cannot preserve the diacritics that distinguish words in an indigenous language, or has never encountered the name of a major regional festival, the harm is not an association. It is an absence. The community does not receive a distorted reflection; it receives none at all.

Northeast India makes this concrete. The region comprises eight states and over 220 distinct languages spanning the Austroasiatic, Tibeto-Burman, and Indo-Aryan families, alongside contact varieties such as Nagamese. Despite tens of millions of speakers, these languages are almost entirely absent from major multilingual NLP systems (Joshi et al., 2020). The consequences are not abstract. NER systems that score F1 = 0.000 on Khasi surnames cannot support legal document processing, government service delivery, or cultural archiving in Khasi. Tokenizers that corrupt Garo morpheme markers at 50% rates cannot serve as reliable infrastructure for any downstream Garo application.

This paper presents evidence of two failure modes that are facets of a single underlying problem: the systematic exclusion of Northeast Indian languages from multilingual NLP infrastructure. First, we demonstrate *entity-level invisibility* through NortheastNER, a domain-specific NER model for Northeast India. Baseline multilingual models score F1 = 0.000 on entities such as *Lyngdoh* (a prominent Khasi surname), *Wangala* (the principal Garo harvest festival), and *Garo* (the tribal community itself). NortheastNER, fine-tuned on domain-specific data, achieves F1 = 0.964 on the same entities. Second, we demonstrate *orthographic erasure* through a systematic evaluation of five LLMs on Khasi diacritics (ï, ñ) and the Garo morpheme boundary marker (U+00B7). Four of five models corrupt these characters at rates between 18.8% and 50%. A custom multilingual tokenizer achieving 26–50% token reduction across five languages demonstrates that both failure modes are addressable through community-grounded infrastructure.

Together, these findings operationalize *stereotype-by-omission* as a measurable harm category, extending existing frameworks for representation disparity (Joshi et al., 2020; Gallegos et al., 2024) toward communities absent

Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026), pages 62–68

from model training entirely, and propose cultural representation audits as a practical response.

2 Background and Related Work

2.1 Bias as Association vs. Bias as Omission

The dominant paradigm in NLP bias research treats stereotyping as an association problem (Blodgett et al., 2020). Models encode associations between demographic groups and attributes, and these associations reflect and amplify societal biases (Gallegos et al., 2024). Hofmann et al. (2024) extend this to covert discrimination: LLMs make systematically worse decisions about speakers of African American English based on dialect cues alone (Hu et al., 2025). Tao et al. (2024) show that LLMs exhibit strong Western value alignment via World Values Survey comparisons, reflecting whose values were encoded during training.

These findings assume the target community is represented in training data. For most Northeast Indian language communities, that assumption does not hold. The failure is not distortion but erasure. The analysis must start earlier, at the level of whether the community appears in the model's representational world at all.

2.2 Low-Resource Languages and Tokenization

Joshi et al. (2020) document the steep gradient of linguistic inclusion in NLP. Northeast Indian languages fall into the lowest resource tiers, with no representation in standard benchmarks such as Flores-101 (Goyal et al., 2022). Tokenization amplifies this exclusion. Rust et al. (2021) show that tokenization fertility correlates strongly with downstream task performance for low-resource languages (Maksymenko and Turuta, 2025). Multilingual tokenizers fragment low-resource language text into suboptimal units, raising inference cost and degrading model understanding. Chang and Bergen (2024) note that multilingual performance gaps in LLMs are often traceable to data sparsity and tokenization artifacts. For languages with semantically meaningful diacritics or morpheme markers, the problem is compounded: tokenizers may corrupt the characters themselves, silently altering meaning.

2.3 NER and Prior NLP Work on Northeast Indian Languages

NER for Northeast Indian languages is almost entirely absent from the literature. Warjri et al. (2021) develop the first POS tagging corpus for Khasi, noting minimal computational resources for the language. Hujon et al. (2024) present neural machine translation systems for English-Khasi, highlighting the unique challenges of Austroasiatic language structure, but focus on translation quality rather than foundational entity recognition. No prior NER system addresses the entity types that matter for regional applications: tribal communities, indigenous festivals, endemic flora and fauna (Radchenko and Drushchak, 2025).

3 Stereotype by Omission: A Framework

We distinguish two forms of stereotype-by-omission relevant to this work.

Entity omission occurs when a model's training distribution renders it incapable of recognizing culturally significant named entities. The failure is not a biased output but an absent one. A model that scores F1 = 0.000 on *Wangala* does not misclassify it. It does not register it at all. The implicit encoding is that this festival, and the community that celebrates it, falls outside the scope of what the model knows.

Orthographic erasure occurs when a tokenizer corrupts or discards characters that carry semantic weight in a community's writing system. For Khasi, the diacritics ï and ñ distinguish word meanings and carry morphological information. A tokenizer that replaces ï with *i* does not produce a near-equivalent. It produces a different word. The community's orthographic conventions are treated as noise to be normalized away.

Both failure modes encode a cultural hierarchy rooted in training data composition. Entities and orthographies that appear frequently in the dominant training corpus are handled correctly; those outside it are erased. Hofmann et al. (2024) show that distributional absences produce measurable real-world harms within a single language. The same logic applies at the level of entire language communities.

4 NortheastNER: Entity-Level Visibility

4.1 Task Design

We developed NortheastNER, a domain-specific NER model for Northeast India, covering six en-

tity categories: PLACES (villages, districts, geographic locations), TRIBES (tribal communities and sub-groups), FESTIVALS (cultural events and traditional celebrations), TOURIST (sites and attractions), FLORA (plant species), and FAUNA (animal species). These categories reflect the named entity landscape of Northeast India rather than the standard PER/LOC/ORG schema, which does not capture the entities that matter for regional applications.

4.2 Data and Model

NortheastNER fine-tunes `xlm-roberta-base` (Conneau et al., 2020) on a weakly supervised corpus of approximately 25,000 labeled sentences. Training data sources: a Northeast India geographic gazetteer (∼45,000 village and district entries), a Northeast India tribal entity dataset (427 entities), Himalayan biodiversity databases, and curated ethnographic and cultural documentation. Gazetteer-based weak supervision with BIO tagging generated training labels; a conflict resolution pipeline resolved overlapping spans by prioritizing more specific entity categories. Splits were performed at sentence level after corpus generation; gazetteer entries themselves were deduplicated, though lexical overlap of high-frequency regional place names across splits cannot be fully excluded. Hyperparameters: learning rate 3e-5, AdamW, batch size 16, max sequence length 256, weight decay 0.01, 3 epochs, single A4500 GPU (20 GB).

4.3 Baselines

The primary comparison is between NortheastNER and MuRIL (Khanuja et al., 2021), both fine-tuned on the same weakly supervised training data under identical hyperparameters. The untrained `xlm-roberta-base` baseline serves as a sanity check, not a competitive system: it confirms that zero-shot multilingual encoders have no representational capacity for these entity types. The scientifically meaningful comparison is between the two fine-tuned models, NortheastNER and MuRIL, trained under identical conditions.

4.4 Results

Table 1 shows entity-level seqeval scores on the held-out development set and on authentic regional texts.

The untrained baseline achieves F1 = 0.000 across all categories, confirming total representational absence. IndicNER, evaluated qualitatively, recognized only high-resource geographic names already present in standard corpora (*Tura*, *Wards Lake*). It failed entirely on *Lyngdoh*, *Wangala*, and *Garo* as a tribal entity. The gap between development set performance (F1 = 0.964) and authentic regional text (F1 = 0.778) is expected and interpretable: development data shares lexical distribution with training gazetteers, while real-world text introduces out-of-gazetteer names and TOURIST/PLACES boundary ambiguity. This gap does not undermine the core finding; it confirms that gazetteer-based supervision has known limits and that domain-specific data collection remains necessary.

Model	P	R	F1
xlm-roberta-base (untrained)	0.000	0.000	0.000
MuRIL (fine-tuned)	0.952	0.950	0.951
NortheastNER (ours)	**0.962**	**0.967**	**0.964**
NortheastNER (real-world)	0.980	0.645	0.778

Table 1: NER comparison. Development set results above the rule; real-world test on authentic regional texts below. P = Precision, R = Recall.

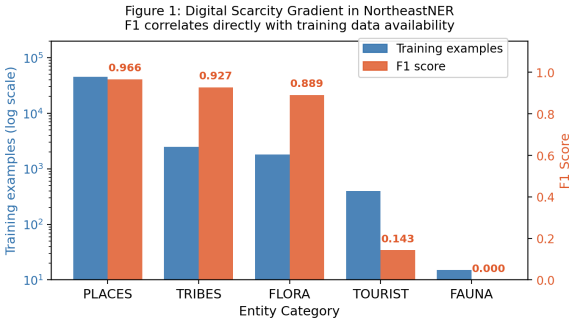

Figure 1: F1 score and training data availability per entity category. Performance tracks data availability directly, illustrating how cultural and ecological marginalization compounds through digital scarcity.

4.5 The Digital Scarcity Gradient

Figure 1 and Table 2 show per-category results.

PLACES achieves F1 = 0.966 supported by 45,000 gazetteer entries. FAUNA achieves F1 = 0.000 with 15 examples. This is not a modeling failure. It is a data failure that reflects a deeper reality: digital records for regionally endemic fauna are almost entirely absent. The scarcity gradient maps directly onto cultural and ecological marginalization that precedes any NLP system.

Category	Train examples	F1
PLACES	45,000	0.966
TRIBES	2,500	0.927
FLORA	1,800	0.889
TOURIST	400	0.143
FAUNA	15	0.000

Table 2: NortheastNER per-category results. F1 correlates directly with training data volume.

Model	Preserved	Corrupted
Gemma-2-2B	100%	0%
Falcon3-3B	81.2%	18.8%
Nemotron-Mini-4B	∼60%	∼40%
Llama-3.2-3B	∼60%	∼40%
Falcon-H1-3B	50%	50%

Table 3: Orthographic preservation of Khasi (ï, ñ) and Garo (U+00B7) special characters across five LLMs.

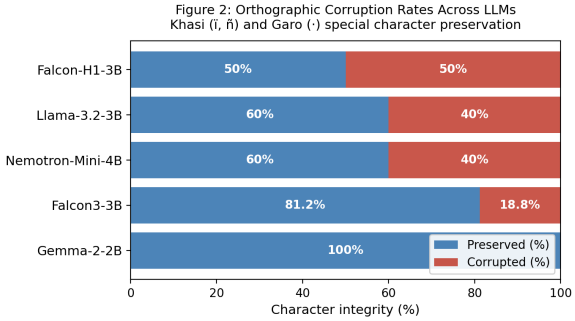

Figure 2: Character integrity vs. corruption rate per model. Only Gemma-2-2B achieves full preservation. Four of five models corrupt semantically meaningful characters at rates between 18.8% and 50%.

5 Orthographic Erasure: Tokenization Study

5.1 Linguistic Background

Khasi (Austroasiatic, ∼1.4M speakers) uses diacritics ï and ñ that carry semantic weight. Their removal changes word meaning, not merely appearance. Garo (Tibeto-Burman, ∼1.2M speakers) uses the middle dot (U+00B7, ·) as a morpheme boundary marker essential for grammatical parsing (Warjri et al., 2021). Both languages use Latin script, which might appear to ensure tokenizer compatibility. This is not the case when specific orthographic conventions fall outside the training distribution.

5.2 Evaluation Setup

We evaluated five LLMs on their preservation of these features: Gemma-2-2B, Falcon3-3B, Nemotron-Mini-4B, Llama-3.2-3B, and Falcon-H1-3B. Text samples were encoded using each model's HuggingFace tokenizer (Transformers v4.36.0) and decoded back to verify round-trip character integrity. Input text was NFKC-normalized prior to encoding to rule out normalization artifacts as a confound. Character integrity rate measures the proportion of special characters (ï, ñ, U+00B7) correctly preserved through encode-decode across standardized Khasi and Garo samples.

5.3 Results

Four of five models corrupt between 18.8% and 50% of semantically meaningful characters (Table 3, Figure 2). Only Gemma-2-2B achieves full character integrity. A developer who selects Falcon-H1-3B as a base model for a Khasi application will silently corrupt half of the diacritics that distinguish word meanings, with no error signal. The variation across models is the key finding: this is not an inherent limitation of LLM tokenization. It is a design and training data choice.

6 Toward Remediation: Custom Multilingual Tokenization

Following community-grounded approaches to building regional language infrastructure (El Mekki et al., 2025; Nyalang, 2026), we present a custom SentencePiece Unigram tokenizer trained on 27,500 sentences (5,500 per language) across five Northeast Indian languages: Khasi, Garo, Mizo, Nyishi, and Nagamese. The 10,000-token vocabulary was constructed to respect morphological boundaries and orthographic conventions of all five languages. NFKC normalization was applied consistently during training and evaluation. Compression efficiency was evaluated on a held-out test set of 2,000 sentences per language (10,000 total), strictly separated from training data. Compression is computed as:

$$Compression(\%) = \left(1 - \frac{T_{custom}}{T_{baseline}}\right) \times 100$$

benchmarked against five baseline LLMs using HuggingFace AutoTokenizer (Rust et al., 2021).

Table 4 shows per-language token length comparisons. The custom tokenizer achieves 26–50% token reduction over all five baselines. Gemma-2-9B shows the best compatibility: 33% vocabulary overlap and the lowest compression gap of 26.63%, consistent with its superior character integrity in Section 5. Nagamese Creole exhibits the highest tokenization inefficiency (31–45% compression gap)

Language	Custom	Gem	Lla	Opn	Mis
Khasi	28.12	36.08	40.18	44.50	44.66
Garo	23.52	32.21	35.72	39.08	38.74
Mizo	29.01	39.90	43.33	47.08	47.49
Nyishi	9.89	13.53	14.86	16.25	16.18
Nagamese	31.84	46.52	53.47	57.16	58.31

Table 4: Mean tokens per sentence: custom tokenizer vs. baseline LLMs (Gem = Gemma-2-9B, Lla = Llama-3.2-8B, Opn = OpenHathi-7B, Mis = Mistral-7B-v0.3). Evaluated on 2,000 held-out sentences per language.

despite approximately 30 million L2 speakers, doubly marginalized by both national and international NLP infrastructure.

The custom tokenizer demonstrates that culturally grounded tokenization for this language family is technically feasible. It is, however, a remediation and not a solution. The structural problem is whose languages are considered at tokenizer design time.

7 Discussion

7.1 Omission as a Distinct Harm Category

Current taxonomies of NLP bias address representation harms, allocation harms, and quality-of-service disparities (Gallegos et al., 2024). Stereotype-by-omission does not fit neatly into these categories. It is not an association. It is a prior failure: the model has no representation of the community at all.

Hofmann et al. (2024) show that covert discrimination can operate through the absence of positive signals rather than the presence of negative ones. Our findings extend this logic further: not dialect variation within a represented language, but the near-total absence of entire language communities from the model's training distribution. A health information system built on a model that cannot recognize Khasi tribal names cannot serve Khasi communities, regardless of whether it produces explicitly biased outputs.

7.2 The Feedback Loop of Digital Scarcity

The digital scarcity gradient in Figure 1 illustrates a self-reinforcing dynamic. Low digital presence leads to low training data, which leads to poor model coverage, which makes it harder to build downstream applications, which reduces incentive to generate more digital content in the language. NER systems that fail on FAUNA with 15 training examples will not improve until more ecological documentation exists in these languages digitally. But the absence of usable NLP tools reduces incentive to produce that documentation.

Breaking this loop requires deliberate investment in data collection for culturally specific entity types, not merely general-purpose text. The category structure of NortheastNER, covering TRIBES, FESTIVALS, FLORA, and FAUNA, is itself an argument that relevant entities must be defined by communities, not inferred from high-resource language taxonomies.

7.3 Toward Cultural Representation Audits

We propose that cultural representation audits become a standard component of multilingual model evaluation. Such audits would measure: (1) entity coverage for culturally specific named entity types in the target region; (2) orthographic integrity for extended character sets used by the target language; and (3) tokenization efficiency relative to a language-specific reference tokenizer.

Standard benchmarks do not include Northeast Indian languages (Goyal et al., 2022), and quantitative disparities across languages in multilingual models remain underreported (Hu et al., 2025). The absence of evaluation infrastructure perpetuates the absence of model capability. NortheastNER and the supporting gazetteers will be released upon acceptance to support further evaluation work.

8 Conclusion

We have presented evidence that Northeast Indian languages are subjected to systematic stereotype-by-omission in current LLM infrastructure, manifesting as entity invisibility and orthographic erasure. NortheastNER demonstrates both the extent of baseline failures (F1 = 0.000 on core cultural entities) and the feasibility of domain-specific remediation (F1 = 0.964). A systematic tokenization evaluation shows orthographic corruption is widespread (18.8–50%) but model-dependent and therefore addressable. A custom multilingual tokenizer demonstrates that language-specific tokenization can partially close the gap.

We call for stereotype-by-omission to be recognized as a distinct harm category in NLP bias research, and for cultural representation audits to become standard evaluation practice. Communities absent from model training should not need to wait for general-purpose multilingual models to catch up. Targeted, community-grounded infrastructure is both faster and more appropriate.

NortheastNER, along with the supporting gazetteers and datasets, will be released upon acceptance.

Limitations

The tokenization study covers Khasi and Garo only; broader coverage across the 220+ languages of Northeast India requires further work. Closed-source models are not evaluated. Downstream task evaluation of the custom tokenizer is deferred to future work. NortheastNER relies on gazetteer-based weak supervision, which may underrepresent entities not captured in existing gazetteers. The FAUNA category (F1 = 0.000, 15 training examples) requires dedicated data collection before it is reliable. Sentence-level train/dev splits may contain lexical overlap of high-frequency regional place names. Exact corruption rates for Nemotron-Mini-4B and Llama-3.2-3B are approximated at $\sim$40%; precise per-character breakdowns are deferred to future work. Community validation studies with affected language communities have not been conducted and are planned as future work.

Acknowledgments

The author thanks the MWire Labs team for their support. We also acknowledge the indigenous language communities of Northeast India whose languages and cultural heritage motivate this work.

References

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. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

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. Association for Computational Linguistics.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025. Quantifying language disparities in multilingual large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Aiusha Vellintihun Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. 2024. Neural machine translation systems for English to Khasi: A case study of an Austroasiatic language. *Expert Systems with Applications*, 238(Part A):121813.

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. Association for Computational Linguistics.

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*.

Daniil Maksymenko and Oleksii Turuta. 2025. Tokenization efficiency of current foundational large language models for the Ukrainian language. *Frontiers in Artificial Intelligence*, 8:1538165.

Badal Nyalang. 2026. NE-BERT: A multilingual language model for nine Northeast Indian languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.

Vladyslav Radchenko and Nazarii Drushchak. 2025. Improving named entity recognition for low-resource languages using large language models: A Ukrainian case study. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 27–35, Vienna, Austria (online). Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Sunita Warjri, Partha Pakray, Saralin A. Lyngdoh, and Arnab Kumar Maji. 2021. Part-of-speech (POS) tagging using deep learning-based approaches on the designed Khasi POS corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(3).