NortheastGenAI 2026: Overview of the First Workshop on AI-Assisted Research for Northeast India

Badal Nyalang[1], Monali Longmailai[2], Clarissa Jane Rajee[3], Teisovi Angami[4], and Dibya Jyoti Bora[5]

[1] MWire Labs, Shillong, Meghalaya, India
[2] Assam University, Silchar, Assam, India
[3] Martin Luther Christian University, Shillong, Meghalaya, India
[4] Nagaland University, Nagaland, India
[5] Assam Kaziranga University, Assam, India

Abstract. We present NortheastGenAI 2026, the first workshop dedicated to AI-assisted research on Northeast India's languages, cultures, and ecosystems. Held virtually on May 29, 2026, the workshop received 10 submissions across three thematic tracks, of which 7 were accepted following an AI-assisted review process with mandatory human editorial oversight on every submission. Reviews were generated independently by two large language models – Claude Sonnet 4.6 (Anthropic) and Command A (Cohere) – across four criteria: Relevance, Plausibility, Novelty, and Clarity. The two systems agreed on 8 of 10 submissions (80% exact agreement). This overview describes the workshop's motivation, the review methodology, the accepted papers by track, and reflections on running an openly transparent, AI-reviewed academic workshop for one of the world's most linguistically diverse and underrepresented regions.

Keywords: Northeast India · low-resource languages · AI-assisted research · AI-assisted peer review · human-in-the-loop · language documentation · digital equity

1 Introduction

Can AI research Northeast India? This is the open question NortheastGenAI 2026 set out to explore. Northeast India is home to more than 200 languages spanning the Tibeto-Burman, Austro-Asiatic, and Tai-Kadai families, as well as extraordinary biodiversity and rich oral traditions that remain almost entirely absent from global AI research. Despite this linguistic and cultural richness, the region is critically underrepresented in natural language processing, digital humanities, and AI-for-society research.

NortheastGenAI 2026 was conceived as a direct response to this gap. Organized by MWire Labs, Shillong, and supported by a Cohere Labs Catalyst Grant, the workshop invited AI-assisted research submissions on Northeast India's languages, cultures, histories, and ecosystems. The workshop draws inspiration from

Agents4Science 2025 [1], the first conference in which AI agents served as both primary authors and reviewers, adapting the spirit of open, AI-assisted inquiry to a domain where the stakes are highest: low-resource, underrepresented, and urgently underdocumented knowledge systems of the Eastern Himalayan and Indo-Burma region.

This workshop does not position AI as a replacement for human scholarship. We are asking an open question and publishing everything that results. All submissions, prompting methods, and AI-generated reviews are publicly visible on OpenReview. Humans remain the final gatekeepers. This overview paper describes the workshop's design, the human-in-the-loop review process, the accepted papers by track, and reflections on this first edition.

2 Workshop Design and Submission Guidelines

NortheastGenAI 2026 was organized around three thematic tracks:

T1: Language, Culture and Heritage — low-resource NLP, ASR, MT, script and OCR, oral traditions, indigenous knowledge, and digital archiving.

T2: Society, History and Anthropology — community structures, oral history, ethnography, indigenous governance, gender studies, and historical documentation.

T3: AI and Technology for NE India — AI applications, EdTech, GovTech, digital infrastructure, multilingual systems, biodiversity monitoring, and climate resilience.

All submissions were required to be substantially AI-assisted or AI-generated with full disclosure of the AI model used, the date of generation, and the primary system prompt. This mandatory disclosure requirement ensures reproducibility of the research process itself, not only its outputs. Submissions without disclosure were desk rejected. Papers were required to engage with at least one track with a clear Northeast India focus and to include at least one empirical grounding element: a dataset, field reference, citation, or empirical observation. The workshop explicitly welcomed exploratory work, negative results, and pilot studies. All submissions were handled through OpenReview, with reviews and decisions made publicly visible.

3 AI-Assisted Review Process with Human-in-the-Loop Oversight

NortheastGenAI 2026 employed AI-assisted peer review supported by a Cohere Labs Catalyst Grant. Drawing on the Agents4Science [1] framework as a reference point, we designed a review process that placed human oversight at every stage — not as an exception, but as a structural requirement.

Each submission received two independent AI-assisted reviews, generated by Claude Sonnet 4.6 (Anthropic) and Command A (Cohere). Reviews assessed each paper across four criteria: Relevance to the workshop theme, Plausibility

of claims, Novelty of contribution, and Clarity of writing, each scored on a 1–5 scale. The two systems agreed on 8 of 10 submissions (80% exact agreement; 90% including partial agreement). The one substantive disagreement — on a submission that was ultimately rejected for insufficient regional focus — was resolved through human editorial review.

Critically, and unlike the Agents4Science model where human expert review was reserved only for the top-scoring papers, every submission at NortheastGenAI 2026 received a human editorial check by the workshop organizing committee before any decision was made. No acceptance or rejection was issued on the basis of AI review alone. The human editor had no prior relationship with any of the submitting authors, and no conflicts of interest were identified.

This design choice was deliberate. Bianchi et al. [1] documented notable sycophancy in LLM-generated reviews, particularly from Gemini 2.5 Pro, which gave scores that diverged substantially from human expert assessments. They also found that approximately 56% of submissions contained at least one hallucinated reference. These findings reinforce the necessity of mandatory human oversight rather than optional human review. At NortheastGenAI 2026, AI reviews served as structured first-pass assessments; the final decision on every paper rested with the workshop chairs and program committee.

All reviews, meta-reviews, and decisions are publicly available on OpenReview at `openreview.net/group?id=NortheastGenAI/2026/Workshop`. This transparency — making AI reviews, human editorial checks, and final decisions all publicly visible — is a deliberate design choice. NortheastGenAI positions itself as an open experiment in AI-assisted regional research, not a claim about AI research quality.

4 Submissions and Acceptance

The workshop received 10 submissions in total. One submission was desk rejected as a duplicate of another submission in the same batch. Of the remaining 9 complete submissions reviewed, 2 were rejected for failing to meet the Northeast India focus requirement and for insufficient empirical grounding. Seven papers were accepted for presentation and inclusion in these proceedings, representing a 78% acceptance rate among complete, non-duplicate submissions. Of the 7 accepted papers, 2 addressed Track T1, 1 addressed Track T2, 3 addressed Track T3, and 1 was cross-track (T1 and T3).

5 Accepted Papers

The seven accepted papers span all three workshop tracks and represent contributions from researchers across India and internationally.

5.1 Track T1: Language, Culture and Heritage

Doley [2] examines Mising, a UNESCO-classified Definitely Endangered Tibeto-Burman language of Assam's Brahmaputra Valley, tracing its endangerment

challenges and exploring how AI and NLP tools — particularly ASR, LLMs, and digital corpus building — can support sustainable revitalisation. The paper presents a community literary conclave as a successful model of community mobilisation for language documentation.

Jadid and Hussain [3] propose AssamBench, the first systematic, human-annotated benchmark for evaluating LLMs on Assamese across five tasks: sentiment analysis, named entity recognition, machine translation, reading comprehension, and hate speech detection.

5.2 Track T2: Society, History and Anthropology

Khan and Dutta [4] argue that Khasi society in Meghalaya is better understood as matrilineal rather than matriarchal: while women occupy a central position in lineage continuity and household security, formal authority in kin mediation, village governance, and public decision-making remains largely male-dominated.

5.3 Track T3: AI and Technology for NE India

Mall [5] proposes a decentralized AI framework for Northeast India emphasizing multilingual systems, mobile-first access, offline functionality, and community participation, arguing that the region can emerge as a model for inclusive, ethical, and low-resource AI innovation across governance, education, and climate resilience domains.

Khan [6] integrates GIS, multi-temporal satellite imagery, and machine learning to automate spatio-temporal analysis of forest cover and carbon dynamics across Assam, Arunachal Pradesh, Manipur, and Mizoram, demonstrating that AI-enabled geospatial workflows provide a scalable decision-support framework for climate resilience in Northeast India.

Sachar and Shirisha [7] propose the ARGF-NE framework, a six-dimensional AI Readiness Gap Framework for Northeast India covering digital connectivity infrastructure, linguistic data availability, compute access, human capital, policy alignment, and institutional capacity.

5.4 Cross-track: T1 and T3

Shirisha and Sachar [8] draw on UDISE+ 2024-25 school education data and Census 2011 language statistics to present six empirically grounded findings on the relationship between language scheduling status and digital infrastructure access in schools, proposing a three-pillar framework for Language AI, Education, and Climate Resilience in Northeast India.

6 Workshop Program and Awards

NortheastGenAI 2026 was held virtually on May 29, 2026, with the following program in IST:

10:30–10:45 Opening Remarks, MWire Labs

10:45–11:05 Keynote I: Bonaventure F. P. Dossou (McGill University and Mila)
Doing More with Less: Efficient Methods for Low-Resource Languages

11:05–11:25 Keynote II: Dr. Prabhat Kumar Bharti (Bennett University)
CO-REVIEWER: Can AI Review Like a Human? An Agentic Framework for LLM-Human Alignment in Peer Review

11:25–12:35 Paper Presentations (7 papers, 10 minutes each)

12:35–12:50 Closing, Awards, and NortheastGenAI 2027 Announcement

Two awards were presented at the closing session. The Best Paper award was given to Shirisha and Sachar [8] for *Language, Artificial Intelligence, and Digital Equity in North East India*, recognised for its empirically grounded policy analysis and cross-disciplinary contribution. The Best Exploratory Paper award was given to Jadid and Hussain [3] for *AssamBench*, recognised for its practical contribution toward closing the LLM evaluation gap for Assamese.

7 Conclusion and Future Work

NortheastGenAI 2026 set out to ask an open question: can AI research Northeast India? The seven accepted papers in these proceedings offer early, grounded, and transparently produced answers across language documentation, social anthropology, geospatial analysis, and digital equity. Together they position Northeast India not as a passive subject of AI research, but as an active testbed for low-resource, high-diversity AI systems.

The workshop's AI-assisted and human-in-the-loop review model demonstrated that transparent, open peer review is achievable at this scale without sacrificing rigor. The 80% agreement rate between two independent LLM reviewers, combined with mandatory human oversight on every submission, produced decisions that were consistent, defensible, and publicly accountable. The lessons from this first edition will inform the design of NortheastGenAI 2027.

NortheastGenAI 2027 is planned as the second edition of the workshop. We invite researchers, practitioners, and institutions working on Northeast India's languages, cultures, and ecosystems to join the next edition. Details will be announced at `northeastgenai.github.io`.

8 Acknowledgments

NortheastGenAI 2026 was organized by MWire Labs, Shillong, Meghalaya. We thank Cohere Labs for the Catalyst Grant that supported the AI-assisted review infrastructure. We thank our program committee members Dr. Biman Debbarma

(Tripura University), Dr. Arup K Nath (Tezpur University), Dr. Antonios Anastasopoulos (George Mason University), Genevieve Chyrmang (Cotton University), Dr. Kamal Kumar Tanti (MSSV Nagaon), and Anisha Jadhav for their service to this first edition. We thank Walmatchi Gabil Momin and Emmi Grace Wanswett for their tireless coordination work that made the workshop possible. We thank all authors who submitted their work and all attendees who joined us virtually on May 29, 2026. Proceedings are published through Northeast Langchive (northeastlangchive.org) under CC-BY-4.0. ISBN 978-81-997217-0-8.

References

1. Bianchi, F., Queen, O., Thakkar, N., Sun, E., Zou, J.: Exploring the use of AI authors and reviewers at Agents4Science. Nature Biotechnology (2025). https://doi.org/10.1038/s41587-025-02963-8
2. Doley, M.: From Oral Tradition to Digital Future: Documenting the Mising Language through Community Initiatives and AI. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)
3. Jadid, J., Hussain, M.A.: AssamBench: Toward a Comprehensive Benchmark for Evaluating Large Language Models on Assamese. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)
4. Khan, A., Dutta, S.: Is Khasi Society Truly Matriarchal? A Critical Study of Matriliny and Gender Power in Meghalaya. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)
5. Mall, A.: AI and Digital Public Infrastructure for Northeast India: A Low Resource Framework for Inclusive Governance, Education, and Climate Resilience. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)
6. Khan, A.: AI-Assisted Spatio-Temporal Analysis of Forest Cover and Carbon Dynamics in Northeast India: A Remote Sensing and GIS Approach. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)
7. Sachar, B., Shirisha, A.: Beyond Connectivity: A Multi-Dimensional AI Readiness Gap Analysis for Northeast India. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)
8. Shirisha, A., Sachar, B.: Language, Artificial Intelligence, and Digital Equity in North East India: A Quantitative and Policy Analysis of Low-Resource Language Exclusion, Educational Infrastructure, and Climate Knowledge Systems. In: Proceedings of the First NortheastGenAI Workshop 2026 (2026)