BHASHA 2025

1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)

Proceedings of the BHASHA Workshop 2025

December 23, 2025

The BHASHA organizers gratefully acknowledge the support from the following sponsors.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-313-5

Preface

We are extremely happy to bring forth the **1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)** as part of *IJCNLP-AACL 2025* conference held during 20th-24th December, 2025.

India, despite being a linguistically rich country with 22 official languages, does not enjoy the benefits of NLP research according to the potential. The special nature of Indian languages, from being inflectional and agglutinative to having a free word order, does not let direct usage of tools built for other languages. In this context, the BHASHA workshop is conceived to focus on creating tools, benchmarks, resources, annotated corpora, evaluation metrics, etc. for Indian languages.

BHASHA is being held as a full-day workshop on *23rd December, 2025*. The program includes two invited talks, multiple research paper oral and poster presentations. In addition, two shared task competitions were held as part of the BHASHA workshop and papers for those will be presented as posters and demonstrations along with a shared task overview talk.

The program committee consisted of 19 eminent researchers from both academia and industry. A total of 26 papers were submitted, out of which 1 was desk rejected. Of the remaining 25 papers, 11 have been accepted to be part of the proceedings, giving an overall acceptance ratio of 11/26 = 42%. While 8 of these papers are being presented orally, two poster sessions are held where all the 11 posters are presented for longer and better interactions between the authors and the audience. Out of the 11 accepted papers, 8 are from India, while 1 each are from Japan, Canada, and USA.

The BHASHA workshop also featured two *shared tasks*, one on **Grammar Error Correction (IndicGEC)** on 5 Indian languages—Hindi, Bangla, Telugu, Tamil, and Malayalam—and the other on **Word Grouping (IndicWG)** on Hindi. While 14 and 2 teams participated respectively in the two tasks for the final stages, 10 papers were received. Out of these, 6 were accepted for the proceedings. A summary paper on the two shared tasks and the different submissions is also included in the proceedings.

We thank the IJCNLP-AACL workshop chairs for helping us in various stages of the workshop. It is my pleasure to also thank the entire organizing team and the different chairs who played their roles to perfection for the successful conduct of this workshop.

Arnab Bhattacharya
General Chair, BHASHA 2025

Organizing Committee

General Chair

Arnab Bhattacharya, Indian Institute of Technology Kanpur, India

Program Chairs

Pawan Goyal, Indian Institute of Technology Kharagpur, India
Arnab Bhattacharya, Indian Institute of Technology Kanpur, India

Publication Chairs

Pramit Bhattacharyya, Indian Institute of Technology Kanpur, India
Shubham Kumar Nigam, Indian Institute of Technology Kanpur, India

Invited Talk Chairs

Saptarshi Ghosh, Indian Institute of Technology Kharagpur, India
Kripabandhu Ghosh, Indian Institute of Science and Research Kolkata, India

Web Chair

Hrishikesh Terdalkar, Birla Institute of Technology and Science Pilani, Hyderabad Campus, India

Local Chair

Ganesh Ramakrishnan, Indian Institute of Technology Bombay, India

Demonstration Chairs

Karthika N J, Indian Institute of Technology Bombay, India
Manoj Balaji Jagadeeshan, Indian Institute of Technology Kharagpur, India

Organiser

Subinay Adhikary, Indian Institute of Science and Research Kolkata, India

Program Committee

Program Committee

Aditya Maheshwari, Indian Institute of Management Indore, India
Arnab Bhattacharya, Indian Institute of Technology Kanpur, India
Chaitali Dangarikar, Indian Institute of Technology Kanpur, India
Hrishikesh Terdalkar, Birla Institute of Technology and Science Pilani, Hyderabad Campus, India
Jivnesh Sandhan, Kyoto University, Japan
Karthika N J, Indian Institute of Technology Bombay, India
Koustav Rudra, Indian Institute of Technology Kharagpur, India
Kripabandhu Ghosh, Indian Institute of Science and Research Kolkata, India
Mahesh V S D S Akavarapu, Universität Tübingen, Germany
Manish Shrivastava, International Institute of Information Technology Hyderabad, India
Manoj Balaji Jagadeeshan, Indian Institute of Technology Kharagpur, India
Maunendra Sankar Desarkar, Indian Institute of Technology Hyderabad, India
Mounika Marreddy, Goethe University, Germany
Pawan Goyal, Indian Institute of Technology Kharagpur, India
Prajna Upadhyay, Birla Institute of Technology and Science Pilani, Hyderabad Campus, India
Pramit Bhattacharyya, Indian Institute of Technology Kanpur, India
Procheta Sen, University of Liverpool, UK
Rohit Saluja, Indian Institute of Technology Mandi, India
Shubham Kumar Nigam, Indian Institute of Technology Kanpur, India

Invited Speakers

Monojit Choudhury, Mohamed bin Zayed University of Artificial Intelligence, UAE
Kalika Bali, Microsoft Research, India

Keynote Talk

Beyond Data: Rethinking Scale, Adaptation, and Culture in AI

Monojit Choudhury
Mohamed bin Zayed University of Artificial Intelligence, UAE
2025-12-23 09:30:00 – Room: **VMCC, IIT Bombay, India**

Abstract: AI learns from data. Better data — richer, cleaner, more diverse — undeniably yields better models and evaluations. This narrative is familiar, almost axiomatic. Yet, these data-driven scaling approaches face two fundamental challenges. First, no corpus, however vast, can capture the infinite variability of human languages, contexts, and preferences. Second, every act of data creation is also an act of omission; each dataset is a boundary between inclusion and exclusion.

A sustainable path forward cannot simply be the endless accumulation of data, but rather the cultivation of models that can learn from less, adapt on the fly, and transfer understanding across contexts. Evaluation, too, must evolve from assessing isolated competencies to probing a model's capacity for learning and adaptation in novel scenarios.

In this talk, I explore these ideas through the lens of culture. It is nearly impossible to define and capture the endless variations of cultures through datasets. I argue that AI models therefore must be trained for *meta-cultural competency*—the ability to serve in any culture rather than a specific pre-defined culture. I also present novel methodologies for evaluating meta-cultural awareness.

Bio: Monojit Choudhury is a faculty member at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE. His research spans computational linguistics, multilinguality, cultural AI, and responsible AI. He has contributed significantly to understanding linguistic diversity, low-resource language modeling, and the socio-cultural dimensions of AI systems. Prior to MBZUAI, he worked at Microsoft Research, where he led multiple projects on multilingual NLP and cultural intelligence in AI models.

Keynote Talk

By the People, For the People: Community-Based Multilingual and Multicultural Evaluation for Responsible AI

Kalika Bali
Microsoft Research, India
2025-12-23 16:00:00 – Room: **VMCC, IIT Bombay, India**

Abstract: AI evaluation often claims "multilingual" coverage but relies on English-centric benchmarks that miss cultural and linguistic realities. To build truly inclusive systems, we need communities—not just datasets—in the loop. This keynote highlights why participatory evaluation matters: co-defining goals with local stakeholders, creating culturally grounded scenarios beyond translation, and combining human insight with scalable tools.
Drawing on initiatives like *Samiksha* and *DOSA*, this talk demonstrates how community-driven approaches uncover hidden biases, improve trust, and align AI with lived experiences of the Global Majority.
The talk concludes with practical models for collaboration between researchers, industry, and civil society to make evaluation democratic, accountable, and impactful.

Bio: Kalika Bali is a Senior Principal Researcher at Microsoft Research India and a prolific researcher working across AI, NLP, Speech Technology, and Technology for Empowerment. Her work focuses on multilingual and multicultural technology, particularly for low-resource language communities, including Indian languages. She also works at the intersection of gender and technology, advocating for holistic approaches to mitigating gender bias in technology and foundational GenAI models. Her deep passion for advancing NLP and speech technologies for Indian languages, among other research areas, is reflected in her publications at top-tier NLP venues. In 2023, Dr. Bali was featured on the inaugural *TIME100 AI* list for her transformative contributions to AI and her commitment to building responsible and inclusive AI technologies.

Table of Contents