LM4UC 2025

The 1st Workshop on
Language Models for Underserved Communities

Proceedings of the Workshop

May 4, 2025

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 979-8-89176-242-8

Introduction

We are delighted to welcome you to the NAACL 2025 Workshop on Language Models for Underserved Communities (LM4UC), held in Albuquerque, New Mexico, on May 4, 2025. This workshop aims to address the persistent gaps in natural language processing (NLP) technologies for underserved communities, ensuring equitable and culturally sensitive advancements in artificial intelligence (AI).

The rapid advancement of natural language processing technologies has unlocked transformative opportunities across numerous domains, from communication to knowledge preservation. However, these benefits are not equitably distributed, leaving many underserved communities—speakers of Indigenous languages, regional dialects, and minority languages spoken by smaller populations—without adequate access to these innovations. This disparity stems from multiple factors, including limited linguistic data, insufficient computational resources, and a lack of commercial prioritization. Such languages, which include examples like Yoruba, Igbo, Native American languages, and dialects in multilingual nations such as India and Indonesia, are often both low-resource and underserved, compounded by challenges in AI governance and cultural representation. To address these inequities, the LM4UC initiative aims to foster rigorous research and dialogue centered on three critical pillars:

- **AI Governance**: Establishing robust legal and ethical frameworks to ensure fairness, transparency, and data sovereignty in the development and deployment of language models.
- **Cultural NLP**: Designing models that preserve linguistic diversity and accurately reflect cultural nuances, safeguarding unique heritage and values embedded in language.
- **Sustainable NLP**: Developing efficient, scalable models optimized for low-resource environments, aligning with environmental sustainability and accessibility goals.

This year, we received numerous high-quality submissions addressing a broad spectrum of topics, including democratizing AI access, preserving linguistic diversity, encoding cultural norms, and building efficient language models. We appreciate the dedication of the authors, reviewers, and program committee members in maintaining the scientific rigor and diversity of perspectives that enrich this workshop.

Invited Speakers

We are privileged to host an exceptional group of invited speakers, featuring Timothy Baldwin from MBZUAI, Timnit Gebru from Google, Pratyusha Ria Kalluri from Stanford, David Ifeoluwa Adelani from McGill, and Genta Indra Winata from Capital One. Their presentations will provide valuable insights into the pressing challenges and pioneering solutions within the field of natural language processing, with a particular focus on addressing the needs of underserved communities.

Acknowledgments

We extend our gratitude to the speakers and organizing committee for their unwavering commitment in making this workshop possible. Special thanks also go to our sponsors and supporters for their invaluable contributions. We hope this workshop serves as a platform for vibrant discussions, meaningful collaborations, and impactful research that advances the inclusivity of language technologies worldwide. Thank you for joining us at LM4UC 2025—we look forward to an engaging and productive event.

Sincerely,
The LM4UC Workshop Organizers

Organizing Committee

Workshop Chairs

Sang Truong, Stanford University, USA
Rifki Afina Putri, Korea Advanced Institute of Science & Technology, Korea
Duc Nguyen, HCM University of Technology - VNU-HCM, Vietnam
Angelina Wang, Stanford University, USA
Daniel Ho, Stanford University, USA
Alice Oh, Korea Advanced Institute of Science & Technology, Korea
Sanmi Koyejo, Stanford University, USA

Table of Contents

Program

Sunday, May 4, 2025

09:00 - 09:30 *(In-person) Opening Remarks: Alice Oh*

09:30 - 10:00 *(In-person) Keynote 1: David Ifeoluwa Adelani*

10:00 - 10:30 *(Virtual) Keynote 2: Timnit Gebru*

10:30 - 11:00 *(Hybrid) Structured Networking Event + Tea Break*

11:00 - 11:30 *(In-person) Keynote 3: Genta Indra Winata*

11:30 - 12:00 *(Virtual) Keynote 4: Timothy Baldwin*

12:00 - 12:30 *(Virtual) Keynote 5: Pratyusha Ria Kalluri*

12:30 - 13:00 *(Hybrid) Structured Networking Event + Lunch Break*

13:00 - 13:50 *(Hybrid) Panel Discussion by Angelina Wang*

13:50 - 15:30 *(Hybrid) Student Oral Presentation*

ABDUL: A New Approach to Build Language Models for Dialects Using Formal Language Corpora Only
Yassine Toughrai, Kamel Smaïli and David Langlois

Untangling the Influence of Typology, Data, and Model Architecture on Ranking Transfer Languages for Cross-Lingual POS Tagging
Enora Rice, Ali Marashian, Hannah Haynie, Katharina Wense and Alexis Palmer

Caption Generation in Cultural Heritage: Crowdsourced Data and Tuning Multimodal Large Language Models
Artem Reshetnikov and Maria-Cristina Marinescu

Preserving Cultural Identity with Context-Aware Translation Through Multi-Agent AI Systems
Mahfuz Anik, Abdur Rahman, Azmine Wasi and Md Ahsan

Sunday, May 4, 2025 (continued)

	Enhancing Small Language Models for Cross-Lingual Generalized Zero-Shot Classification with Soft Prompt Tuning Fred Philippy, Siwen Guo, Cedric Lothritz, Jacques Klein and Tegawendé Bissyandé
	Cognate and Contact-Induced Transfer Learning for Hamshentsnag: A Low-Resource and Endangered Language Onur Keleş, Baran Günay and Berat Doğan
	On Tables with Numbers, with Numbers Konstantinos Kogkalidis and Stergios Chatzikyriakidis
	Direct Preference Optimization with Unobserved Preference Heterogeneity Keertana Chidambaram, Karthik Seetharaman and Vasilis Syrgkanis
15:30 - 16:30	*(Hybrid) Poster Session*
16:30 - 17:00	*(Virtual) Awards Ceremony and Closing Remarks: Sanmi Koyejo*