AfricaNLP 2025

Sixth Workshop on African Natural Language Processing (AfricaNLP 2025): Multilingual and Multicultural-aware LLMs

Proceedings of the Workshop

July 31, 2025

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

Sponsored By

Google DeepMind

DAIR

let us

build together.

Meta

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 979-8-89176-257-2

Introduction

We are pleased to present the proceedings of the Sixth Workshop on African Language Processing (AfricaNLP 2025), held on July 31st, 2025 in Vienna, Austria. The theme for this year's workshop is "Multilingual and Multicultural-aware LLMs," reflecting the need for language technology to be tailored to all users, especially those on the African continent.

These proceedings are the first archival proceedings of the AfricaNLP workshop, and this is the first time the workshop has been held at ACL. We accepted approximately 60% of accepted papers, reflecting our desire to balance inclusion and selectivity. In addition to the 28 archival papers that appear in the proceedings, 7 non-archival papers were also presented at the workshop.

We would like to thank the sponsors of the workshop for their generous support: Google DeepMind, Apple, Distributed AI Research Institute (DAIR), Masakhane, and Meta.

With gratitude,
The AfricaNLP 2025 Organizers

Organizing Committee

General Chair

David Ifeoluwa Adelani, McGill University and Mila

Program and Publication Chair

Constantine Lignos, Brandeis University

Sponsorship Chairs

Henok Ademtew, Vella AI
Shamsuddeen Muhammad, Imperial College London
Clemencia Siro, University of Amsterdam

Mentoring and Communications Chair

Everlyn Asiko Chimoto, University of Cape Town and Lelapa AI

Local and Virtual Chair

Israel Abebe Azime, Saarland University

Organizers

Aremu Anuoluwapo, Lelapa AI
Happy Buzaaba, Princeton University
Rooweither Mabuya, SADiLaR
Andiswa Bukula, SADiLaR
Bonaventure F. P. Dossou, McGill University and Mila
Mmbasibidi Setaka, SADiLaR
Idris Abdulmumin, University of Pretoria

Program Committee

Reviewers

Idris Abdulmumin, David Ifeoluwa Adelani, Henok Biadglign Ademtew, Simbiat Ajao, Emmanuel Akanji, Bunmi Akinremi, Jesujoba Oluwadara Alabi, Felermino D. M. A. Ali, Victor Jotham Ashioya, Busayo Awobade

Edward Bayes, Tadesse Destaw Belay, Happy Buzaaba

Yonas Chanie, Emmanuel Kigen Chesire

Sudhansu Bala Das, Muhammad Umar Diginsa, Emmanuel Dorley, Bonaventure F. P. Dossou

Khalid Elmadani, Naome A Etori

Elodie Gauthier, Gideon George, Agam Goyal, David Guzmán, Tajuddeen Gwadabe

Cari Beth Head

Raphael Iyamu

Sandeep Kumar Jha, Adejumobi Monjolaoluwa Joshua

Sulaiman Kagumire, Börje F. Karlsson, Aditi Khandelwal, Alfred Malengo Kondoro, Sujay S Kumar

Melaku Lake, Sven Lampe, Eric Le Ferrand, En-Shiun Annie Lee, Senyu Li, Weiran Lin

Dunstan Matekenya, Evans Gesura Mecha, Francois Meyer, Anjishnu Mukherjee, Elie Mulamba, Raghavan Muthuregunathan

Abdou Mohamed Naira, Antony Ndolo, Mulubrhan Abebe Nerea, Gebregziabihier Nigusie

Perez Ogayo, Kelechi Ogueji, Odunayo Ogundepo, Jessica Ojo, Ifeoma Okoh, Akintunde Oladipo, Flora Oladipupo, Jeffrey Otoibhi

Chester Palen-Michel, Ted Pedersen, Van-Thuy Phi

Stephen D. Richardson, Nathaniel Romney Robinson

Elizabeth Salesky, Fabian David Schmidt, Tajwaa Scott, Walelign Tewabe Sewunetie, Olamide Shogbamu, Rashidat Damilola Sikiru, Yueqi Song

Jiayi Wang

Seid Muhie Yimam, Hao Yu

Tolúlopé Ògúnrèmí

Invited Talk
Building with Africa: Afrocentric Natural Language Processing

Muhammad Abdul-Mageed
The University of British Columbia

Abstract: Africa's linguistic landscape is one of the richest in the world, with over 2,000 languages and dialects spoken across the continent. This diversity creates a unique environment for innovation in natural language technologies. In this talk, I will describe our collaborative journey to close the technology gap and bring African languages into mainstream NLP research. I will focus on seven key publications—Towards Afrocentric NLP, AfroLID, SERENGETI, Cheetah, Toucan, Sahara, and Voice of a Continent—outlining the goals that drove each project, the obstacles we overcame and the insights we gained along the way. Finally, I will examine the impact that culturally rooted NLP systems can have on African communities, from richer digital communication and the preservation of linguistic heritage to more inclusive and equitable technological innovation.

Bio: Muhammad Abdul-Mageed is the Canada Research Chair in Natural Language Processing and Machine Learning and is an Associate Professor at the University of British Columbia. As director of the UBC Deep Learning & NLP Group, co-director of the SSHRC I Trust Artificial Intelligence partnership and co-lead of the SSHRC Ensuring Full Literacy initiative, his work develops multilingual, multimodal and cross-cultural large-language models that are culturally sensitive, equitable, efficient and socially aware. These models advance applications across speech, language and vision—supporting improved human health, more engaging learning, safer social networking and reduced information overload. His research has been funded by the Gates Foundation (through Clear Global), NSERC, the Canada Foundation for Innovation, with additional support from Google, AMD and Amazon. He has authored over 180 peer-reviewed publications, advised the Government of Canada on generative AI policy, and delivered invited lectures, keynotes and panel presentations in more than 25 countries. His work has been featured in outlets such as MIT Technology Review, The Globe and Mail, Euronews and Libération.

Invited Talk
Mapping Progress in African NLP

Jesujoba Oluwadara Alabi
Saarland University

Abstract: NLP research on African languages is active and growing, even though recent efforts—including work on large language models—have primarily focused on high-resource languages. In the past 5 years, there has been a surge of interest in African NLP, which we recently surveyed. In this talk, I will present key takeaways from that work: where research has been concentrated, and where new efforts are most needed. I will also present our recent efforts to address some of these gaps and future directions: AFRIDOC-MT, a multilingual document-level translation benchmark targeting health and tech domains, and AfriHuBERT, a compact self-supervised speech model designed to help close the speech technology gap for African languages. Overall, these insights and projects showcase the progress made and the path forward to more inclusive and impactful NLP for African languages.

Bio: Jesujoba Oluwadara Alabi is a PhD candidate and researcher at Saarland University, Germany, advised by Prof. Dr. Dietrich Klakow. His research focuses on natural language processing (NLP) for low-resource (African) languages, with interests in machine translation, speech processing, NLP model adaptation, and interpretability of model adaptation methods. He is a member of the Masakhane community and has contributed to several key projects advancing NLP for African languages. Notably, one of his publications received a Best Paper Award (Global Challenges) at COLING 2022 for developing AfroXLMR, a multilingual pre-trained language model for African languages. Other notable awards include an Area Chair Award at IJCNLP-AACL 2023 and Outstanding Paper Award at NAACL 2025.

Invited Talk
Scaling Speech Recognition for African Languages

Joyce Nakatumba-Nabende
Makerere University

Abstract: Automatic speech recognition (ASR) for African languages remains challenging due to limited labeled data and a lack of practical guidance to build effective systems in low-resource settings. Although pretrained models such as Whisper, XLS-R, MMS, and W2v-BERT have improved access, their comparative performance across languages, training scales, and decoding strategies remains understudied. In this talk, I will discuss the evaluation of ASR models on thirteen African languages, fine-tuning each on training subsets. The talk will also cover the assessment of the impact of language model decoding using n-gram models trained on open-source text. Finally, I will delve into a framework and results for evaluation of ASR models beyond WER and CER metrics.

Bio: Dr. Joyce Nakatumba-Nabende is a senior lecturer in the Department of Computer Science at Makerere University and the current director for the Makerere University Center for Artificial Intelligence. She is a research scientist addressing global and African challenges as part of "CoRE-AI" Africa-Europe Clusters of Research Excellence on Innovation and Technology. Dr. Nakatumba-Nabende has worked on research in the development and application of Artificial Intelligence and machine learning models and contributes to sustainable and equitable outcomes in health and agriculture, advancing digital inclusion, and improving African language representation in AI.

Invited Talk
Building Language Technologies for Low-Resourced Languages

Hellina Hailu Nigatu
University of California, Berkeley

Abstract: In recent years, we have seen an increase in the number of languages included in NLP research. Particularly, “low-resource languages” are gaining attention after decades of neglect from mainstream research. While inclusion in NLP research certainly has benefits for speakers of these languages, there are also some risks in how we design and build NLP systems. In this talk, we will first cover background on what low-resourced languages are and what gaps exist in current NLP research when designing language technologies for speakers of these languages. Then, we will take a magnifying lens and look at a pre-processing step performed in Amharic NLP and its impact on monolingual and cross-lingual model performance for Machine Translation. We will end by connecting to literature on technology-facilitated language change and why it is important for us to critically reflect on each stage of the NLP pipeline.

Bio: Hellina Hailu Nigatu is a PhD Candidate at UC Berkeley. She received her BSc from Addis Ababa University in Electrical Engineering and her MSc from UC Berkeley in Computer Science. Her research is at the intersection of HCI, NLP, and AI Ethics, with a specific focus on languages with limited data available online. Hellina studies how current language technology design fails for speakers of these languages and how we can design better, contextual language technologies with users’ needs in mind. Hellina holds fellowships from SIGHPC and FAccT. Her research has won several awards, including the Outstanding Paper Award at EMNLP 2024, the Best Paper Award at Black in AI 2024, and the Research of the Year Award from the Wikimedia Foundation.

Invited Talk
Multilingual Modeling and Evaluation in Llama 4 and Beyond

Sebastian Ruder
Meta

Abstract: Multilingual LLMs have become so powerful that they can be used in real-world conversations in a variety of applications. While this presents many opportunities, it also poses challenges associated with the complexity of natural language. In this talk, I will seek to connect academic research to real-world challenges of multilingual conversational AI. I will first provide an overview of multilinguality in Llama 4, highlighting the importance of evaluation. I will then discuss what it takes to bridge the gap between academic and real-world evaluations. Finally, I will discuss how we can develop models that are useful to speakers in their local context, across the globe and for African languages.

Bio: Sebastian Ruder is a research scientist at Meta based in Berlin, Germany where he is working on multilingual LLMs. Previously, he led the Multilingual team at Cohere and worked as a research scientist at Google DeepMind. He completed his PhD in Natural Language Processing Insight Research Centre for Data Analytics, while working as a research scientist at Dublin-based text analytics startup AYLIEN and studied Computational Linguistics at the University of Heidelberg, Germany and at Trinity College, Dublin.

Table of Contents