LoResLM 2025

The First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)

Proceedings of the Workshop

January 20, 2025

©2025 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-215-2

The workshop is supported in part by CLARIN-UK, funded by the Arts and Humanities Research Council as part of the Infrastructure for Digital Arts and Humanities programme.

Preface

We are pleased to present the proceedings of the first Workshop on Language Models for Low-Resource Languages (LoResLM 2025), co-located at the 31st International Conference on Computational Linguistics (COLING 2025) in Abu Dhabi, United Arab Emirates.

There has been rapid growth in natural language processing (NLP) over the past few years, particularly with the invention of neural language models, such as transformers and large language models, which achieved state-of-the-art results in many tasks with diverse emerging capabilities. However, since the capabilities of language models (LMs) are primarily determined by the characteristics of their pre-trained language corpora, these models tend to be more focused on high-resource languages. They often struggle with low-resource languages, which are estimated to be around 7,000. Despite their worldwide usage, these languages generally receive little research attention and lack sufficient digital data and resources to support NLP tasks. Following this bias towards high-resource languages, which negatively affects a significant portion of the global community, there has been a growing trend in developing and adopting LMs for low-resource languages to promote linguistic fairness. To support and strengthen this movement, we initiated LoResLM this year to provide a forum for researchers to share and discuss their ongoing work on LMs for low-resource languages.

Primarily focusing on developing and evaluating neural language models for low-resource languages, LoResLM 2025 invited submissions on a broad range of topics, including creating corpora, developing benchmarks, building or adapting LMs, and exploring LM applications for low-resource languages. In total, we received 52 submissions, including 40 long papers and 12 short papers. Among these, we accepted 35 papers, including 28 long papers and seven short papers, to appear in the workshop proceedings following the review process.

The accepted papers cover a broad spectrum of low-resource languages spanning eight language families. The majority representation (47.2%) is from the Indo-European family, with contributions across its four first-level/major branches. In total, 28 low-resource languages were focused on in these studies. The papers also represent 13 diverse research areas, with the top three being Language Modelling, Machine Translation and Translation Aids, and Lexical Semantics. We are pleased to see such a wide range of contributions, with the potential to inspire diverse and impactful future research on low-resource languages.

LoResLM 2025 would not be successful without several wonderful people who joined this initiative. First of all, we would like to thank the authors who submitted their work to the workshop, encouraging research in many low-resource languages that span diverse research areas. We are very grateful for the programme committee members who played a crucial role towards this workshop's success with their timely engagement with the review process, providing constructive feedback to help authors improve the quality of their papers to meet the general standards. We are also particularly thankful to Prof Jose Camacho-Collados for accepting our invitation to serve as the keynote speaker, sharing his knowledge and experience, and providing valuable insights to the NLP community. Our sincere appreciation also goes to CLARIN-UK for sponsoring the workshop. We are very grateful to everybody for supporting us to make LoResLM 2025 successful.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage
(LoResLM 2025 Organisers)

https://loreslm.github.io/

Organizing Committee

Hansi Hettiarachchi, Lancaster University, UK
Tharindu Ranasinghe, Lancaster University, UK
Paul Rayson, Lancaster University, UK
Ruslan Mitkov, Lancaster University, UK
Mohamed Gaber, Birmingham City University, UK
Damith Premasiri, Lancaster University, UK
Fiona Anting Tan, National University of Singapore, Singapore
Lasitha Uyangodage, University of Münster, Germany

Program Committee

Gábor Bella, IMT Atlantique, France
Samuel Cahyawijaya, The Hong Kong University of Science and Technology, Hong Kong
Burcu Can, University of Stirling, UK
Çağrı Çöltekin, University of Tübingen, Germany
Raj Dabre, National Institute of Information and Communications Technology, Japan
Vera Danilova, Uppsala University, Sweden
Debashish Das, Birmingham City University, UK
Ona de Gibert, University of Helsinki, Finland
Alphaeus Dmonte, George Mason University, USA
Bonaventure F. P. Dossou, McGill University, Canada
Daan van Esch, Google
Ignatius Ezeani, Lancaster University, UK
Anna Furtado, University of Galway, Ireland
Amal Htait, Aston University, UK
Ali Hürriyetoğlu, Wageningen University & Research, Netherlands
Danka Jokic, University of Belgrade, Serbia
Diptesh Kanojia, University of Surrey, UK
Daisy Lal, Lancaster University, UK
Colin Leong, University of Dayton, USA
Veronika Lipp, Hungarian Research Centre for Linguistics, Hungary
Muhidin Mohamed, Aston University, UK
Farhad Nooralahzadeh, University of Zurich, Switzerland
Rrubaa Panchendrarajan, Queen Mary University of London, UK
Nadeesha Pathirana, Aston University, UK
Alistair Plum, University of Luxembourg, Luxembourg
Nishat Raihan, George Mason University, USA
Omid Rohanian, University of Oxford, UK
Sandaru Seneviratne, Australian National University, Australia
Ravi Shekhar, University of Essex, UK
Archchana Sindhujan, University of Surrey, UK
Claytone Sikasote, University of Cape Town, South Africa
Marjana Prifti Skenduli, University of New York Tirana, Albania
Uthayasanker Thayasivam, University of Moratuwa, Sri Lanka
Taro Watanabe, Nara Institute of Science and Technology, Japan
Edlira Vakaj, Birmingham City University, UK
John Vidler, Lancaster University, UK
Phil Weber, Aston University, UK
Bryan Wilie, Hong Kong University of Science & Technology, Hong Kong
Artūrs Znotiņš, University of Latvia, Latvia

Table of Contents

Overview of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)
Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan and Lasitha Randunu Chandrakantha Uyangodage 1

Atlas-Chat: Adapting Large Language Models for Low-Resource Moroccan Arabic Dialect
Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine ABBAHADDOU, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis and Eric Xing 9

Empowering Persian LLMs for Instruction Following: A Novel Dataset and Training Approach
Hojjat Mokhtarabadi, Ziba Zamani, Abbas Maazallahi and Mohammad Hossein Manshaei 31

BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis
Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain and Abu Raihan Mostofa Kamal 68

Using Language Models for assessment of users' satisfaction with their partner in Persian
Zahra Habibzadeh and Masoud Asadpour 78

Enhancing Plagiarism Detection in Marathi with a Weighted Ensemble of TF-IDF and BERT Embeddings for Low-Resource Language Processing
Atharva Mutsaddi and Aditya Prashant Choudhary 89

Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa
Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad and Devon Jarvis 101

Automated Collection of Evaluation Dataset for Semantic Search in Low-Resource Domain Language
Anastasia Zhukova, Christian E. Matt and Bela Gipp 112

Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia
Lance Calvin Lim Gamboa and Mark Lee 123

Exploiting Word Sense Disambiguation in Large Language Models for Machine Translation
Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka and Masao Utiyama 135

Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek
Maciej Rapacz and Aleksander Smywiński-Pohl 145

Language verY Rare for All
Ibrahim Merad, Amos Wolf, Ziad Mazzawi and Yannick Léo 166

Improving LLM Abilities in Idiomatic Translation
Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu and Sean O'Brien 175

A Comparative Study of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters
Yifan Liu, Gelila Tilahun, Xinxiang Gao, Qianfeng Wen and Michael Gervers 182

Bridging Literacy Gaps in African Informal Business Management with Low-Resource Conversational Agents
Maimouna Ouattara, Abdoul Kader Kaboré, Jacques Klein and Tegawendé F. Bissyandé 193

Social Bias in Large Language Models For Bangla: An Empirical Study on Gender and Religious Bias
Jayanta Sadhu, Maneesha Rani Saha and Rifat Shahriyar 204

Extracting General-use Transformers for Low-resource Languages via Knowledge Distillation
Jan Christian Blaise Cruz 219

Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models
Sina Bagheri Nezhad, Ameeta Agrawal and Rhitabrat Pokharel 225

BabyLMs for isiXhosa: Data-Efficient Language Modelling in a Low-Resource Context
Alexis Matzopoulos, Charl Hendriks, Hishaam Mahomed and Francois Meyer 240

Mapping Cross-Lingual Sentence Representations for Low-Resource Language Pairs Using Pre-trained Language Models
Tsegaye Misikir Tashu and Andreea Ioana Tudor 249

How to age BERT Well: Continuous Training for Historical Language Adaptation
Anika Harju and Rob van der Goot 258

Exploiting Task Reversibility of DRS Parsing and Generation: Challenges and Insights from a Multilingual Perspective
Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei 268

BBPOS: BERT-based Part-of-Speech Tagging for Uzbek
Latofat Bobojonova, Arofat Akhundjanova, Phil Sidney Ostheimer and Sophie Fellenz 287

When Every Token Counts: Optimal Segmentation for Low-Resource Language Models
Vikrant Dewangan, Bharath Raj S, Garvit Suri and Raghav Sonavane 294

Recent Advancements and Challenges of Turkic Central Asian Language Processing
Yana Veitsman and Mareike Hartmann 309

CaLQuest.PT: Towards the Collection and Evaluation of Natural Causal Ladder Questions in Portuguese for AI Agents
Uriel Anderson Lasheras and Vladia Pinheiro 325

PersianMCQ-Instruct: A Comprehensive Resource for Generating Multiple-Choice Questions in Persian
Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini and Marco Gori 344

Stop Jostling: Adaptive Negative Sampling Reduces the Marginalization of Low-Resource Language Tokens by Cross-Entropy Loss
Galim Turumtaev 373

Towards Inclusive Arabic LLMs: A Culturally Aligned Benchmark in Arabic Large Language Model Evaluation
Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader and Anis Koubaa 387

Controlled Evaluation of Syntactic Knowledge in Multilingual Language Models
Daria Kryvosheieva and Roger Levy 402

Evaluating Large Language Models for In-Context Learning of Linguistic Patterns In Unseen Low Resource Languages
Hongpu Zhu, Yuqi Liang, Wenjing Xu and Hongzhi Xu 414

Next-Level Cantonese-to-Mandarin Translation: Fine-Tuning and Post-Processing with LLMs
Yuqian Dai, Chun Fai Chan, Ying Ki Wong and Tsz Ho Pun 427

When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages
Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan and Shenbin Qian 437

Program

Monday, January 20, 2025

08:45–09:00 ***Opening Remarks***

09:00–10:00 **Invited Talk: Jose Camacho-Collados (Cardiff University)**

10:00–10:30 **Session 1: Language Modelling**

10:00–10:15 *Atlas-Chat: Adapting Large Language Models for Low-Resource Moroccan Arabic Dialect*
Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine ABBAHADDOU, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis and Eric Xing

10:15–10:30 *Empowering Persian LLMs for Instruction Following: A Novel Dataset and Training Approach*
Hojjat Mokhtarabadi, Ziba Zamani, Abbas Maazallahi and Mohammad Hossein Manshaei

10:30–11:00 ***Coffee Break***

11:00–12:00 **Poster Session 1: Language Model Applications/ Sentiment Analysis/ Machine Translation**

BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis
Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain and Abu Raihan Mostofa Kamal

Using Language Models for assessment of users' satisfaction with their partner in Persian
Zahra Habibzadeh and Masoud Asadpour

Enhancing Plagiarism Detection in Marathi with a Weighted Ensemble of TF-IDF and BERT Embeddings for Low-Resource Language Processing
Atharva Mutsaddi and Aditya Prashant Choudhary

Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa
Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad and Devon Jarvis

Automated Collection of Evaluation Dataset for Semantic Search in Low-Resource Domain Language
Anastasia Zhukova, Christian E. Matt and Bela Gipp

Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia
Lance Calvin Lim Gamboa and Mark Lee

Does Machine Translation Impact Offensive Language Identification? The Case of Indo-Aryan Languages
Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe and Marcos Zampieri

Exploiting Word Sense Disambiguation in Large Language Models for Machine Translation
Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka and Masao Utiyama

Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek
Maciej Rapacz and Aleksander Smywiński-Pohl

Language verY Rare for All
Ibrahim Merad, Amos Wolf, Ziad Mazzawi and Yannick Léo

Improving LLM Abilities in Idiomatic Translation
Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu and Sean O'Brien

12:00–13:00 Session 2: Language Model Applications

12:00–12:15 *A Comparative Study of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters*
Yifan Liu, Gelila Tilahun, Xinxiang Gao, Qianfeng Wen and Michael Gervers

12:15–12:30 *From Arabic Text to Puzzles: LLM-Driven Development of Arabic Educational Crosswords*
Kamyar Zeinalipour, Moahmmad Saad, Marco Maggini and Marco Gori

12:30–12:45 *Bridging Literacy Gaps in African Informal Business Management with Low-Resource Conversational Agents*
Maimouna Ouattara, Abdoul Kader Kaboré, Jacques Klein and Tegawendé F. Bissyandé

12:45–13:00 *Social Bias in Large Language Models For Bangla: An Empirical Study on Gender and Religious Bias*
Jayanta Sadhu, Maneesha Rani Saha and Rifat Shahriyar

13:00–14:00 ***Lunch Break***

Monday, January 20, 2025 (continued)

14:00–15:00 Poster Session 2: Language Modelling/ Linguistic Insights, Parsing and Semantic Tagging with Language Models

Extracting General-use Transformers for Low-resource Languages via Knowledge Distillation
Jan Christian Blaise Cruz

Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models
Sina Bagheri Nezhad, Ameeta Agrawal and Rhitabrat Pokharel

BabyLMs for isiXhosa: Data-Efficient Language Modelling in a Low-Resource Context
Alexis Matzopoulos, Charl Hendriks, Hishaam Mahomed and Francois Meyer

Mapping Cross-Lingual Sentence Representations for Low-Resource Language Pairs Using Pre-trained Language Models
Tsegaye Misikir Tashu and Andreea Ioana Tudor

How to age BERT Well: Continuous Training for Historical Language Adaptation
Anika Harju and Rob van der Goot

Exploiting Task Reversibility of DRS Parsing and Generation: Challenges and Insights from a Multi-lingual Perspective
Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei

BBPOS: BERT-based Part-of-Speech Tagging for Uzbek
Latofat Bobojonova, Arofat Akhundjanova, Phil Sidney Ostheimer and Sophie Fellenz

When Every Token Counts: Optimal Segmentation for Low-Resource Language Models
Vikrant Dewangan, Bharath Raj S, Garvit Suri and Raghav Sonavane

IsiZulu noun classification based on replicating the ensemble approach for Runyankore
Zola Mahlaza, C. Maria Keet, Imaan Sayed and Alexander Van Der Leek

Recent Advancements and Challenges of Turkic Central Asian Language Processing
Yana Veitsman and Mareike Hartmann

15:00–15:30 Session 3: Language Models for Question Answering

15:00–15:15 *CaLQuest.PT: Towards the Collection and Evaluation of Natural Causal Ladder Questions in Portuguese for AI Agents*
Uriel Anderson Lasheras and Vladia Pinheiro

Monday, January 20, 2025 (continued)

15:15–15:30 *PersianMCQ-Instruct: A Comprehensive Resource for Generating Multiple-Choice Questions in Persian*
Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini and Marco Gori

15:30–16:00 ***Coffee Break***

16:00–17:00 **Session 4: Language Modelling and Evaluation**

16:00–16:15 *Stop Jostling: Adaptive Negative Sampling Reduces the Marginalization of Low-Resource Language Tokens by Cross-Entropy Loss*
Galim Turumtaev

16:15–16:30 *Towards Inclusive Arabic LLMs: A Culturally Aligned Benchmark in Arabic Large Language Model Evaluation*
Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader and Anis Koubaa

16:30–16:45 *Controlled Evaluation of Syntactic Knowledge in Multilingual Language Models*
Daria Kryvosheieva and Roger Levy

16:45–17:00 *Evaluating Large Language Models for In-Context Learning of Linguistic Patterns In Unseen Low Resource Languages*
Hongpu Zhu, Yuqi Liang, Wenjing Xu and Hongzhi Xu

17:00–17:30 **Session 5: Machine Translation with Language Models**

17:00–17:15 *Next-Level Cantonese-to-Mandarin Translation: Fine-Tuning and Post-Processing with LLMs*
Yuqian Dai, Chun Fai Chan, Ying Ki Wong and Tsz Ho Pun

17:15–17:30 *When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages*
Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan and Shenbin Qian

17:30–18:00 ***Awards and Closing Remarks***