

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

Abu Dhabi, United Arab Emirates
20 January 2025

ISBN: 979-8-3313-1370-8

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Overview of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)</i> Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan and Lasitha Randunu Chandrantha Uyangodage	1
<i>Atlas-Chat: Adapting Large Language Models for Low-Resource Moroccan Arabic Dialect</i> 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
<i>Empowering Persian LLMs for Instruction Following: A Novel Dataset and Training Approach</i> Hojjat Mokhtarabadi, Ziba Zamani, Abbas Maazallahi and Mohammad Hossein Manshaei	31
<i>BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis</i> Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain and Abu Raihan Mostofa Kamal	68
<i>Using Language Models for assessment of users' satisfaction with their partner in Persian</i> Zahra Habibzadeh and Masoud Asadpour	78
<i>Enhancing Plagiarism Detection in Marathi with a Weighted Ensemble of TF-IDF and BERT Embeddings for Low-Resource Language Processing</i> Atharva Mutsaddi and Aditya Prashant Choudhary	89
<i>Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa</i> Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad and Devon Jarvis	101
<i>Automated Collection of Evaluation Dataset for Semantic Search in Low-Resource Domain Language</i> Anastasia Zhukova, Christian E. Matt and Bela Gipp	112
<i>Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia</i> Lance Calvin Lim Gamboa and Mark Lee	123
<i>Exploiting Word Sense Disambiguation in Large Language Models for Machine Translation</i> Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka and Masao Utiyama	135
<i>Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek</i> Maciej Rapacz and Aleksander Smywiński-Pohl	145
<i>Language very Rare for All</i> Ibrahim Merad, Amos Wolf, Ziad Mazzawi and Yannick Léo	166
<i>Improving LLM Abilities in Idiomatic Translation</i> Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu and Sean O'Brien	175
<i>A Comparative Study of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters</i> Yifan Liu, Gelila Tilahun, Xinxiang Gao, Qianfeng Wen and Michael Gervers	182

<i>Bridging Literacy Gaps in African Informal Business Management with Low-Resource Conversational Agents</i>	
Maimouna Ouattara, Abdoul Kader Kaboré, Jacques Klein and Tegawendé F. Bissyandé	193
<i>Social Bias in Large Language Models For Bangla: An Empirical Study on Gender and Religious Bias</i>	
Jayanta Sadhu, Maneesha Rani Saha and Rifat Shahriyar	204
<i>Extracting General-use Transformers for Low-resource Languages via Knowledge Distillation</i>	
Jan Christian Blaise Cruz	219
<i>Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models</i>	
Sina Bagheri Nezhad, Ameeta Agrawal and Rhitabrat Pokharel	225
<i>BabyLMs for isiXhosa: Data-Efficient Language Modelling in a Low-Resource Context</i>	
Alexis Matzopoulos, Charl Hendriks, Hishaam Mahomed and Francois Meyer	240
<i>Mapping Cross-Lingual Sentence Representations for Low-Resource Language Pairs Using Pre-trained Language Models</i>	
Tsegaye Misikir Tashu and Andreea Ioana Tudor	249
<i>How to age BERT Well: Continuous Training for Historical Language Adaptation</i>	
Anika Harju and Rob van der Goot	258
<i>Exploiting Task Reversibility of DRS Parsing and Generation: Challenges and Insights from a Multi-lingual Perspective</i>	
Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei	268
<i>BBPOS: BERT-based Part-of-Speech Tagging for Uzbek</i>	
Latofat Bobojonova, Arofat Akhundjanova, Phil Sidney Ostheimer and Sophie Fellenz	287
<i>When Every Token Counts: Optimal Segmentation for Low-Resource Language Models</i>	
Vikrant Dewangan, Bharath Raj S, Garvit Suri and Raghav Sonavane	294
<i>Recent Advancements and Challenges of Turkic Central Asian Language Processing</i>	
Yana Veitsman and Mareike Hartmann	309
<i>CaLQuest.PT: Towards the Collection and Evaluation of Natural Causal Ladder Questions in Portuguese for AI Agents</i>	
Uriel Anderson Lasheras and Vladia Pinheiro	325
<i>PersianMCQ-Instruct: A Comprehensive Resource for Generating Multiple-Choice Questions in Persian</i>	
Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini and Marco Gori	344
<i>Stop Jostling: Adaptive Negative Sampling Reduces the Marginalization of Low-Resource Language Tokens by Cross-Entropy Loss</i>	
Galim Turumtaev	373
<i>Towards Inclusive Arabic LLMs: A Culturally Aligned Benchmark in Arabic Large Language Model Evaluation</i>	
Omer Nacar, Serry Taiseer Sibae, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Al-habashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader and Anis Koubaa	387

<i>Controlled Evaluation of Syntactic Knowledge in Multilingual Language Models</i>	
Daria Kryvosheieva and Roger Levy	402
<i>Evaluating Large Language Models for In-Context Learning of Linguistic Patterns In Unseen Low Resource Languages</i>	
Hongpu Zhu, Yuqi Liang, Wenjing Xu and Hongzhi Xu	414
<i>Next-Level Cantonese-to-Mandarin Translation: Fine-Tuning and Post-Processing with LLMs</i>	
Yuqian Dai, Chun Fai Chan, Ying Ki Wong and Tsz Ho Pun	427
<i>When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages</i>	
Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan and Shenbin Qian	437