

Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)

Rabat, Morocco
29 March 2026

ISBN: 979-8-3313-3463-5

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© (2026) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2026)

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>NE-BERT: A Multilingual Language Model for Nine Northeast Indian Languages</i> Badal Nyalang	1
<i>Do Tokenizers Fail on Informal Hindi Expressions? Evidence from Static, Downstream, and Robustness Analyses</i> Manikandan Ravikiran, Tanmay Tiwari, Vibhu Gupta, Rakesh Prakash, Rohit Saluja and Shayan Mohanty	13
<i>Competence Collapse in Code-Mixed Generation: Spectral Evidence and Mechanistic Recovery via Cross-Lingual Activation Steering</i> Tanushree Ravindra Pratap Yadav	29
<i>When Multilingual Evaluation Assumptions Fail: Tokenization Effects Across Scripts</i> Manodyna K H and Luc De Nardi	41
<i>Making Large Language Models Speak Tulu: Structured Prompting for an Extremely Low-Resource Language</i> Prathamesh Devadiga and Paras Chopra	50
<i>To make someone do something: mining alert-style directives in Bulgarian social media for low-resource language modelling</i> Ruslana Margova and Stanislav Penkov	62
<i>BanglaLlama: LLaMA for Bangla Language</i> Abdullah Khan Zehady, Shubhashis Roy Dipta, Naymul Islam, Safi Al Mamun and Santu Karma-ker	73
<i>Evaluating Retrieval-Augmented Generation for Medication Question Answering on Nigerian Drug Labels in Yorùbá</i> Zainab Tairu and Aramide Adebessin	90
<i>Grammatical Error Correction for Low-Resource Languages: The Case of Zarma</i> Mamadou K. Keita, Marcos Zampieri, Christopher M Homan, Adwoa Asantewaa Bremang, Den- nis Asamoah Owusu and Huy Le	98
<i>Quantifying Cross-Lingual Interference: Algorithmic Standardization of Kamtapuri in Large Language Models</i> Roumak Das	110
<i>SinhaLegal: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts</i> Minduli Lasandi and Nevidu Jayatilleke	114
<i>BanglaIPA: Towards Robust Text-to-IPA Transcription with Contextual Rewriting in Bengali</i> Jakir Hasan, Shrestha Datta, Md Saiful Islam, Shubhashis Roy Dipta and Ameya Debnath ...	132
<i>Improving Romanian LLM Pretraining Data using Diversity and Quality Filtering</i> Vlad-Andrei Negoită, Mihai Masala and Traian Rebedea	140
<i>Tone in Yoruba ASR: Evaluating the Impact of Tone Recognition on Transformer-Based ASR Models</i> Joy Olusanya	149

<i>LLM-as-a-Judge for Low-Resource Languages: Adapting Ragas and Comparative Ranking for Romanian</i>	
Claudiu Creanga and Liviu P Dinu	157
<i>QARI: Neural Architecture for Urdu Extractive Machine Reading Comprehension</i>	
Samreen Kazi and Shakeel Ahmed Khoja	168
<i>When LLMs Annotate: Reliability Challenges in Low-Resource NLI</i>	
Solmaz Panahi, John Kelleher and Vasudevan Nedumpozhimana	178
<i>Qomhrá: A Bilingual Irish and English Large Language Model</i>	
Joseph McInerney, Khanh-Tung Tran, Liam Loneragan, Neasa Ní Chiaráin, Ailbhe Ni Chasaide and Barry Devereux	189
<i>Anchoring the Judge: Curriculum-Based Adaptation and Reference-Anchored MQM for LLM-Based Machine Translation of an Unseen Low-Resource Language - A Case of Nupe</i>	
Umar Baba Umar, Sulaimon Adebayo Bashir and Abdulmalik Danlami Mohammed	200
<i>TeluguEval: A Comprehensive Benchmark for Evaluating LLM Capabilities in Telugu</i>	
Revanth Kumar Gundam and Radhika Mamidi	212
<i>Cross-Lingual Emotion Recognition in Balinese Text using Multilingual-LLMs under Peer-Collaborations Settings</i>	
Putu Kussa Laksana Utama, Tsegaye Misikir Tashu and Jilles Steeve Dibangoye	225
<i>Enabling Structured Reasoning in Sindhi with Culturally Grounded Instruction Tuning</i>	
Mehak Mehak, Kamyar Zeinalipour, Pireh Soomro, Cristiano Chesi, Marco Gori and Marco Maggini	239
<i>Targeted Syntactic Evaluation of Language Models on Georgian Case Alignment</i>	
Daniel Gallagher and Gerhard Heyer	259
<i>So, How Much Do LLMs Hallucinate on Low-Resource Languages? A Quantitative and Qualitative Analysis</i>	
Kushal Trivedi, Murtuza Shaikh and Sriyansh Sharma	271
<i>Learning from Scarcity: Building and Benchmarking Speech Technology for Sukuma.</i>	
Macton Mgonzo, Kezia Oketch, Naome A Etori, Winnie Mang’eni, Elizabeth Fabian Nyaki and Michael Samwel Mollel	288
<i>We Are (Language) Family”: Adapting Transformer models to related minority languages with linguistic data</i>	
Miguel López-Otal and Jorge Gracia	297
<i>A Comprehensive Evaluation of Chain-of-Thought Faithfulness in Persian Classification Tasks</i>	
Shakib Yazdani, Cristina España-Bonet, Eleftherios Avramidis, Yasser Hamidullah and Josef Van Genabith	311
<i>Under-resourced studies of under-resourced languages: lemmatization and POS-tagging with LLM annotators for historical Armenian, Georgian, Greek and Syriac</i>	
Chahan Vidal-Gorène, Bastien Kindt and Florian Cafiero	324
<i>Large Language Models for Mental Health: A Multilingual Evaluation</i>	
Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor and Marcos Zampieri	335

<i>Serbian SuperGLUE: Towards an Evaluation Benchmark for South Slavic Language Models</i>	
Mitar Perovic and Teodora Mihajlov	347
<i>Less is More: Adapting Text Embeddings for Low-Resource Languages with Small Scale Noisy Synthetic Data</i>	
Zaruhi Navasardyan, bagratuni@metric.am bagratuni@metric.am, Spartak Bughdaryan and Hrant Davtyan	362
<i>Evaluating Large Language Models on Lithuanian Grammatical Cases</i>	
Urtė Jakubauskaitė and Raquel G. Alhama	371
<i>Tracking the evolution of LLM capabilities for Belarusian with OpenAI Evals</i>	
Vladislav Poritski, Oksana Volchek, Maksim Aparovich, Volha Harytskaya and Pavel Smrz .	378
<i>Beyond Many-Shot Translation: Scaling In-Context Demonstrations For Low-Resource Machine Translation</i>	
Luis Frentzen Salim, Esteban Carlin, Alexandre Morinvil, Xi Ai and Lun-Wei Ku	388
<i>Bootstrapping Embeddings for Low Resource Languages</i>	
Merve Basoz, Andrew Horne and Mattia Oppè	408
<i>The Indonesian Religiolect Corpus: Data Curation for Muslim, Protestant, and Catholic Language Varieties</i>	
Dan Sachs	426
<i>Representation-Aware Prompting for Zero-Shot Marathi Text Classification: IPA, Romanization, Repetition</i>	
Van-Hien Tran, Huy Hien Vu, Hideki Tanaka and Masao Utiyama	436
<i>MaiBERT: A Pre-training Corpus and Language Model for Low-Resourced Maithili Language</i>	
Sumit Yadav, Raju Kumar Yadav, Utsav Maskey, Gautam Siddharth Kashyap, Ganesh Gautam and Usman Naseem	444
<i>KyrText: A Multi-Domain Large-Scale Corpus for Kyrgyz Language</i>	
Tilek Chubakov	453
<i>Pretraining and Benchmarking Modern Encoders for Latvian</i>	
Arturs Znotins	461
<i>Escaping the Probability Trap: Mitigating Semantic Drift in Cantonese-Mandarin Translation</i>	
Yuzhi Liang and Fangqi Chen	471
<i>How multilingual are multilingual LLMs? A case study in Northern Sámi-Finnish Translation</i>	
Jonne Sälevä and Constantine Lignos	484
<i>Tokenization and Morphological Fidelity in Uralic NLP: A Cross-Lingual Evaluation</i>	
Nuo Xu and Ahrii Kim	493
<i>Hebrew Diacritics Restoration using Visual Representation</i>	
Yair Elboher and Yuval Pinter	504
<i>Out-Of-Tune rather than Fine-Tuned: How Pre-training, Fine-tuning and Tokenization Affect Semantic Similarity in a Historical, Non-Standardized Domain</i>	
Stella Verkijk and Piek Vossen	515
<i>LuxDiagRC: A Diagnostic Reading Comprehension Corpus for Luxembourgish with Linguistic and Cognitive Annotation Layers</i>	
Christophe Friezas Gonçalves, Salima Lamsiyah and Christoph Schommer	532

<i>Cross-Lingual and Cross-Domain Transfer Learning for POS Tagging in Historical Germanic Low-Resource Languages</i>	
Irene Miani, Sara Stymne and Gregory R. Darwin	542
<i>MTQE.en-he: Machine Translation Quality Estimation for English-Hebrew</i>	
Andy Rosenbaum, Assaf Siani and Ilan Kernerman	559
<i>Tokenization Cost, Retention, and Orthography Robustness for Ladin and Italian Varieties</i>	
Alessio Staffini	570
<i>UrHiOdSynth: A Multilingual Synthetic Corpus for Speech-to-Speech Translation in Low-Resource Indic Languages</i>	
Jamaluddin, Subhankar Panda, Aditya Narendra, Kamanksha Prasad Dubey and Mohammad Na-deem	584
<i>BanglaSummEval: Reference-Free Factual Consistency Evaluation for Bangla Summarization</i>	
Ahmed Rafid, Rumman Adib, Fariya Ahmed, Ajwad Abrar and Mohammed Saidul Islam . . .	595
<i>Parameter-Efficient Quality Estimation via Frozen Recursive Models</i>	
Umar Abubacar, Roman Bauer and Diptesh Kanojia	609
<i>Neural Machine Translation for French–Mooré: Adapting Large Language Models to Low-Resource Languages</i>	
walkercompaore972@gmail.com walkercompaore972@gmail.com, Maimouna Ouattara, Rodri-que Kafando, Tegawendé F. Bissyandé, Abdoul Kader Kabore and aminata.sabane@ujkz.bf amina-ta.sabane@ujkz.bf	615
<i>Contributing to Speech-to-Speech Translation for African Low-Resource Languages : Study of French-Mooré Pair</i>	
Fayçal S. A. Ouedraogo, Maimouna Ouattara, Rodrique Kafando, Abdoul Kader Kabore, amina-ta.sabane@ujkz.bf aminata.sabane@ujkz.bf and Tegawendé F. Bissyandé	623
<i>Domain-Specific Quality Estimation for Machine Translation in Low-Resource Scenarios</i>	
Namrata Bhalchandra Patil Gurav, Akashdeep Ranu, Archchana Sindhujan and Diptesh Kanojia	630
<i>Overview of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)</i>	
Hansi Hettiarachchi, Tharindu Ranasinghe, Alistair Plum, Paul Rayson, Ruslan Mitkov, Mohamed Medhat Gaber, Damith Premasiri, Fiona Anting Tan and Lasitha Uyangodage	651