

MRL 2025

**The 5th Workshop on Multilingual Representation Learning
(MRL 2025)**

Proceedings of the Workshop

November 8-9, 2025

©2025 Association for Computational Linguistics

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 979-8-89176-345-6

Organizing Committee

Workshop Organizers

David Ifeoluwa Adelani
Catherine Arnett
Duygu Ataman
Tyler A. Chang
Hila Gonen
Rahul Raja
Fabian Schmidt
David Stap
Jiayi Wang

Program Committee

Program Chairs

David Ifeoluwa Adelani
Catherine Arnett
Duygu Ataman
Tyler A. Chang
Hila Gonen
Rahul Raja
Fabian David Schmidt
David Stap
Jiayi Wang

Reviewers

Victor Olalekan Akinode
Solomon Oluwole Akinola
Muhammad Arif
Catherine Arnett
Nigus Wereta Asnake
Duygu Ataman
Ali Athar
Fatemeh Azadi
Travis M. Bartley
Vishal Bhalla
Nischal Reddy Chandra
Xupeng Chen
Jiajing Chen
Koel Dutta Chowdhury
Benedikt Ebing
Yassine El Kheir
Gregor Geigle
Tommaso Green
David Guzmán
Yusif Ibrahimov
Ainaz Jamshidi
Gaganpreet Jhajj
Jiby Mariya Jose
Haeji Jung
Haeji Jung
Zhengjian Kang
Yixiao Kang
Hikmat Khan
Christopher Klamm
Prashant Kodali
Hongzhi Kuai
Xuchen Li
Senyu Li
Tomasz Limisiewicz

Tomasz Limisiewicz
Pranita Yogesh Mahajan
Anish Mahishi
Yan Meng
Moseli Mots'oezli
Usman Nawaz
Esther Odunayo Oduntan
Peter Oseghale Ohue
Yewande Ojo
Jessica Ojo
Tobi Olatunji
Ejiro Onose
Udita Patel
Rahul Raja
Manikant Roy
Shaibal Saha
Shubham Shukla
David Stap
Janet Yunchen Sung
Wenjia Tan
Shailja Thakur
Vajratiya Vajrobol
Vajratiya Vajrobol
Arpita Vats
Deepali Verma
Sahil Walia
Azmine Toushik Wasi
Song-Li Wu
Zonghao Ying
Dokyoonyoon Yoon
Hao Yu
Zhehao Zhang
Xufeng Zhao
Ziqi Zhou

Table of Contents

<i>No Language Data Left Behind: A Cross-Cultural Study of CJK Language Datasets in the Hugging Face Ecosystem</i>	
Dasol Choi, Woomyoung Park and Youngsook Song	1
<i>Cross-Document Cross-Lingual NLI via RST-Enhanced Graph Fusion and Interpretability Prediction</i>	
Mengying Yuan, WenHao Wang, Zixuan Wang, Yujie Huang, Kangli Wei, Fei Li, Chong Teng and Donghong Ji	11
<i>Universal Patterns of Grammatical Gender in Multilingual Large Language Models</i>	
Andrea Schröter and Ali Basirat	34
<i>Cross-lingual Transfer Dynamics in BLOOMZ: Insights into Multilingual Generalization</i>	
Sabyasachi Samantaray and Preethi Jyothi	47
<i>CoCo-CoLa: Evaluating and Improving Language Adherence in Multilingual LLMs</i>	
Elnaz Rahmati, Alireza Salkhordeh Ziabari and Morteza Dehghani	62
<i>Understand, Solve and Translate: Bridging the Multilingual Mathematical Reasoning Gap</i>	
Hyunwoo Ko, Guijin Son and Dasol Choi	78
<i>Unlocking LLM Safeguards for Low-Resource Languages via Reasoning and Alignment with Minimal Training Data</i>	
Zhuowei Chen, Bowei Zhang, Nankai Lin, Tian Hou and Lianxi Wang	96
<i>Meta-Pretraining for Zero-Shot Cross-Lingual Named Entity Recognition in Low-Resource Philippine Languages</i>	
David Demitri Africa, Suchir Salhan, Yuval Weiss, Paula Buttery and Richard Diehl Martinez	106
<i>Extended Abstract for Linguistic Universals": Emergent Shared Features in Independent Monolingual Language Models via Sparse Autoencoders</i>	
Ej Zhou and Suchir Salhan	128
<i>The Unreasonable Effectiveness of Model Merging for Cross-Lingual Transfer in LLMs</i>	
Lucas Bandarkar and Nanyun Peng	131
<i>Reassessing Speech Translation for Low-Resource Languages: Do LLMs Redefine the State-of-the-Art Against Cascaded Models?</i>	
Jonah Dauvet, Min Ma, Jessica Ojo and David Ifeoluwa Adelani	149
<i>Quality-Aware Translation Tagging in Multilingual RAG system</i>	
Hoyeon Moon, Byeolhee Kim and Nikhil Verma	161
<i>Improving Language Transfer Capability of Decoder-only Architecture in Multilingual Neural Machine Translation</i>	
Zhi Qu, Yiran Wang, Chenchen Ding, Hideki Tanaka, Masao Utiyama and Taro Watanabe	178
<i>How Can We Relate Language Modeling to Morphology?</i>	
Wessel Poelman, Thomas Bauwens and Miryam de Lhoneux	196
<i>On the Consistency of Multilingual Context Utilization in Retrieval-Augmented Generation</i>	
Jirui Qi, Raquel Fernández and Arianna Bisazza	199
<i>CLIRudit: Cross-Lingual Information Retrieval of Scientific Documents</i>	
Francisco Valentini, Diego Kozłowski and Vincent Larivière	226

<i>TenseLoC: Tense Localization and Control in a Multilingual LLM</i> Ariun-Erdene Tumurchuluun, Yusser Al Ghussin, David Mareček, Josef Van Genabith and Koel Dutta Chowdhury	243
<i>Reversible Disentanglement of Meaning and Language Representations from Multilingual Sentence Encoders</i> Keita Fukushima, Tomoyuki Kajiwara and Takashi Ninomiya.....	265
<i>Alif: Advancing Urdu Large Language Models via Multilingual Synthetic Data Distillation</i> Muhammad Ali Shafique, Kanwal Mehreen, Muhammad Arham, Maaz Amjad, Sabur Butt and Hamza Farooq.....	271
<i>Pragyaan: Designing and Curating High-Quality Cultural Post-Training Datasets for Indian Languages</i> Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri and Shubham Agarwal	285
<i>SOI Matters: Analyzing Multi-Setting Training Dynamics in Pretrained Language Models via Subsets of Interest</i> Shayan Vassef, Amirhossein Dabiriaghdam, Mohammadreza Bakhtiari and Yadollah Yaghoobzadeh.....	322
<i>When Scripts Diverge: Strengthening Low-Resource Neural Machine Translation Through Phonetic Cross-Lingual Transfer</i> Ammon Shurtz, Christian Richardson and Stephen D. Richardson.....	336
<i>Conditions for Catastrophic Forgetting in Multilingual Translation</i> Danni Liu and Jan Niehues.....	347
<i>Monolingual Adapter Networks for Efficient Cross-Lingual Alignment</i> Pulkit Arya.....	360
<i>Culturally-Nuanced Story Generation for Reasoning in Low-Resource Languages: The Case of Javanese and Sundanese</i> Salsabila Zahirah Pranida, Rifo Ahmad Genadi and Fajri Koto.....	369
<i>Zero-Shot Cross-Lingual Transfer using Prefix-Based Adaptation</i> Sneha A, Sayambhu Sen, Piyush Singh Pasi, Abhishek Singhania and Preethi Jyothi	385
<i>Exploring the Role of Transliteration in In-Context Learning for Low-resource Languages Written in Non-Latin Scripts</i> Chunlan Ma, Yihong Liu, Haotian Ye and Hinrich Schuetze.....	397
<i>Type and Complexity Signals in Multilingual Question Representations</i> Robin Kokot and Wessel Poelman	411
<i>Entropy2Vec: Crosslingual Language Modeling Entropy as End-to-End Learnable Language Representations</i> Patrick Amadeus Irawan, Ryandito Diandaru, Belati Jagad Bintang Syuhada, Randy Zakya Suchrady, Alham Fikri Aji, Genta Indra Winata, Fajri Koto and Samuel Cahyawijaya.....	426
<i>Language Surgery in Multilingual Large Language Models</i> Joanito Agili Lopo, Muhammad Ravi Shulthan Habibi, Tack Hwa Wong, Muhammad Ilham Ghazali, Fajri Koto, Genta Indra Winata, Peerat Limkonchotiwat, Alham Fikri Aji and Samuel Cahyawijaya	438

<i>Relevant for the Right Reasons? Investigating Lexical Biases in Zero-Shot and Instruction-Tuned Rerankers</i>	
Yuchen Mao, Barbara Plank and Robert Litschko	468
<i>Cross-Lingual Knowledge Augmentation for Mitigating Generic Overgeneralization in Multilingual Language Models</i>	
Sello Ralethe and Jan Buys	483
<i>What if I ask in alia lingua? Measuring Functional Similarity Across Languages</i>	
Debangana Mishra, Arihant Rastogi, Agyeya Singh Negi, Shashwat Goel and Ponnurangam Kumaraguru	496
<i>Multilingual Learning Strategies in Multilingual Large Language Models</i>	
Ali Basirat	507
<i>Sub-1B Language Models for Low-Resource Languages: Training Strategies and Insights for Basque</i>	
Gorka Urbizu, Ander Corral, Xabier Saralegi and Iñaki San Vicente	519
<i>jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval</i>	
Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang and Han Xiao	531
<i>RoBiologyDataChoiceQA: A Romanian Dataset for improving Biology understanding of Large Language Models</i>	
Dragos-Dumitru Ghinea, Adela-Nicoleta Corbeanu and Marius-Adrian Dumitran	551
<i>Mind the (Language) Gap: Towards Probing Numerical and Cross-Lingual Limits of LVLMS</i>	
Somraj Gautam, Abhirama Subramanyam Penamakuri, Abhishek Bhandari and Gaurav Harit	568
<i>MUG-Eval: A Proxy Evaluation Framework for Multilingual Generation Capabilities in Any Language</i>	
Seyoung Song, Seogyong Jeong, Eunsu Kim, Jiho Jin, Dongkwan Kim, Jamin Shin and Alice Oh	585
<i>Scaling, Simplification, and Adaptation: Lessons from Pretraining on Machine-Translated Text</i>	
Dan John Velasco and Matthew Theodore Roque	612
<i>A Federated Approach to Few-Shot Hate Speech Detection for Marginalized Communities</i>	
Haotian Ye, Axel Wisioerek, Antonis Maronikolakis, Özge Alaçam and Hinrich Schütze	631
<i>Training of LLM-Based List-Wise Multilingual Reranker</i>	
Hao Yu and David Ifeoluwa Adelani	652