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

Hansi Hettiarachchi[1], Tharindu Ranasinghe[1], Paul Rayson[1], Ruslan Mitkov[1]
Mohamed Gaber[2], Damith Premasiri[1], Fiona Anting Tan[3], Lasitha Uyangodage[4]
[1]Lancaster University, UK [2]Birmingham City University, UK
[3]National University of Singapore, Singapore [4]University of Münster, Germany
loreslm2025@gmail.com

Abstract

The first Workshop on Language Models for Low-Resource Languages (LoResLM 2025) was held in conjunction with the 31st International Conference on Computational Linguistics (COLING 2025) in Abu Dhabi, United Arab Emirates. This workshop mainly aimed to provide a forum for researchers to share and discuss their ongoing work on language models (LMs) focusing on low-resource languages, following the recent advancements in neural language models and their linguistic biases towards high-resource languages. LoResLM 2025 attracted notable interest from the natural language processing (NLP) community, resulting in 35 accepted papers from 52 submissions. These contributions cover a broad range of low-resource languages from eight language families and 13 diverse research areas, paving the way for future possibilities and promoting linguistic inclusivity in NLP.

1 Introduction

Language models (LMs) have been a long-standing research topic, originating with simple n-gram models in the 1950s (Shannon, 1951). They are computational models that use the generative likelihood of word sequences to perform natural language processing (NLP) tasks (Zhao et al., 2023). Recent advancements in LMs have significantly shifted towards neural language models due to their more robust capabilities (Zhao et al., 2023; Minaee et al., 2024). Developing pre-trained neural language models/transformers is a key milestone in LM research that notably enhanced NLP performance (Vaswani et al., 2017; Devlin et al., 2019). This breakthrough has also prompted the development of more advanced large language models (LLMs), such as GPT, which consist of vast numbers of parameters pre-trained on extensive text corpora, resulting in state-of-the-art natural language understanding and generation across various applications (Touvron et al., 2023; Jiang et al., 2023).

There are approximately 7,000 spoken languages worldwide (van Esch et al., 2022). However, most NLP research focuses on about 20 languages with high resources (Magueresse et al., 2020). For example, 63% of the papers published at ACL 2008 focused on English (Bender, 2011), and even a decade later, 70% of the papers at ACL 2021 were evaluated only in English (Ruder et al., 2022). The remaining numerous languages that receive little research attention are commonly referred to as low-resource languages. These languages generally lack sufficient digital data and resources to support NLP tasks. They are also known as resource-scarce, resource-poor, less computerised, low-data, or low-density languages (Ranathunga et al., 2023).

Since the capabilities of LMs are primarily determined by the characteristics of their pre-trained language corpora, disparities in language resources are also evident within the models. For instance, many widely used transformer models (e.g., `BERT` (Devlin et al., 2019), `RoBERTa` (Liu et al., 2019), `BART` (Lewis et al., 2020), and `T5` (Raffel et al., 2020)) only support English. However, the cross-lingual capabilities of transformers have paved the way for multilingual models (e.g., `mBERT` (Devlin et al., 2019), `XLM-R` (Conneau et al., 2020), `mT5` (Xue et al., 2021), and `BLOOM` (Scao et al., 2022)), allowing low-resource languages to benefit from other languages through joint learning approaches. Despite this progress, these models are typically limited to up to 100 languages due to the curse of multilingualism (Conneau et al., 2020). In light of this challenge, developing monolingual models (e.g., `SinBERT` for Sinhala (Dhananjaya et al., 2022), and `PhoBERT` for Vietnamese (Nguyen and Tuan Nguyen, 2020)) is another growing trend recently established to promote research in low-resource languages.

There are several common factors which impede low-resource language research. One major issue is limited data availability, as the performance of

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 1–8
January 20, 2025.

most models depends heavily on the amount of training data (Hettiarachchi et al., 2024). Even recent neural LMs with multilingual capabilities tend to perform poorly when pre-training data for a particular language is limited or unseen (Ahuja et al., 2022; Hettiarachchi et al., 2023). Data quality also plays a pivotal role in research outcomes, yet the absence of recommended guidelines hinders the quality of low-resource language data (Lignos et al., 2022). Additionally, the scarcity of benchmark datasets tailored for low-resource languages tends to bias most model evaluations towards high-resource languages (Blasi et al., 2022; Ranasinghe et al., 2024).

Interestingly, there are several ongoing efforts that aim to encourage research on low-resource languages and mitigate the bias in NLP approaches towards high-resource languages (Chakravarthi et al., 2022; Ojha et al., 2023; Melero et al., 2024). We organised the first Workshop on Language Models for Low-Resource Languages (LoResLM 2025) to further strengthen this trend. LoResLM 2025[1] specifically focused on LM-based approaches for low-resource languages, inviting submissions on a broad range of topics, including creating corpora, developing benchmarks, building or adapting LMs, and exploring LM applications for low-resource languages. Section 2 provides a summary of the workshop contributions, highlighting language and task/research area coverage. We invite you to refer to the full papers available in the proceedings for more detailed information.

2 Workshop Contributions

LoResLM 2025 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. We provide a detailed summary of the distribution of accepted papers across various languages and research areas below.

2.1 Languages

As illustrated in Figure 1, the papers accepted to LoResLM 2025 mainly span eight language families. The majority representation is from Indo-European family, while Koreanic, Sino-Tibetan and Isolate language families have equal minority representation. Languages with no relationships with others were considered under the Isolate family.

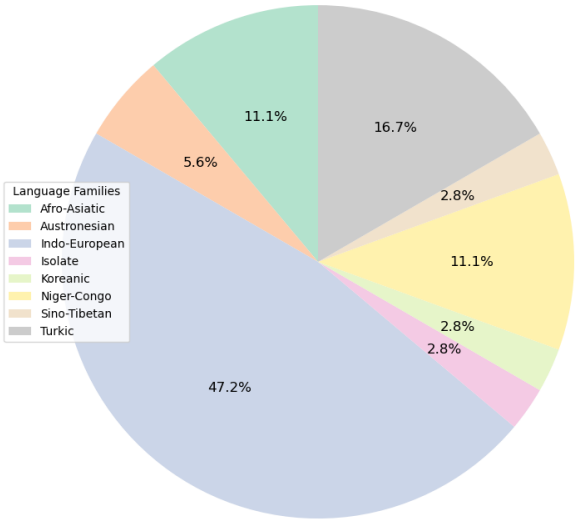

Figure 1: Distribution of workshop contributions across language families

We present a detailed language-level analysis in Table 1. We further divided the Indo-European family into its first branch level for a comprehensive exploration, given its wide contributions. Overall, there were contributions from four distinct branches of the Indo-European language family. During this analysis, we focused exclusively on low-resource languages, excluding high-resource languages involved in comparison studies. However, some languages that would typically classify as high-resource considering the general resource distribution across popular research areas (e.g. Arabic, German, etc.) were considered low-resource in specific contexts where resources are limited, such as particular domains, research areas, or dialects. In total, contributions covered 28 low-resource languages. Additionally, a few papers experimented with multiple languages (more than five) from various language families. These were categorised under 'Multiple' but excluded from the language count given above, as their focus was more on the task level rather than the language level.

2.2 Research Areas

Table 2 shows the distribution of the accepted papers across various NLP research areas. These areas were adopted based on the topics of call for papers from leading NLP conferences in 2024.

Overall, the accepted papers contributed to 13 NLP research areas. As expected, the most popular topic among the accepted papers was *'Language Modelling'* with eleven papers. *'Machine Trans-*

[1]Available at https://loreslm.github.io/

Language Family	Language	Papers
Afro-Asiatic	Arabic	Nacar et al. (2025); Shang et al. (2025); Zeinalipour et al. (2025b)
	Hausa	Sani et al. (2025)
Austronesian	Filipino	Gamboa and Lee (2025)
	Tagalog	Cruz (2025)
Indo-European (Germanic)	German	Zhukova et al. (2025)
	Old English	Harju and van der Goot (2025)
Indo-European (Hellenic)	Ancient Greek	Rapacz and Smywiński-Pohl (2025)
Indo-European (Indo-Iranian)	Bengali	Alam et al. (2025); Sadhu et al. (2025)
	Marathi	Mutsaddi and Choudhary (2025); Dmonte et al. (2025)
	Persian	Habibzadeh and Asadpour (2025); Mokhtarabadi et al. (2025); Zeinalipour et al. (2025a)
	Sinhala	Dmonte et al. (2025)
	Urdu	Amin et al. (2025); Donthi et al. (2025)
Indo-European (Italic)	Italian	Amin et al. (2025)
	Medieval Latin	Liu et al. (2025)
	Monégasque	Merad et al. (2025)
	Portuguese	Lasheras and Pinheiro (2025)
Isolate	Basque	Kryvosheieva and Levy (2025)
Koreanic	Korean	Tran et al. (2025)
Niger-Congo	isiXhosa	Matzopoulos et al. (2025)
	IsiZulu	Mahlaza et al. (2025)
	Mooré	Ouattara et al. (2025)
	Swahili	Kryvosheieva and Levy (2025)
Sino-Tibetan	Cantonese	Dai et al. (2025)
Turkic	Kazakh	Veitsman and Hartmann (2025)
	Kyrgyz	Veitsman and Hartmann (2025)
	Turkish	Veitsman and Hartmann (2025)
	Turkmen	Veitsman and Hartmann (2025)
	Uzbek	Veitsman and Hartmann (2025); Bobojonova et al. (2025)
Multiple		Bagheri Nezhad et al. (2025); Zhu et al. (2025); Tashu and Tudor (2025); Sindhujan et al. (2025); Dewangan et al. (2025)

Table 1: Coverage of workshop papers across different languages. The final row ('Multiple') represents the scenario where more than five languages from multiple language families are experimented with.

lation and Translation Aids' was the second most popular topic with six papers. The other topics approximately had a similar number of papers. Apart from the papers mentioned in Table 2, Veitsman and Hartmann (2025) provided a survey on Central Asian Turkic languages spanning across several research areas.

3 Conclusions

The first Workshop on Language Models for Low-Resource Languages (LoResLM 2025) attracted a lot of interest from the NLP community, having 35 accepted papers from 52 submissions. The accepted papers mainly span eight language families, with the majority representation being from Indo-European families. Furthermore, the accepted papers contributed to 13 NLP research areas, with major contributions to *'Language Modelling'* and *'Machine Translation and Translation Aids'*. We believe the findings and resources from LoResLM will open exciting new avenues to empower linguistic diversity for millions of low-resource languages.

For the future iterations of LoResLM, we expect better representation from more diverse linguistic groups, particularly those from underrepresented families such as Uralic, Dravidian and Indigenous languages of the Americas. Furthermore, we aim to diversify research topics, encouraging work in areas such as speech processing, information extraction, and dialogue systems, which are critical for many practical applications.

Paper	Dialogue and Interactive Systems	Ethics, Bias, and Fairness	Information Retrieval and Text Mining	Language Modelling	Linguistic Insights Derived using Computational Techniques	Machine Translation and Translation Aids	NLP and LLM Applications	Offensive Speech Detection and Analysis	Phonology, Morphology and Word Segmentation	Question Answering	Lexical Semantics	Sentiment Analysis, Stylistic Analysis, Opinion and Argument Mining	Syntactic analysis (Tagging, Chunking, Parsing)
Liu et al. (2025)											✓		
Gamboa and Lee (2025)		✓											
Alam et al. (2025)												✓	
Cruz (2025)				✓									
Dai et al. (2025)						✓							
Turumtaev (2025)				✓									
Sani et al. (2025)											✓		
Mutsaddi and Choudhary (2025)							✓						
Amin et al. (2025)													✓
Bagheri Nezhad et al. (2025)				✓									
Ouattara et al. (2025)	✓												
Zhu et al. (2025)					✓								
Matzopoulos et al. (2025)				✓									
Rapacz and Smywiński-Pohl (2025)						✓							
Habibzadeh and Asadpour (2025)												✓	
Dmonte et al. (2025)						✓		✓					
Tashu and Tudor (2025)				✓									
Mokhtarabadi et al. (2025)				✓									
Tran et al. (2025)											✓		
Merad et al. (2025)						✓							
Mahlaza et al. (2025)					✓								
Nacar et al. (2025)				✓									
Kryvosheieva and Levy (2025)				✓									
Harju and van der Goot (2025)				✓									
Shang et al. (2025)				✓									
Donthi et al. (2025)						✓							
Sadhu et al. (2025)		✓											
Sindhujan et al. (2025)						✓							
Bobojonova et al. (2025)													✓
Dewangan et al. (2025)									✓				
Zeinalipour et al. (2025a)										✓			
Lasheras and Pinheiro (2025)										✓			
Zeinalipour et al. (2025b)							✓						
Zhukova et al. (2025)			✓										

Table 2: Coverage of workshop papers across different NLP areas.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2025. BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Muhammad Saad Amin, Luca Anselma, and Alessandro Mazzei. 2025. Exploiting Task Reversibility of DRS Parsing and Generation: Challenges and Insights from a Multi-lingual Perspective. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. 2025. Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language

Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Latofat Bobojonova, Arofat Akhundjanova, Phil Sidney Ostheimer, and Sophie Fellenz. 2025. BBPOS: BERT-based Part-of-Speech Tagging for Uzbek. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors. 2022. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jan Christian Blaise Cruz. 2025. Extracting General-use Transformers for Low-resource Languages via Knowledge Distillation. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, and Tsz Ho Pun. 2025. Next-Level Cantonese-to-Mandarin Translation: Fine-Tuning and Post-Processing with LLMs. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vikrant Dewangan, Bharath Raj S, Garvit Suri, and Raghav Sonavane. 2025. When Every Token Counts: Optimal Segmentation for Low-Resource Language Models. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala - A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385, Marseille, France. European Language Resources Association.

Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe, and Marcos Zampieri. 2025. Does Machine Translation Impact Offensive Language Identification? The Case of Indo-Aryan Languages. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM Abilities in Idiomatic Translation. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Lance Calvin Lim Gamboa and Mark Lee. 2025. Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Zahra Habibzadeh and Masoud Asadpour. 2025. Using Language Models for Assessment of Users' Satisfaction with Their Partner in Persian. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anika Harju and Rob van der Goot. 2025. How to age BERT Well: Continuous Training for Historical Language Adaptation. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2023. TTL: transformer-based two-phase transfer learning for cross-lingual news event detection. *International Journal of Machine Learning and Cybernetics*, 14(8):2739–2760.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Daria Kryvosheieva and Roger Levy. 2025. Controlled Evaluation of Syntactic Knowledge in Multilingual Language Models. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Uriel Anderson Lasheras and Vladia Pinheiro. 2025. CaLQuest.PT: Towards the Collection and Evaluation of Natural Causal Ladder Questions in Portuguese for AI Agents. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward More Meaningful Resources for Lower-resourced Languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yifan Liu, Gelila Tilahun, Xinxiang Gao, Qianfeng Wen, and Michael Gervers. 2025. A Comparative Study of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Zola Mahlaza, C. Maria Keet, Imaan Sayed, and Alexander Van Der Leek. 2025. IsiZulu Noun Classification Based on Replicating the Ensemble Approach for Runyankore. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Alexis Matzopoulos, Charl Hendriks, Hishaam Mahomed, and Francois Meyer. 2025. BabyLMs for isiXhosa: Data-Efficient Language Modelling in a Low-Resource Context. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Maite Melero, Sakriani Sakti, and Claudia Soria, editors. 2024. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Ibrahim Merad, Amos Wolf, Ziad Mazzawi, and Yannick Léo. 2025. Language verY Rare for All. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Hojjat Mokhtarabadi, Ziba Zamani, Abbas Maazallahi, and Mohammad Hossein Manshaei. 2025. Empowering Persian LLMs for Instruction Following: A Novel Dataset and Training Approach. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Atharva Mutsaddi and Aditya Prashant Choudhary. 2025. Enhancing Plagiarism Detection in Marathi with a Weighted Ensemble of TF-IDF and BERT Embeddings for Low-Resource Language Processing. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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. 2025. Towards Inclusive Arabic LLMs: A Culturally Aligned Benchmark in Arabic Large Language Model Evaluation. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, editors. 2023. *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*. Association for Computational Linguistics, Dubrovnik, Croatia.

Maimouna Ouattara, Abdoul Kader Kaboré, Jacques Klein, and Tegawendé F. Bissyandé. 2025. Bridging Literacy Gaps in African Informal Business Management with Low-Resource Conversational Agents. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*, pages 1–41.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.*, 55(11).

Maciej Rapacz and Aleksander Smywiński-Pohl. 2025. Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2025. Social Bias in Large Language Models for Bangla: An Empirical Study on Gender and Religious Bias. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad, and Devon Jarvis. 2025. Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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. 2025. Atlas-Chat: Adapting Large Language Models for Low-Resource Moroccan Arabic Dialect. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell system technical journal*, 30(1):50–64.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Tsegaye Misikir Tashu and Andreea Ioana Tudor. 2025. Mapping Cross-Lingual Sentence Representations for Low-Resource Language Pairs Using Pre-trained Language Models. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka, and Masao Utiyama. 2025. Exploiting Word Sense Disambiguation in Large Language Models for Machine Translation. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Galim Turumtaev. 2025. Stop Jostling: Adaptive Negative Sampling Reduces the Marginalization of Low-Resource Language Tokens by Cross-Entropy Loss. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing System and Speaker Metadata for 2,800+ Language Varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz

Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yana Veitsman and Mareike Hartmann. 2025. Recent Advancements and Challenges of Turkic Central Asian Language Processing. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini, and Marco Gori. 2025a. PersianMCQ-Instruct: A Comprehensive Resource for Generating Multiple-Choice Questions in Persian. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Kamyar Zeinalipour, Moahmmad Saad, Marco Maggini, and Marco Gori. 2025b. From Arabic Text to Puzzles: LLM-Driven Development of Arabic Educational Crosswords. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hongpu Zhu, Yuqi Liang, Wenjing Xu, and Hongzhi Xu. 2025. Evaluating Large Language Models for In-Context Learning of Linguistic Patterns In Unseen Low Resource Languages. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anastasia Zhukova, Christian E. Matt, and Bela Gipp. 2025. Automated Collection of Evaluation Dataset for Semantic Search in Low-Resource Domain Language. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages (LoResLM2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.