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

Hansi Hettiarachchi[1], Tharindu Ranasinghe[1], Alistair Plum[2], Paul Rayson[1], Ruslan Mitkov[1], Mohamed Gaber[3], Damith Premasiri[1], Fiona Anting Tan[4], Lasitha Uyangodage[5]

[1]Lancaster University, UK [2]University of Luxembourg, Luxembourg
[3]Queensland University of Technology, Australia [4]Instagram, Meta
[5]University of Münster, Germany
`loreslm.contact@gmail.com`

Abstract

The second workshop on Language Models for Low-Resource Languages (LoResLM 2026) was held in conjunction with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026) in Rabat, Morocco. 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 and dialects, following recent advancements in neural language models and their linguistic biases towards high-resource languages. LoResLM 2026 attracted a notable interest from the natural language processing (NLP) community, resulting in 55 accepted papers from 79 submissions. These contributions cover a broad range of low-resource languages from 13 language families and 11 diverse research areas, paving the way for future possibilities and promoting linguistic inclusivity in NLP.

1 Introduction

Language models (LMs) have constituted a fundamental research area, beginning with basic n-gram models in the 1950s (Shannon, 1951). These are computational frameworks that leverage the generative probability of word sequences to execute natural language processing (NLP) tasks (Zhao et al., 2023). Recent developments in LMs have predominantly transitioned toward neural language models owing to their superior capabilities (Zhao et al., 2023; Minaee et al., 2024). The creation of pretrained neural language models/transformers represents a pivotal achievement in LM research that substantially improved NLP performance (Vaswani et al., 2017; Devlin et al., 2019). This advancement has additionally catalysed the emergence of more sophisticated large language models (LLMs), such as GPT, which comprise enormous numbers of parameters pre-trained on massive text corpora, achieving state-of-the-art natural language understanding and generation across diverse applications (Touvron et al., 2023; Jiang et al., 2023).

Approximately 7,000 spoken languages exist globally (van Esch et al., 2022). Nevertheless, the majority of NLP research concentrates on roughly 20 high-resource languages (Magueresse et al., 2020). For instance, 63% of publications at ACL 2008 centred on English (Bender, 2011), and even ten years later, 70% of articles at ACL 2021 were evaluated exclusively in English (Ruder et al., 2022). The countless other languages that attract minimal research interest are typically termed low-resource languages. These languages usually lack adequate digital data and resources to facilitate NLP tasks. They are alternatively known as resource-scarce, resource-poor, less computerised, low-data, or low-density languages (Ranasinghe et al., 2025a).

Given that LM capabilities are fundamentally shaped by the properties of their pre-training language corpora, resource disparities are similarly reflected within the models. For example, numerous prevalent 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)) accommodate only English. Nonetheless, the cross-lingual capacities of transformers have enabled 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)), permitting low-resource languages to gain advantages from other languages via joint learning strategies. Notwithstanding this advancement, these models are generally constrained to approximately 100 languages due to the curse of multilingualism (Conneau et al., 2020). In response to this limitation, creating monolingual models (e.g., `Sinhala-BERT` for Sinhala (Ranasinghe et al., 2025b), `PhoBERT` for Vietnamese (Nguyen and Tuan Nguyen, 2020), and `BanglaBERT` for Bangla (Bhattacharjee et al., 2022)) represents another emerging trend recently established to advance research in low-resource languages.

Several common factors constrain low-resource NLP research. A primary challenge is insufficient data availability, as most model performance relies substantially on the volume of training data (Hettiarachchi et al., 2024). Even recent neural LMs with multilingual capabilities typically exhibit sub-optimal performance when pre-training data for a specific language is scarce or absent (Ahuja et al., 2022; Hettiarachchi et al., 2023). Furthermore, the lack of benchmark datasets designed for low-resource languages tends to skew most model evaluations toward high-resource languages (Blasi et al., 2022; Ranasinghe et al., 2024; Chen et al., 2026).

Notably, numerous ongoing initiatives aim to promote research on low-resource languages and reduce bias in NLP methodologies that favour high-resource languages (Chakravarthi et al., 2022; Ojha et al., 2023; Melero et al., 2024). We organised the inaugural Workshop on Language Models for Low-Resource Languages (LoResLM 2025) to further advance this momentum. LoResLM 2025 (Hettiarachchi et al., 2025a) concentrated specifically on LM-based methodologies for low-resource languages, soliciting submissions across a wide array of topics, including corpus creation, benchmark development, LM construction or adaptation, and exploring LM applications for low-resource languages (Hettiarachchi et al., 2025b). Building on the success of LoResLM 2025, we organised the second edition of this workshop (LoResLM 2026)[1] co-located with EACL 2026 to continue strengthening research in this vital area. Section 2 presents an overview of the workshop contributions, emphasising language and task/research area coverage. We encourage you to refer to the complete papers available in the proceedings for more comprehensive information.

2 Workshop Contributions

As the second iteration of the workshop, LoResLM 2026 received 79 submissions. Among these, we accepted 55 papers, including 44 long papers and 11 short papers, to appear in the workshop proceedings, following the review process. A detailed overview of the accepted papers, including their distribution across languages and research areas, is presented in the following sections.

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

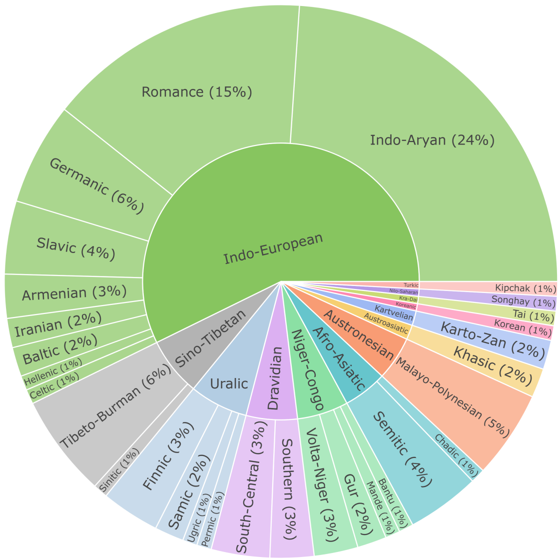

Figure 1: Distribution of LoResLM 2026 papers across language families.

2.1 Languages

As illustrated in Figure 1, the papers accepted to LoResLM 2026 cover languages from 13 distinct language families, reflecting a broad linguistic scope. Compared to the previous edition, submissions this year introduced six additional families: Austroasiatic, Dravidian, Kartvelian, Kra-Dai, Nilo-Saharan, and Uralic, demonstrating an expansion in both participation and research diversity.

Similar to the previous edition, the Indo-European family remains the most represented, accounting for approximately half of the overall language coverage. Contributions span nine branches within this family, with Indo-Aryan languages emerging as the most prominent subgroup, representing 24% of the total coverage within the workshop's focus area.

The remaining papers are distributed across a range of other language families. Sino-Tibetan, Uralic, Dravidian, Niger-Congo, Afro-Asiatic, and Austronesian each contribute roughly 5–7% of the coverage, forming a middle tier of representation. Additional families, including Austroasiatic, Kartvelian, Koreanic, Kra-Dai, Nilo-Saharan, and Turkic, appear with smaller shares of 1–2%, providing comparatively limited yet valuable diversity across linguistic contexts.

Table 1: Coverage of NLP areas and languages in workshop papers. Languages are colour-coded by language family, using the same colour scheme as in Figure 1.

Paper	Ethics, Bias, and Fairness	Information Retrieval and Text Mining	Language Modelling	Lexical Semantics	Linguistic Insights Derived using Computational Techniques	Machine Translation and Translation Aids	NLP and LLM Applications	Phonology, Morphology and Word Segmentation	Question Answering	Sentiment Analysis, Stylistic Analysis, Opinion and Argument Mining	Syntactic analysis (Tagging, Chunking, Parsing)	Languages
(Nyalang, 2026)			✓									Garo, Kokborok, Meitei, Mizo, Naga, Nyishi Assamese Khasi, Pnar
(Ravikiran et al., 2026)								✓				Hindi
(Yadav, 2026)							✓					Hindi
(H and Nardi, 2026)								✓				Hindi, Nepali, Urdu, Arabic
(Devadiga and Chopra, 2026)							✓					Tulu
(Margova and Penkov, 2026)							✓					Bulgarian
(Zehady et al., 2026)			✓									Bangla
(Tairu and Adebesin, 2026)									✓			Yoruba
(Keita et al., 2026)							✓					Zarma, Bambara
(Das, 2026)	✓											Kamtapuri
(Lasandi and Jayatilleke, 2026)		✓										Sinhala
(Hasan et al., 2026)								✓				Bangla
(Negoiță et al., 2026)			✓									Romanian
(Olusanya, 2026)								✓				Yoruba
(Creanga and Dinu, 2026)							✓					Romanian
(Kazi and Khoja, 2026)									✓			Urdu
(Panahi et al., 2026)							✓					Farsi
(McInerney et al., 2026)			✓									Irish
(Umar et al., 2026)						✓						Nupe
(Gundam and Mamidi, 2026)			✓									Telugu
(Utama et al., 2026)										✓		Balinese
(Mehak et al., 2026)			✓									Sindhi
(Gallagher and Heyer, 2026)											✓	Georgian
(Trivedi et al., 2026)							✓					Burmese, Armenian
(Mgonzo et al., 2026)								✓				Sukuma
(López-Otal and Gracia, 2026)			✓									Aragonese
(Yazdani et al., 2026)							✓					Persian
(Vidal-Gorène et al., 2026)											✓	Armenian, Greek Georgian Syriac
(Raihan et al., 2026)							✓					Bangla, Russian, Portuguese, Spanish Thai Arabic
(Perovic and Mihajlov, 2026)			✓									Serbian
(Navasardyan et al., 2026)				✓								Armenian
(Jakubauskaitė and Alhama, 2026)											✓	Lithuanian
(Poritski et al., 2026)							✓					Belarusian

Paper	Ethics, Bias, and Fairness	Information Retrieval and Text Mining	Language Modelling	Lexical Semantics	Linguistic Insights Derived using Computational Techniques	Machine Translation and Translation Aids	NLP and LLM Applications	Phonology, Morphology and Word Segmentation	Question Answering	Sentiment Analysis, Stylistic Analysis, Opinion and Argument Mining	Syntactic analysis (Tagging, Chunking, Parsing)	Languages
(Salim et al., 2026)						✓						Indonesian, Javanese, Sundanese
(Basoz et al., 2026)			✓									Afrikaans, Hindi, Marathi Telugu, Indonesian, Hausa, Korean
(Sachs, 2026)	✓				✓							Indonesian
(Tran et al., 2026)							✓			✓		Marathi
(Yadav et al., 2026)			✓									Maithili
(Chubakov, 2026)			✓									Kyrgyz
(Znotins, 2026)			✓									Latvian
(Liang and Chen, 2026)						✓						Cantonese
(Sälevä and Lignos, 2026)						✓						Finnish, North Sámi
(Xu and Kim, 2026)								✓				Estonian, Finnish Hungarian, Komi-Zyrian, North Sámi Russian
(Elboher and Pinter, 2026)								✓				Hebrew
(Verkijk and Vossen, 2026)				✓								Early Modern Dutch
(Gonçalves et al., 2026)					✓							Luxembourgish
(Miani et al., 2026)											✓	Old English, Old High German, Old Norse, Old Saxon
(Rosenbaum et al., 2026)						✓						Hebrew
(Staffini, 2026)								✓				Emilian-Romagnol, Friulian, Ladin, Ligurian, Lombard, Neapolitan, Piedmontese, Sardinian, Sicilian, Tarantino, Venetian
(Jamaluddin et al., 2026)						✓						Hindi, Odia, Urdu
(Rafid et al., 2026)							✓					Bangla
(Abubacar et al., 2026)						✓	✓					Gujarati, Hindi, Marathi, Sinhala Tamil, Telugu, Estonian
(Compaore et al., 2026)						✓	✓					Mooré French
(Ouedraogo et al., 2026)						✓	✓					Mooré French
(Gurav et al., 2026)						✓	✓					Gujarati, Hindi, Marathi Tamil, Telugu

A detailed summary of LoResLM 2026 papers, including language-level information, is presented in Table 1. In total, the accepted papers cover 82 distinct languages across the 13 families, highlighting strong community engagement with a wide range of languages and promoting opportunities for expanded multilingual and inclusive modelling efforts in future.

2.2 Research Areas

The papers accepted to LoResLM 2026 span 11 NLP research areas, in alignment with the call-for-papers themes commonly adopted by leading NLP conferences during 2024–2025. Table 1 presents the distribution of accepted papers across these areas, illustrating the breadth of topics addressed within the workshop.

The largest share of contributions falls under *NLP and LLM Applications*, with 16 accepted papers, making it the most prominent research area this year. *Language Modelling* emerges as the second most represented topic with 12 papers, followed by *Machine Translation and Translation Aids*, with 10 papers. All remaining research areas received fewer than 10 accepted papers, reflecting a more distributed yet diverse range of contributions.

A comparison with LoResLM 2025 (Figure 2) reveals notable shifts in research focus. *NLP and LLM Applications* demonstrates substantial growth, indicating increasing interest in applied and deployment-oriented work. *Phonology, Morphology and Word Segmentation* also shows considerable expansion compared to the previous edition. Positive growth trends are further observed in *Language Modelling*, *Machine Translation and Translation Aids*, and *Syntactic Analysis (Tagging, Chunking, Parsing)*. In contrast, areas such as *Ethics, Bias, and Fairness*, *Information Retrieval and Text Mining*, *Linguistic Insights Derived using Computational Techniques*, *Question Answering*, and *Sentiment Analysis, Stylistic Analysis, Opinion and Argument Mining* remain relatively stable across the two editions, suggesting a steady engagement, while *Lexical Semantics* exhibits a slight decline in representation this year.

3 Conclusions

The second workshop on Language Models for Low-Resource Languages (LoResLM 2026) reflects growing engagement from the NLP community, receiving 79 submissions and accepting 55 papers. The accepted papers demonstrate substantial linguistic diversity, covering 82 distinct languages across 13 language families. Consistent with broader trends in NLP research, Indo-European languages remain the most represented. In addition to linguistic breadth, the workshop spans 11 NLP research areas, with *NLP and LLM Applications* and *Language Modelling* emerging as the most prominent themes within the context of neural language models for low-resource languages. We believe the findings and resources from LoResLM 2026 will open exciting new avenues for advancing linguistic diversity and inclusivity across a wide range of low-resource languages.

The future iterations of LoResLM aim to further broaden linguistic and research diversity. We encourage stronger representation from underrepresented language families, such as Turkic, Kra-Dai, and Indigenous languages of the Americas. Expanding beyond text-based approaches remains another key direction, with increased interest in multimodal and speech-oriented research, as well as in studies of low-resource dialects and regional language varieties. In terms of research areas, we aim to diversify research topics, encouraging work in areas such as information extraction, lexical semantics, and dialogue systems, which are critical for many practical applications. Through these directions, we hope LoResLM will continue to expand the reach and inclusivity of language technologies for diverse linguistic communities worldwide.

References

Umar Abubacar, Roman Bauer, and Diptesh Kanojia. 2026. Parameter-Efficient Quality Estimation via Frozen Recursive Models. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

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.

Merve Basoz, Andrew Horne, and Mattia Opper. 2026. Bootstrapping Embeddings for Low Resource Languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

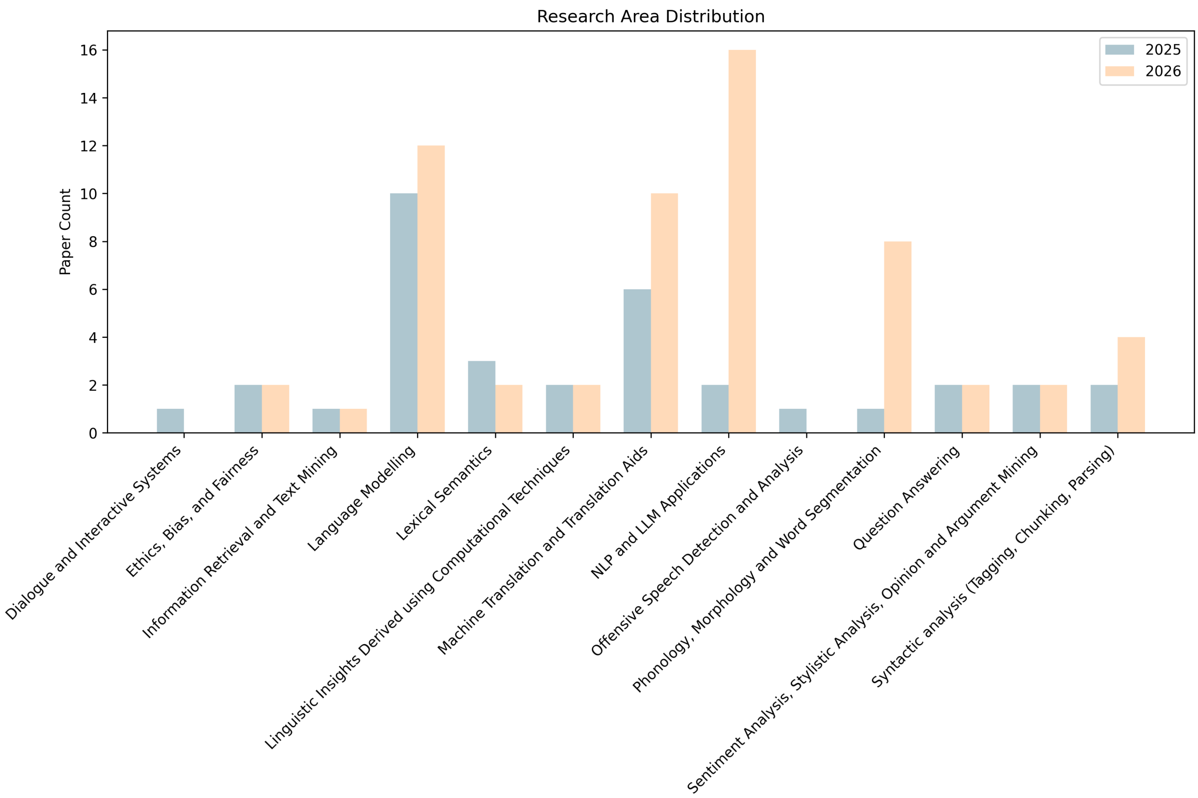

Figure 2: Distribution of LoResLM 2025, 2026 papers across research areas.

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

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

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.

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.

Claire Chen, Alistair Plum, Hansi Hettiarachchi, Diptesh Kanojia, Saroj Basnet, Marcos Zampieri, and Tharindu Ranasinghe. 2026. MUNIChus: Multilingual News Image Captioning Benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Tilek Chubakov. 2026. KyrText: A Multi-Domain Large-Scale Corpus for Kyrgyz Language. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Walker Stanislas Rocksane Compaore, Maimouna Ouattara, Rodrique Kafando, Tegawendé F. Bissyandé, Abdoul Kader Kabore, and Aminata Sabane. 2026. Neural Machine Translation for French–Mooré: Adapting Large Language Models to Low-Resource Languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

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.

Claudiu Creanga and Liviu P Dinu. 2026. LLM-as-a-Judge for Low-Resource Languages: Adapting Ragas and Comparative Ranking for Romanian. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Roumak Das. 2026. Quantifying Cross-Lingual Interference: Algorithmic Standardization of Kamtapuri in Large Language Models. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Prathamesh Devadiga and Paras Chopra. 2026. Making Large Language Models Speak Tulu: Structured Prompting for an Extremely Low-Resource Language. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Yair Elboher and Yuval Pinter. 2026. Hebrew Diacritics Restoration using Visual Representation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Daniel Gallagher and Gerhard Heyer. 2026. Targeted Syntactic Evaluation of Language Models on Georgian Case Alignment. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Christophe Friezas Gonçalves, Salima Lamsiyah, and Christoph Schommer. 2026. LuxDiagRC: A Diagnostic Reading Comprehension Corpus for Luxembourgish with Linguistic and Cognitive Annotation Layers. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Revanth Kumar Gundam and Radhika Mamidi. 2026. TeluguEval: A Comprehensive Benchmark for Evaluating LLM Capabilities in Telugu. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Namrata Bhalchandra Patil Gurav, Akashdeep Ranu, Archchana Sindhujan, and Diptesh Kanojia. 2026. Domain-Specific Quality Estimation for Machine Translation in Low-Resource Scenarios. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Manodnya K H and Luc De Nardi. 2026. When Multilingual Evaluation Assumptions Fail: Tokenization Effects Across Scripts. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Jakir Hasan, Shrestha Datta, Md Saiful Islam, Shubhashis Roy Dipta, and Ameya Debnath. 2026. BanglaIPA: Towards Robust Text-to-IPA Transcription with Contextual Rewriting in Bengali. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage. 2025a. Proceedings of the first workshop on language models for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Randunu Chandrakantha Uyangodage. 2025b. Overview of the first workshop on language models for low-resource languages (LoResLM 2025). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 1–8, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Urtė Jakubauskaitė and Raquel G. Alhama. 2026. Evaluating Large Language Models on Lithuanian Grammatical Cases. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Jamaluddin, Subhankar Panda, Aditya Narendra, Kamanksha Prasad Dubey, and Mohammad Nadeem. 2026. UrHiOdSynth: A Multilingual Synthetic Corpus for Speech-to-Speech Translation in Low-Resource Indic Languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Samreen Kazi and Shakeel Ahmed Khoja. 2026. QARI: Neural Architecture for Urdu Extractive Machine Reading Comprehension. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Mamadou K. Keita, Marcos Zampieri, Christopher M Homan, Adwoa Asantewaa Bremang, Dennis Asamoah Owusu, and Huy Le. 2026. Grammatical Error Correction for Low-Resource Languages: The Case of Zarma. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Minduli Lasandi and Nevidu Jayatilleke. 2026. SinhaLegal: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Yuzhi Liang and Fangqi Chen. 2026. Escaping the Probability Trap: Mitigating Semantic Drift in Cantonese-Mandarin Translation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. 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.

Miguel López-Otal and Jorge Gracia. 2026. "We Are (Language) Family": Adapting Transformer models to related minority languages with linguistic data. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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*.

Ruslana Margova and Stanislav Penkov. 2026. To make someone do something: mining alert-style directives in Bulgarian social media for low-resource language modelling. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Joseph McInerney, Khanh-Tung Tran, Liam Lonergan, Neasa Ní Chiaráin, Ailbhe Ni Chasaide, and Barry Devereux. 2026. Qomhrá: A Bilingual Irish and English Large Language Model. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Mehak, Kamyar Zeinalipour, Pireh Soomro, Cristiano Chesi, Marco Gori, and Marco Maggini. 2026. Enabling Structured Reasoning in Sindhi with Culturally Grounded Instruction Tuning. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Macton Mgonzo, Kezia Oketch, Naome A Etori, Winnie Mang'eni, Elizabeth Fabian Nyaki, and Michael Samwel Mollel. 2026. Learning from Scarcity: Building and Benchmarking Speech Technology for Sukuma. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Irene Miani, Sara Stymne, and Gregory R. Darwin. 2026. Cross-Lingual and Cross-Domain Transfer Learning for POS Tagging in Historical Germanic Low-Resource Languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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*.

Zaruhi Navasardyan, Bagratuni Minsayan, Spartak Bughdaryan, and Hrant Davtyan. 2026. Less is More: Adapting Text Embeddings for Low-Resource Languages with Small Scale Noisy Synthetic Data. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Vlad-Andrei Negoiță, Mihai Masala, and Traian Rebedea. 2026. Improving Romanian LLM Pretraining Data using Diversity and Quality Filtering. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Badal Nyalang. 2026. NE-BERT: A Multilingual Language Model for Nine Northeast Indian Languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. 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.

Joy Olusanya. 2026. Tone in Yoruba ASR: Evaluating the Impact of Tone Recognition on Transformer-Based ASR Models. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Soutongnoma Alex Fayçal Ouedraogo, Maimouna Ouattara, Rodrique Kafando, Abdoul Kader Kabore, Aminata Sabane, and Tegawendé F. Bissyandé. 2026. Contributing to Speech-to-Speech Translation for African Low-Resource Languages : Study of French-Mooré Pair. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Solmaz Panahi, John Kelleher, and Vasudevan Nedumpozhimana. 2026. When LLMs Annotate: Reliability Challenges in Low-Resource NLI. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Mitar Perovic and Teodora Mihajlov. 2026. Serbian SuperGLUE: Towards an Evaluation Benchmark for South Slavic Language Models. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Vladislav Poritski, Oksana Volchek, Maksim Aparovich, Volha Harytskaya, and Pavel Smrz. 2026. Tracking the evolution of LLM capabilities for Belarusian with OpenAI Evals. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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).

Ahmed Rafid, Rumman Adib, Fariya Ahmed, Ajwad Abrar, and Mohammed Saidul Islam. 2026. BanglaSummEval: Reference-Free Factual Consistency Evaluation for Bangla Summarization. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, and Marcos Zampieri. 2026. Large Language Models for Mental Health: A Multilingual Evaluation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

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.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025a. MUSTS: MUltilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353, Vienna, Austria. Association for Computational Linguistics.

Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, Isuri Nanomi Arachchige, Alistair Plum, Paul Rayson, and Ruslan Mitkov. 2025b. Sinhala encoder-only language models and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8623–8636, Vienna, Austria. Association for Computational Linguistics.

Manikandan Ravikiran, Tanmay Tiwari, Vibhu Gupta, Rakesh Prakash, Rohit Saluja, and Shayan Mohanty. 2026. Do Tokenizers Fail on Informal Hindi Expressions? Evidence from Static, Downstream, and Robustness Analyses. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Andy Rosenbaum, Assaf Siani, and Ilan Kernerman. 2026. MTQE.en-he: Machine Translation Quality Estimation for English-Hebrew. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Dan Sachs. 2026. The Indonesian Religiolect Corpus: Data Curation for Muslim, Protestant, and Catholic

Language Varieties. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Luis Frentzen Salim, Esteban Carlin, Alexandre Morinvil, Xi Ai, and Lun-Wei Ku. 2026. Beyond Many-Shot Translation: Scaling In-Context Demonstrations For Low-Resource Machine Translation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Alessio Staffini. 2026. Tokenization Cost, Retention, and Orthography Robustness for Ladin and Italian Varieties. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Jonne Sälevä and Constantine Lignos. 2026. How multilingual are multilingual LLMs? A case study in Northern Sámi-Finnish Translation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Zainab Tairu and Aramide Adebesin. 2026. Evaluating Retrieval-Augmented Generation for Medication Question Answering on Nigerian Drug Labels in Yorùbá. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van-Hien Tran, Huy Hien Vu, Hideki Tanaka, and Masao Utiyama. 2026. Representation-Aware Prompting for Zero-Shot Marathi Text Classification: IPA, Romanization, Repetition. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Kushal Trivedi, Murtuza Shaikh, and Sriyansh Sharma. 2026. "So, How Much Do LLMs Hallucinate on Low-Resource Languages?" A Quantitative and Qualitative Analysis. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Umar Baba Umar, Sulaimon Adebayo Bashir, and Mohammed Danlami Abdulmalik. 2026. 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. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Putu Kussa Laksana Utama, Tsegaye Misikir Tashu, and Jilles Steeve Dibangoye. 2026. Cross-Lingual Emotion Recognition in Balinese Text using Multilingual-LLMs under Peer-Collaborations Settings. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Stella Verkijk and Piek Vossen. 2026. Out-Of-Tune rather than Fine-Tuned: How Pre-training, Fine-tuning and Tokenization Affect Semantic Similarity in a Historical, Non-Standardized Domain. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Chahan Vidal-Gorène, Bastien Kindt, and Florian Cafiero. 2026. Under-resourced studies of under-resourced languages: lemmatization and POS-tagging with LLM annotators for historical Armenian, Georgian, Greek and Syriac. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Nuo Xu and Ahrii Kim. 2026. Tokenization and Morphological Fidelity in Uralic NLP: A Cross-Lingual Evaluation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for 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.

Sumit Yadav, Raju Kumar Yadav, Utsav Maskey, Gautam Siddharth Kashyap, Ganesh Gautam, and Usman Naseem. 2026. MaiBERT: A Pre-training Corpus and Language Model for Low-Resourced Maithili Language. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Tanushree Ravindra Pratap Yadav. 2026. Competence Collapse in Code-Mixed Generation: Spectral Evidence and Mechanistic Recovery via Cross-Lingual Activation Steering. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Shakib Yazdani, Cristina España-Bonet, Eleftherios Avramidis, Yasser Hamidullah, and Josef van Genabith. 2026. A Comprehensive Evaluation of Chain-of-Thought Faithfulness in Persian Classification Tasks. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Abdullah Khan Zehady, Shubhashis Roy Dipta, Naymul Islam, Safi Al Mamun, and Santu Karmaker. 2026. BanglaLlama: LLaMA for Bangla Language. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Arturs Znotins. 2026. Pretraining and Benchmarking Modern Encoders for Latvian. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, Rabat, Morocco. Association for Computational Linguistics.