Large Language Models and Natural Language Processing On Minority Languages: A Systematic Review

Rachel Edita Roxas

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Large Language Models and Natural Language Processing On Minority Languages: A Systematic Review

Rachel Edita Roxas
University of the Philippines Los Baños
Philippines
***roroxas2*@up.edu.ph**

Abstract

This study presents a systematic literature review on publications on minority languages in large language models and natural language processing. Using the Bibliometrics approach on Scopus-indexed documents published prior to November 2024, analyses and visualization were conducted. Aside from the surge on the number of publications in recent years, collaboration among countries/territories, and the predominance of the computer science subject area are noticeable. The keyword co-occurrence network revealed the prevalence of keywords related to the field of computer science. Schools of thought identified were: 1) Multilingualism and closely-related languages; 2) Performance Evaluation Approaches, and 3) Cross-lingual approaches. We identified the natural language considered in these studies, NLP tasks, technologies used, and social issues and concerns. Conclusions and recommendations for future work are presented.

1 Introduction

Artificial intelligence (AI) technologies have impacted our world. It has influenced various areas of our society such as in the field of finance and accounting (Biju, et al., 2024; Shakdwipee, et al., 2023; Cao, 2020), education (Alqahtani et al., 2023; Chen, et al., 2020; Tikhonova & Raitskaya, 2023), and health (Bajwa, et al., 2021; Li, 2024; Pagallo et al., 2023).

Generative AI uses large language models (LLMs) for natural language processing (NLP) applications. These LLMs have been trained on existing datasets which are predominantly in the majority languages. This underrepresentation of minority language has been shown to affect performance of these AI systems, and to have introduced a myriad of challenges including various sorts of biases (Hedderich et al., 2020).

Thus, the primary objective of this study is to investigate the landscape of the current body of knowledge on LLMs and NLP, with a focus on minority languages. Specifically, the research aims to assess the structure and dynamics of research work on LLM and NLP on minority languages using the bibliometric analysis approach.

2 Related Literature

Bibliometric analysis is "a scientific computer-assisted review methodology that can identify core research or authors, as well as their relationship, by covering all the publications related to a given topic or field" (as cited by Han, et al., 2020). Publication data in various fields such as health policy (Fusco et al., 2020), education (Meyer et al. 2023; Song & Wang, 2020), and nursing (Jabonete & Roxas, 2022), using different research databases, such as Scopus (Roxas & Recario, 2024; Song & Wang, 2020), Google Scholar, and Web of Science (Fan, et al, 2023; Martín-Martín, et al., 2018; Şahin & Candan, 2018).

Several works on bibliometric analysis in LLMs and NLP have been done in the recent past (Roxas & Recario, 2024; Tiwari et al., 2023), with a review paper focusing on low-resource languages (Krasadakis, et al., 2024), but with an emphasis on the legal domain. This shows that there is a gap on capturing the scientific landscape of LLMs and NLP on low resource languages, but across various domains.

3 Data Collection and Methods

We employed both quantitative and qualitative analyses in undertaking a systematic review of publications on LLMs and NLP on minority or low resource languages using the Scopus database.

3.1 Conceptual Framework

The conceptual framework of this study (Hallinger & Kovačević, 2022) constitutes the size, time, space, and composition of the scientific landscape: 1) size (or the quantity and quality) of publications on LLMs and NLP for minority languages, 2) the time and space for the extraction of documents were not defined, and 3) composition (or intellectual structure), which is captured using the visualizations as generated by Scopus, VOSviewer (Nees, et al., 2019) and Biblioshiny (Aria, et al., 2022).

3.2 Collection of Publication Data

We used the Preferred Reporting Items for Systematic Reviews and Meta-analysis (or PRISMA) (Page et al., 2021), which has identified stages: identification, screening, and included (as presented on Table 1).

At the identification phase, documents were retrieved through search Scopus functions of Scopus on October 2024. The search function TITLE-ABS-KEY((“large language model” OR LLM) AND (“natural language processing” OR NLP)) (or called LLM AND NLP hereon) yielded 4,683 documents. During screening, we refined the search function by including the keywords: low-resource, indigenous OR minority languages (simply called minority languages from hereon).

In the included stage, only 101 conference papers and 27 articles were included from the 135 documents, while we excluded 5 conference reviews, 1 editorial and 1 review paper. Further inclusion included 126 English documents, and the exclusion of 2 Chinese documents. These resulted in the 126 final included documents which will be used in this study for further analyses.

3.3 Visualizations

The intellectual structure (or composition) of the 126 Scopus-indexed publications on LLM and NLP on minority languages was captured by visualizations as generated by Biblioshiny (Aria, et al., 2022), Scopus, VOSviewer (Nees, et al., 2019) using the csv file of the 126 documents as exported from the Scopus database. The main information was generated by Biblioshiny (Aria, et al., 2022). Then we used the visualizations of the Scopus function Analyze-Results for the visualizations for documents by year, by country/territory, and by subject area. Then, we used VOSviewer (Nees, et al., 2019) to generate the keyword co-occurrence and co-citation research networks. Countries’ collaboration world map was also generated using Biblioshiny (Aria, et al., 2022).

Keyword Search	**# of Publications**
Identification: LLM AND NLP	4,683
Screening: LLM AND NLP AND Minority languages	135
Conference paper	101
Article	27
English	126
Included	**126**

Table 1: Scopus search functions.

4 Results and Discussion

4.1 LLMs and NLP: 4,683 documents

As shown in the main information (Figure 1), the 4,683 LLM and NLP Scopus documents were published from 1998 to 2025, showing an interesting near 25% international collaboration among the 13,778 authors, and a staggering 148,704 references.

Among the countries/territories (Figure 2), the US leads with 1,540 out of 4,683 documents (or 32.9%) followed by China with 919 (or 19.6%) and the United Kingdom with 346 (or 7.4%), Germany with 344 (or 7.3%), and India with 310 (or 6.6%).

In the subject area (Figure 3), computer science leads with 3720 documents (or at 79.4%) followed by other subject areas with not more than a quarter of the documents.

4.2 LLM and NLP: 126 documents on minority languages

As shown in the main information (Figure 4), the 126 LLM and NLP Scopus documents on minority languages were recently published from 2021 to 2025, showing a 31.75% international

collaboration (even greater than the LLM and NLP documents) among the 564 authors, and 5,064 references.

Figure 1. Main information: 4,683 LLM and NLP documents.

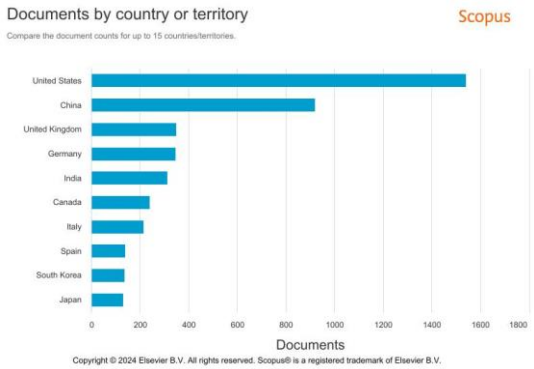

Figure 2. Documents by country/territory: 4,683 LLM and NLP.

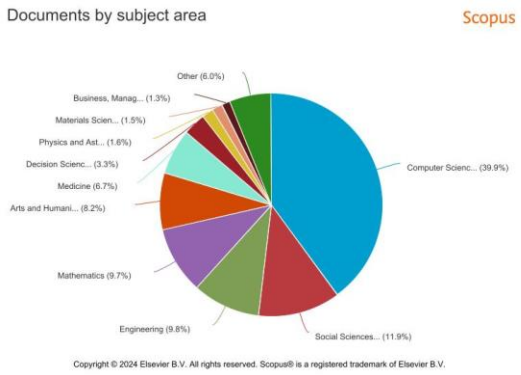

Figure 3. Documents by subject area: 4,683 LLM and NLP.

Figure 4. Main information: 126 LLM and NLP Minority Languages.

Among the countries/territories (Figure 5a), now China leads with 24 out of 126 documents (or 19.0%) followed closely by the US with 22 (or 17.5%), with the other countries with less than 10% contribution. Countries' collaboration and participation (shown in a world map in Figure 5b) reiterates the domination of the US and China.

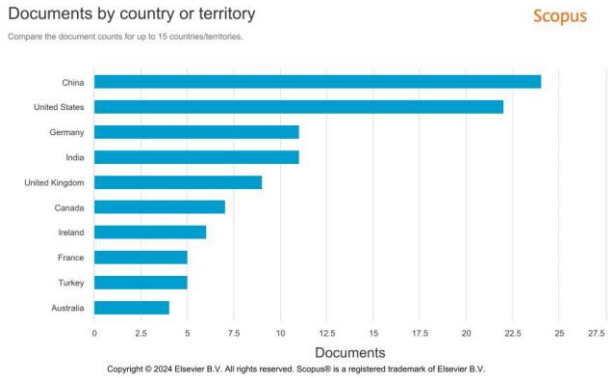

Figure 5a. Documents by country/territory: 126 LLM and NLP Minority Languages.

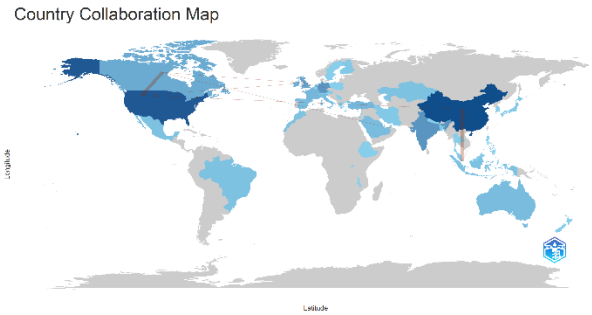

Figure 5b. Countries' Collaboration World Map: 126 LLM and NLP Minority Languages.

In subject area (Figure 6), computer science leads with 118 out of 126 documents (or 93.6%) showing the rigor and strength of the area of computer science in the usage of LLMs and NLP for minority languages, followed by other subject areas with 30% or less contribution. This implies that most publications on LLM and NLP focus more on the computational aspects of these new approaches.

A keyword co-occurrence network (Figure 7) was generated from the 126 documents on LLM and NLP focusing on minority languages, using all keywords, full counting, with a minimum number of occurrences of a keyword=5, of the 790 keywords, 60 meet the threshold, and produced 4 clusters showing the predominance of computer science related keywords.

A co-citation network (Figure 8) was generated from the 126 documents on on LLM and NLP focusing on minority languages, using cited references, with a minimum number of citations of a cited reference=5, of the 5,027 cited references, 27 meet the threshold, and only 26 are connected,

produced 3 clusters. The clusters in the co-citation networks are called by Hallinger and Nguyen (2020) as Schools of Thought, which we label as: 1) Multilingualism and closely-related languages; 2) Performance Evaluation Approaches, and 3) Cross-lingual approaches.

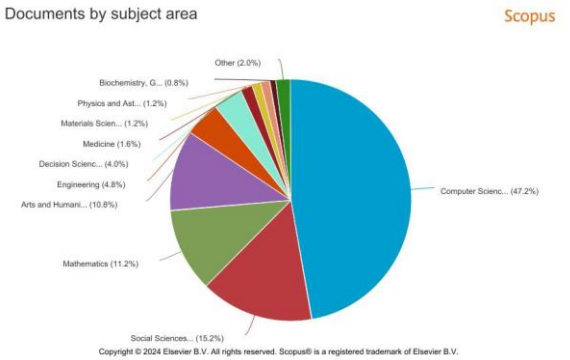

Figure 6. Documents by subject area: 126 LLM and NLP Minority Languages

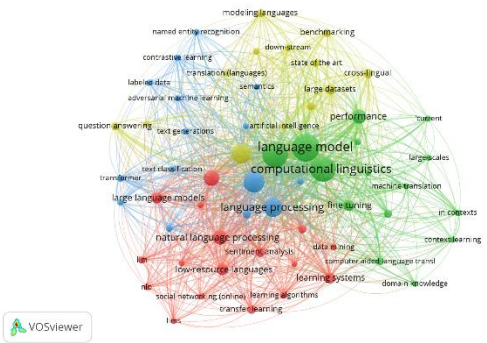

Figure 7. Keyword co-occurrence network: 126 LLM and NLP Minority languages.

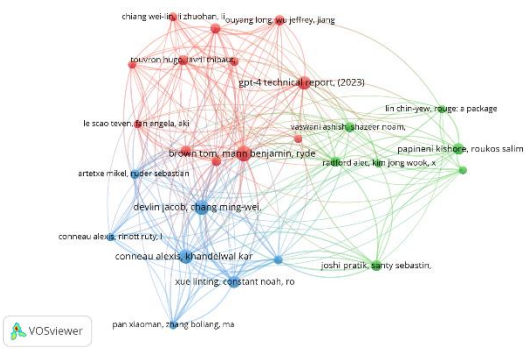

Figure 8. Co-citation: 126 LLM and NLP Minority languages.

4.3 Intellectual Structure: Minority Languages

The composition (or intellectual structure) of the 126 documents on LLM and NLP on minority languages is presented in this section, and is organized as follows: 1) natural languages considered in these studies, with a consideration of the dataset used; 2) NLP tasks focus; 3) technologies used; and 4) social issues and concerns.

Natural Languages: Publications considered specific minority languages, with some focusing on multiple languages in their experiments. Various datasets have been considered, such as Babel-670 with 670 languages representing 24 language families spoken in five continents (Vlantis et al., 2024), Glot500-m with 511 mostly low-resource languages (Imani et al., 2023), BELEBELE, a multiple-choice machine reading comprehension (MRC) dataset spanning 122 language variants (Bandarkar et al., 2024), and SIB-200 with 205 languages and dialects (Adelani et al., 2024). Other publications worked on closely-related language families such as 5 Ethiopian languages (Amharic, Ge'ez, Afan Oromo, Somali, and Tigrinya) (Tonja et al., 2024), Bengali, Gujarati, Hindi, Kannada, Maithili, Marathi, Tamil, Telugu, and Urdu (Dwivedi et al., 2024), Iberian languages (Cerezo-Costas et al., 2024), Comorian dialects (Naira et al., 2024), Chinese-centric languages (Zhang et a., 2024), Malawi and Chichewa (Lewis et al., 2024), and South and North Korea (Berthelier, 2023). This particular approach for multilinguality among closely-related languages is consistent with the findings of Pires et al. (2019) in that the models work best with similar languages.

More than one (1) publication focused on specific languages, such as Bangla (Sadhu et al., 2024; Hasan et al., 2024), Indonesia (Kim et al., 2023; 82), Pashto (Haq et al., 2023a; Haq et al., 2023b), Swahili (Muraoka et al., 2023; Liao & Wu, 2023), and Vietnamese (Pham et al., 2024; Nguyen et al., 2023), while one (1) document each on Armenian (Avetisyan & Broneske, 2023), Assamese script (Baruah et al., 2024), Brazil (Kim et al., 2023), Comorian dialects (Naira et al., 2024), Filipino (Cosme & de Leon, 2024), Jopara (Agüero-Torales et al., 2023), Kazakh (Shymbayev & Alimzhanov, 2023), Kannada (Aparna et al., 2024), Lao (Wang et al., 2024), Marathi (Gaikwad et al., 2024), Minangkabau (Nasution & Onan, 2024), Nigeria (Kim et al., 2023), Occitan (Vergez-Couret et al., 2024), Singlish (Tan et al., 2023), Sinhala (59), Urdu (Muraoka et al., 2023), and Uyrghur (Pan et al., 2024). Code switching languages were also considered in the studies such

as Jopara (combines Guarani and Spanish) (Agüero-Torales et al., 2023), and Filipino-English (Cosme & de Leon, 2024).

NLP Tasks: NLP tasks that were focused on by these 126 studies are led by question-answer generation or chatbots with 14 out of the 126 documents or 11.1%, sentiment analysis with 9 out of the 126 documents or 7.1%, and machine translation with 8 out of the 126 documents or 6.3%. Other NLP tasks include text classification, information retrieval, text summarization, syllabication, NLU, and NLG, automatic profiling of individuals, transliteration, sarcasm detection, offensive language detection, product matching, event argument extraction, entity extraction. The publications also focused on low-level NLP tasks such as tokenization, word segmentation, spelling correction, named entity recognition, and part of speech tagging, semantic parsing, and automatic speech recognition, and transcription. Other applications include machine-generated text detection system, LLM compression, prompt engineering, and synthetic data generation, and security concerns on the "jailbreak" problem (Deng et al., 2024), to address manipulation of LLMs towards undesirable behavior.

Due to the current status of minority languages that are still classified as low resource languages, some publications covered the construction and building of language resources such as: dataset collection and documentation of indigenous languages, and dataset labeling, lexicon construction, speech data collection, and offensive language dataset, and some for specific domains such as agriculture, covid, medicine, education, and for domain adaptation.

Technologies used: Technologies that were mentioned include the fine tuning of existing LLMs, with the leading LLM GPT, and BERT (or its variants). Other publications used BART, BARD, Bloomz, Electra, Flan-T5, Gemma, Llama2/3, LoRA, Mistral, PEGASUS, ProphetNet, and T5, mT5, Zephyr, and using particular technologies such as cross-lingual transfer learning, RNN, CNN, LSTM, BiLSTM, and NLTK.

Most of the 126 publications performed comparisons of evaluations and performance on particular datasets and chosen domains. Most advocated for open resources such as in (Batista et al., 2024).

Social issues and concerns: Social concerns include gender inclusivity NLP (Ovalle et al., 2024), gendered emotion attribution (Sadhu et al., 2024), balancing social impact, opportunities, and ethical constraints (Pinhanez et al., 2023), and regional bias of English LLMs (Lyu et al., 2024). The development of resource-limited devices or applications (Alyafeai & Ahmad, 2021) using lightweight LLMs (Urbizu et al., 2023) was also mentioned as LLMs require both computational speed and heavy storage.

5 Conclusions and Recommendations

We present a systematic literature review of Scopus publications published prior to November 2024 on LLM NLP focusing on minority or low-resource languages. Although that it has been shown that the US dominates the research work on LLM NLP, China leads on publications on LLM NLP on minority languages. Results also show that computer science subject area is still the focus of publications both on LLM NLP and LLM NLP on minority languages, where technology still dominates. Analyses show experiments on multilingual datasets, cross lingual approaches on closely-related languages, across various NLP tasks. Concerns have been raised in these publications on LLM NLP on minority languages on security concerns such as the "jailbreak" problem (Deng et al., 2024), and regional bias of English LLMs (Lyu et al., 2024), to name a few.

Since this study has focused on the publications from the Scopus research database, it is recommended to expand the publication dataset by considering other research databases.

References

D.I. Adelani, Liu H., Shen X., Vassilyev N., Alabi J.O., Mao Y., Gao H., and Lee E.-S.A. 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects

M.M. Agüero-Torales, López-Herrera A.G., and Vilares D. 2023. Multidimensional Affective Analysis for Low-Resource Languages: A Use Case with Guarani-Spanish Code-Switching Language. Cogn. Comput.

T. Alqahtani et al. 2023. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*,

vol. 19, no. 8, Jun. 2023, doi: https://doi.org/10.1016/j.sapharm.2023.05.016.

Sultan Alsarra, Alsarra, Parker Whitehead, Luay Abdeljaber, Naif Alatrash, Latifur Khan, Patrick T. Brandt, Javier Osorio, and Vito D'Orazio. 2024. Extractive Question Answering for Spanish and Arabic Political Text. *International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS)*. Pittsburgh, PA. Winner of the Best Paper Award, 2024, SBP-BRIMS.

Z. Alyafeai, and Ahmad I. 2021. Arabic Compact Language Modelling for Resource Limited Devices

M. Aparna, Srivatsa S., Sai Madhavan G., Dinesh T.B., and Srinivasa S. 2024. AI-Based Assistance for Management of Oral Community Knowledge in Low-Resource and Colloquial Kannada Language

M. Aria, C. Cuccurullo, L. D'Aniello, M. Misuraca, and M. Spano. 2022. Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy. *Sustainability*, vol. 14, no. 6, p. 3643, Mar. 2022, doi: https://doi.org/10.3390/su14063643.

H. Avetisyan, and Broneske, D. 2023. Large Language Models and Low-Resource Languages: An Examination of Armenian NLP. *International Joint Conference on Natural Language Processing.*

J. Bajwa, U. Munir, A. Nori, and B. Williams. 2021. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare Journal*, vol. 8, no. 2, pp. e188–e194, 2021, doi: https://doi.org/10.7861/fhj.2021-0095.

L. Bandarkar Liang D., Muller B., Artetxe M., Shukla S.N., Husa D., Goyal N., Krishnan A., Zettlemoyer L., and Khabsa M. 2024. The BELEBELE Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants

H. Baruah, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024. AssameseBackTranslit: Back Transliteration of Romanized Assamese Social Media Text. *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2014*). Torino, Italy, 20-25 May 2024.

V.A. Batista, Gomes D.S.M., and Evsukoff A. 2024. SESAME - self-supervised framework for extractive question answering over document collections

B. Berthelier. 2023. Division and the Digital Language Divide: A Critical Perspective on Natural Language Processing Resources for the South and North Korean Languages

A.K.V.N. Biju, Thomas, A.S. and Thasneem, J. 2024. Examining the research taxonomy of artificial intelligence, deep learning & machine learning in the financial sphere—a bibliometric analysis. *Qual Quant* 58, 849–878, 2024, doi: https://doi.org/10.1007/s11135-023-01673-0

L. Cao. 2020. AI in Finance: A Review. papers.ssrn.com, Jul. 10, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3647625

H. Cerezo-Costas, Alonso-Doval P., Hormazábal-Lagos M., and Creo A. 2024. Telescope: Discovering Multilingual LLM Generated Texts with Small Specialized Language Models

L. Chen, P. Chen, and Z. Lin. 2020. Artificial Intelligence in Education: a Review. *IEEE Access*, vol. 8, no. 2169–3536, pp. 75264–75278, Apr. 2020, doi: https://doi.org/10.1109/ACCESS.2020.2988510.

Camilla Johnine Cosme, and Marlene M. De Leon. 2024. Sentiment Analysis of Code-Switched Filipino-English Product and Service Reviews Using Transformers-Based Large Language Models. Archum Ateneo.

Y. Deng, Zhang W.; Pan S.J.; Bing L. 2024. Multilingual jailbreak challenges in large language models.

Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. 2024. Navigating Linguistic Diversity: *In-Context Learning and Prompt Engineering for Subjectivity Analysis in Low-Resource Languages.* SN Comput. Sci. 5, 4 (Apr 2024). https://doi.org/10.1007/s42979-024-02770-z

L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill. 2023. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *arXiv* (Cornell University), Apr. 2023, doi: https://doi.org/10.48550/arxiv.2304.02020.

F. Fusco, M. Marsilio, and C. Guglielmetti. 2020. Co-production in health policy and management: a comprehensive bibliometric review. *BMC Health Services Research*, vol. 20, no. 1, Jun. 2020, doi: https://doi.org/10.1186/s12913-020-05241-2.

H. Gaikwad, Kiwelekar A., Laddha M., and Shahare S. 2024. Adopting Pre-trained Large Language Models for Regional Language Tasks: A Case Study

P. Hallinger and V.-T. Nguyen. 2020. Mapping the Landscape and Structure of Research on Education for Sustainable Development: A Bibliometric Review. *Sustainability*, vol. 12, no. 5, p. 1947, Mar. 2020, doi: https://doi.org/10.3390/su12051947.

J. Han, H.-J. Kang, M. Kim, and G. H. Kwon. 2020. Mapping the intellectual structure of research on surgery with mixed reality: Bibliometric network analysis (2000–2019). *Journal of Biomedical Informatics*, vol. 109, p. 103516, Sep. 2020, doi: https://doi.org/10.1016/j.jbi.2020.103516.

P. Hallinger and J. Kovacevic. 2021. Mapping the intellectual lineage of educational management, administration and leadership, 1972–2020. *Educational Management Administration & Leadership*, p. 174114322110060, Apr. 2021, doi: https://doi.org/10.1177/17411432211006093.

I. Haq, Qiu W., Guo J., Tang P. 2023a. Pashto offensive language detection: a benchmark dataset and monolingual Pashto BERT

I. Haq, Qiu W., Guo J., Tang P. 2023b. NLPashto: NLP Toolkit for Low-resource Pashto Language

M.A. Hasan, Das S., Anjum A., Alam F., Anjum A., and Sarker A., and Noori S.R.H. 2024. Zero- and Shot Prompting with LLMs: A Comparative Study with Fine-tuned Models for Bangla Sentiment Analysis

M.A. Hedderich, Lange, L., Adel, H., Strötgen, J., and Klakow, D. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. arXiv 2020, arXiv:2010.12309.

H. Hettiarachchi, Premasiri D., Uyangodage L., and Ranasinghe T. 2024. NSina: A News Corpus for Sinhala

A. Imani, Lin P., Kargaran A.H., Severini S., Sabet M.J., Kassner N., Ma C., Schmid H., Martins A.F.T., Yvon F., and Schütze H. 2023. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages

F. G. V. Jabonete and R. E. O. Roxas. 2022. Barriers to Research Utilization in Nursing: A Systematic Review (2002–2021). *SAGE Open Nursing*, vol. 8, no. 8, p. 23779608221091073, May 2022, doi: https://doi.org/10.1177/23779608221091073.

Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika Wirtz, Traci Hong, and Derry Wijaya. 2023. COVID-19 Vaccine Misinformation in Middle Income Countries. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3903–3915, Singapore. Association for Computational Linguistics.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 83–94, Torino, Italia. ELRA and ICCL.

Panteleimon Krasadakis 1 , Evangelos Sakkopoulos 1,* and Vassilios S. Verykios. 2024. *A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages* , *Electronics* **2024**, *13*, 648. https://doi.org/10.3390/electronics13030648

D.-M. Lewis, Derenzi B., Misomali A., Nyirenda T.; Phiri E., Chifisi L., Makwenda C., and Lesh N. 2024. Human Review for Post-Training Improvement of Low-Resource Language Performance in Large Language Models

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023. Crosslingual Retrieval Augmented In-context Learning for Bangla. *In Proceedings of the First Workshop on Bangla Language Processing (BLP-2023),* pages 136–151, Singapore. Association for Computational Linguistics.

J. Li, A. Dada, J. Kleesiek, and J. Egger. 2024. ChatGPT in Healthcare: A Taxonomy and Systematic Review. Mar. 2024, doi: https://doi.org/10.1101/2023.03.30.23287899.

C. Liao, and X. Wu. 2023. On the Evaluation of ChatGPT-3.5 on Swahili Classification Tasks

J. Lyu, Dost K.; Koh Y.S.; Wicker J. 2024. Regional bias monolingual English language models

A. Martín-Martín, E. Orduna-Malea, and E. Delgado López-Cózar. 2018. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, vol. 116, no. 3, pp. 2175–2188, Jun. 2018, doi: https://doi.org/10.1007/s11192-018-2820-9.

J. G. Meyer et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, vol. 16, no. 1, Jul. 2023, doi: https://doi.org/10.1186/s13040-023-00339-9.

M. Muraoka, Bhattacharjee B., Merler M., Blackwood G., Li Y., and Zhao Y. 2023. Cross-Lingual Transfer of Large Language Model by Visually-Derived Supervision Toward Low-Resource Languages.

A.M. Naira, Bahafid A., Erraji Z., and Benelallam I. 2024. Datasets Creation and Empirical Evaluations of Cross-Lingual Learning on Extremely Low-Resource Languages: A Focus on Comorian Dialects

A.H., Nasution, and Onan A. 2024. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks

J. Nees, L. Van Eck, and Waltman. 2019. *VOSviewer Manual.* Available: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.11.pdf

M.-T., Nguyen, K.-T. Tran, Nguyen V., and Vu X.-S. 2023. ViGPTQA - State-of-the-Art LLMs for Vietnamese Question Answering: System Overview, Core Models Training, and Evaluations

Ovalle A.; Mehrabi N.; Goyal P.; Dhamala J.; Chang K.-W.; Zemel R.; Galstyan A.; Pinter Y.; Gupta R. 2024. Tokenization Matters: Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies

Ugo Pagallo et al. 2023. The underuse of AI in the health sector: Opportunity costs, success stories, risks and recommendations. *Health and Technology*, Dec. 2023, doi: https://doi.org/10.1007/s12553-023-00806-7.

M. J. Page et al.. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, vol. 10, no. 1, Mar. 2021, doi: https://doi.org/10.1186/s13643-021-01626-4.

Kun Pan, Xiaogang Zhang, and Liping Chen. 2024. Research on the Training and Application Methods of a Lightweight Agricultural Domain-Specific Large Language Model Supporting Mandarin

Chinese and Uyghur. *Applied Sciences* 14, no. 13: 5764. https://doi.org/10.3390/app14135764

Quoc-Hung Pham, Huu-Loi Le, Minh Dang Nhat, Khang Tran T., Manh Tran-Tien, Viet-Hung Dang, Huy-The Vu, Minh-Tien Nguyen, and Xuan-Hieu Phan. 2024. Towards Vietnamese Question and Answer Generation: An Empirical Study. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*. 23, 9, Article 132 (September 2024), 28 pages. https://doi.org/10.1145/3675781

C.S. Pinhanez, Cavalin P., Vasconcelos M., and Nogima J. 2023. Balancing Social Impact, Opportunities, and Ethical Constraints of Using AI in the Documentation and Vitalization of Indigenous Languages

T. Pires, Schlinger, E., and Garrette, D. 2019. How Multilingual is Multilingual BERT? In Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4996–5001.

REO, Roxas, and R. N. Recario. 2024. *Scientific landscape on opportunities and challenges of large language models and natural language processing*. Indonesian Journal of Electrical Engineering and Computer Science (IJEECS). [S.l.], v. 36, n. 1, p. 252-263, Oct. 2024. ISSN 2502-4760. https://ijeecs.iaescore.com/index.php/IJEECS/article/view/36861/18630 doi:http://doi.org/10.11591/ijeecs.v36.i1.pp252-263.

J. Sadhu, Saha M.R., and Shahriyar R. 2024. An Empirical Study of Gendered Stereotypes in Emotional Attributes for Bangla in Multilingual Large Language Models

K. Şahin and G. Candan. 2018. Scientific productivity and cooperation in Turkic world: a bibliometric analysis. *Scientometrics*, vol. 115, no. 3, pp. 1199–1229, Apr. 2018, doi: https://doi.org/10.1007/s11192-018-2730-x.

Pushpkant Shakdwipee, K. Agarwal, Hemlata Kunwar, and S. Singh. 2023. Artificial Intelligence in Finance and Accounting: Opportunities and Challenges. *Lecture notes in networks and systems*, pp. 165–177, Jan. 2023, doi: https://doi.org/10.1007/978-981-99-5652-4_17.

P. Song and X. Wang. 2020. A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review*, vol. 21, no. 3, pp. 473–486, Aug. 2020, doi: https://doi.org/10.1007/s12564-020-09640-2.

B. Subedi, Regmi, S., Bal, B.K., and Acharya, P. 2024. Exploring the Potential of Large Language Models (LLMs) for Low-resource Languages: A Study on Named-Entity Recognition (NER) and Part-Of-Speech (POS) Tagging for Nepali Language. *International Conference on Language Resources and Evaluation.*

D. Suhartono, Wongso W., and Tri Handoyo A. 2024. IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection

M. Shymbayev, Alimzhanov Y. 2023. Extractive Question Answering for Kazakh Language

S. H. Amanda Tan, E. S. Aung and H. YAMANA, "Two-stage fine-tuning for Low-resource English-based Creole with Pre-Trained LLMs," in 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, 2023, pp. 1-6, doi: 10.1109/CSDE59766.2023.10487143.

E. Tikhonova and L. Raitskaya. 2023. ChatGPT: Where Is a Silver Lining? Exploring the realm of GPT and large language models. *Journal of language and education*, vol. 9, no. 3, pp. 5–11, Sep. 2023, doi: https://doi.org/10.17323/jle.2023.18119.

A. Tiwari, S. Bardhan, and V. Kumar. 2023. A Bibliographic Study on Artificial Intelligence Research: Global Panorama and Indian Appearance. *arXiv* (Cornell University), Jul. 2023, doi: https://doi.org/10.48550/arxiv.2308.00705.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation. *In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* pages 6341–6352, Torino, Italia. ELRA and ICCL.

G. Urbizu G., Vicente I.S., Saralegi X., Agerri R., and Soroa A. 2023. Scaling Laws for BERT in Low-Resource Settings

Marianne Vergez-Couret, Myriam Bras, Aleksandra Miletić, and Clamença Poujade. 2024. Loflòc: A Morphological Lexicon for Occitan using Universal Dependencies. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 10716–10724, Torino, Italia. ELRA and ICCL.

D. Vlantis, Gornishka I., and Wang S. 2024. Benchmarking the Simplification of Dutch Municipal Text

W. Wang, Dong L., Yu Z., Huang Y., Guo J., and Gao S. 2024. Knowledge-Guided Reinforcement Learning for Low-Resource Unsupervised Syllabification

J. Zhang, Su K., Li H., Mao J., Tian Y., Wen F., Guo C., and T. Matsumoto. 2024. Neural Machine Translation for Low-Resource Languages from a Chinese-centric Perspective: A Survey