Data and Model Centric Approaches for Expansion of Large Language Models to New languages

Anoop Kunchukuttan, Raj Dabre, Rudra Murthy,
Mohammed Safi Ur Rahman Khan and Thanmay Jayakumar

Despite the increasing pace of Large Language Model (LLM) research, a vast majority of existing LLMs mainly support English alongside a handful of high resource languages, leaving a major gap for most low-resource languages. In this tutorial, we focus on approaches to expand the language coverage of LLMs. This provides an efficient and viable path to bring LLM technologies to low-resource languages, instead of training from scratch. We look at approaches at various stages of the LLM training pipeline, like tokenizer training, pre-training, instruction tuning, alignment, evaluation, etc., where adaptations are made to support new languages. We look at data-oriented approaches as well as model-oriented approaches. We hope that our tutorial enables researchers and practitioners to work on incorporating additional languages and tasks into existing LLMs to enhance inclusivity and coverage.

Anoop Kunchukuttan, Principal Applied Researcher, Microsoft
email: ankunchu@microsoft.com
website: https://anoopkunchukuttan.github.io
Anoop Kunchukuttan is a Principal Applied Researcher in the Core AI group at Microsoft and has worked on Azure Translator for a long time. He is also a co-founder and co-lead at AI4Bharat. His research interests include multilingual learning, machine translation, language modeling and low-resource NLP. He received his Ph.D from IIT Bombay. He has published in ACL, EMNLP, NAACL, TACL, TMLR, AAAI, IJCNLP and CSUR.

Raj Dabre, Research Scientist, Google Deepmind
email: prajdabre@google.com
website: https://prajdabre.github.io
Raj Dabre is a Research Scientist at Google DeepMind and an Adjunct Faculty at IIT Madras and IIT Bombay, India. He received his Ph.D. from Kyoto University and his Master's from IIT Bombay. His primary interests

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 12–13

are in low-resource NLP, language modeling and efficiency. He has published in ACL, EMNLP, NAACL, TMLR, AAAI, AACL, IJCNLP and CSUR. He is one of the senior leads at the AI4Bharat lab.

Rudra Murthy V, Research Scientist, IBM Research
email: rmurthyv@in.ibm.com
website: https://murthyrudra.github.io
Rudra Murhty is a research scientist at IBM Research since May 2020. He completed his PhD at IIT Bombay under the guidance of Prof. Pushpak Bhattacharyya. He has published in key AI/NLP conferences such as ACL, EMNLP, and NAACL. He has worked in machine translation, named entity recognition, and information retrieval with a focus on Indic languages.

Mohammed Safi Ur Rahman Khan, PhD Student, IIT Madras
email: mohammed.safi@dsai.iitm.ac.in
website: https://safikhansoofiyani.github.io
Mohammed Safi is a PhD student at the Wadhwani School of Data Science and AI, IIT Madras, under the supervision of Prof. Mitesh M. Khapra. He conducts research at the AI4Bharat Lab, with a focus on data-centric approaches and evaluation methodologies for multilingual LLMs. His work on developing large-scale multilingual pre-training and fine-tuning datasets earned the ACL 2024 Outstanding Paper award.

Thanmay Jayakumar, MS Student, IIT Madras
email: thanmayjayakumar@gmail.com
website: https://thanmayj.github.io
Thanmay Jayakumar is an MS student at the Wadhwani School of Data Science and AI, IIT Madras, under the supervision of Prof. Mitesh M. Khapra. His current interests lie in investigating the multilingual capabilities of LLMs and extending their support to low-resource languages, and his work has received the ACL 2024 Senior Area Chair award. He received his B.Tech. from VNIT Nagpur, India, where he worked on image captioning and open-ended information extraction.