Reading List for *Extending English Large Language Models to New Languages: A Survey*

Last updated: 6nd August 2024

Use these PAPER KEYS to identify papers cited in THIS PRESENTATION.

1. [AceGPT] Huang, Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen et al. "AceGPT, Localizing Large Language Models in Arabic." arXiv preprint arXiv:2309.12053 (2023).
2. [Airavata] Gala, Jay, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. "Airavata: Introducing Hindi Instruction-tuned LLM." *arXiv preprint arXiv:2401.15006* (2024).
3. [ALMA] Haoran Xu, Young Jin Kim, Amr Sharaf, Hany Hassan Awadalla. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. ICLR (2024).
4. [ALMA-R] Xu, Haoran, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. "Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation." *arXiv preprint arXiv:2401.08417* (2024).
5. [AveInit] John Hewitt. 2021. Initializing New Word Embeddings for Pretrained Language Models. Blog (https://nlp.stanford.edu/~johnhew/vocab-expansion.html)
6. [BactrianX] Li, Haonan, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. "Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation." *arXiv preprint arXiv:2305.15011*.
7. [BayLing] Zhang, Shaolei, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu et al. "Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models." *arXiv preprint arXiv:2306.10968* (2023).
8. [BigTrans] Yang, Wen, Chong Li, Jiajun Zhang, and Chengqing Zong. "Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages." *arXiv preprint arXiv:2305.18098* (2023).
9. [BLOOM+1] Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

10. [BLOOMZ] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.
11. [BUFFET] Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer. Arxiv pre-print 2305.14857.
12. [ChatGptMLing] Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
13. [ChatGptMT] Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
14. [ChineseLlama] Cui, Yiming, Ziqing Yang, and Xin Yao. "Efficient and effective text encoding for chinese llama and alpaca." *arXiv preprint arXiv:2304.08177* (2023).
15. [CIT] Li, Chong, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. "Align after Pre-train: Improving Multilingual Generative Models with Cross-lingual Alignment." *arXiv preprint arXiv:2311.08089* (2023).
16. [CLConsist] Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.
17. [COMMIT] Jaeseong Lee, YeonJoon Jung, and Seung-won Hwang. 2024. COMMIT: Code-Mixing English-Centric Large Language Model for Multilingual Instruction Tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3130–3137, Mexico City, Mexico. Association for Computational Linguistics.
18. [ConstrainedW2V] Mundra, Nandini, Aditya Nanda Kishore, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M. Khapra. "An Empirical Comparison of Vocabulary Expansion and Initialization Approaches for Language Models." *arXiv preprint arXiv:2407.05841* (2024).
19. [DocMTLLM] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language

Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

20. [ExpandChoices] Tejaswi, Atula, Nilesh Gupta, and Eunsol Choi. "Exploring Design Choices for Building Language-Specific LLMs." *arXiv preprint arXiv:2406.14670* (2024).
21. [Focus] Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.
22. [GPT3] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
23. [InciBiling] Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.
24. [IndicLLMSuite] Khan, Mohammed Safi Ur Rahman, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. "IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages." ACL 2024. *arXiv preprint arXiv:2403.06350* (2024).
25. [IndicMonoDoc] Meet Doshi, Raj Dabre, Pushpak Bhattacharyya. "Do Not Worry if You Do Not Have Data: Building Pretrained Language Models Using Translationese." *arXiv preprint arXiv: 2403.13638* (2024).
26. [InstructGPT] Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.
27. [LlmByndEng] Lai, Wen, Mohsen Mesgar, and Alexander Fraser. "LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback." *arXiv preprint arXiv:2406.01771* (2024).
28. [LmaByndEng] Zhao, Jun, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. "Llama beyond English: An empirical study on language capability transfer." *arXiv preprint arXiv:2401.01055* (2024).
29. [LmaLatent] Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. "Do Llamas Work in English? On the Latent Language of Multilingual Transformers." *arXiv preprint arXiv:2402.10588* (2024).
30. [LMPpl] Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying Prompts in Language Models via Perplexity Estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

31. [LSP] Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, Yutaka Matsuo. 2023. On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons. In *NAACL 2024*.
32. [MEGA] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
33. [mFTI] Li, Jiahuan, Hao Zhou, Shujian Huang, Shanbo Chen, and Jiajun Chen. "Eliciting the translation ability of large language models via multilingual finetuning with translation instructions." *arXiv preprint arXiv:2305.15083* (2023).
34. [mGPT] Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics.*
35. [MLAMA] Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.
36. [MonoMultiIFT] Chen, Pinzhen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. "Monolingual or multilingual instruction tuning: Which makes a better alpaca." *EACL (2024).*
37. [MSGptMT] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. Arxiv pre-print 2302.09210.
38. [MTDataPretrain] Ji, Shaoxiong, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. "Can Machine Translation Bridge Multilingual Pretraining and Cross-lingual Transfer Learning?." *arXiv preprint arXiv:2403.16777* (2024).
39. [Navarasa1] Navarasa Team. 2024. Introducing Indic Gemma 7B/2B Instruction tuned model on 9 Indian Languages - Navarasa. Blog (https://ravidesetty.medium.com/introducing-indic-gemma-7b-2b-instruction-tuned-model-on-9-indian-languages-navarasa-86bc81b4a282).
40. [Navarasa2] Navarasa Team. 2024. Introducing Navarasa 2.0 - Indic Gemma 7B/2B Instruction tuned model on 15 Indian Languages. Blog (https://ravidesetty.medium.com/introducing-navarasa-2-0-indic-gemma-7b-2b-instruction-tuned-model-on-15-indian-languages-31f6565b2750).
41. [OFA] Liu, Yihong, Peiqin Lin, Mingyang Wang, and Hinrich Schütze. 2023. "OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining." *arXiv preprint arXiv:2311.08849*.
42. [Okapi] Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In *Proceedings*

of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 318–327, Singapore. Association for Computational Linguistics.

43. [OpenHathi] Sarvam Team. 2023. OpenHathi Series: An Approach To Build Bilingual LLMs Frugally. Blog (https://www.sarvam.ai/blog/announcing-openhathi-series).
44. [Palm2] Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri et al. "Palm 2 technical report." *arXiv preprint arXiv:2305.10403* (2023).
45. [ParroT] Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during Chat using Large Language Models tuned with Human Translation and Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.
46. [Pinch] Shaham, Uri, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. "Multilingual instruction tuning with just a pinch of multilinguality." *arXiv preprint arXiv:2401.01854* (2024).
47. [PLUG] Zhang, Zhihan, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. "Plug: Leveraging pivot language in cross-lingual instruction tuning." *arXiv preprint arXiv:2311.08711* (2023).
48. [PNLD] Zhao, Yiran, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. "How do Large Language Models Handle Multilingualism?." *arXiv preprint arXiv:2402.18815* (2024).
49. [Polyglot] Kew, Tannon, Florian Schottmann, and Rico Sennrich. "Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed?." *arXiv preprint arXiv:2312.12683* (2023).
50. [PrimerPMLM] Doddapaneni, S., Ramesh, [PolyLM] Wei, Xiangpeng, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li et al. 2023. "Polylm: An open source polyglot large language model." *arXiv preprint arXiv:2307.06018* (2023).
51. G., Khapra, M.M., Kunchukuttan, A. and Kumar, P., 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.
52. [RomanSetu] Husain, Jaavid Aktar, Raj Dabre, Aswanth Kumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. "RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization." ACL 2024. (*arXiv preprint arXiv:2401.14280)*
53. [SharingNeurons] Weixuan Wang and Barry Haddow and Wei Peng and Alexandra Birch. 2024. Sharing Matters: Analysing Neurons Across Languages and Tasks in LLMs. *arXiv preprint arXiv: 406.09265* (2024).
54. [SDRRL] Zhang, Yuanchi, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. "Enhancing Multilingual Capabilities of Large Language Models through Self-Distillation from Resource-Rich Languages." *arXiv preprint arXiv:2402.12204* (2024).
55. [SeaLLM] Nguyen, Xuan-Phi, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen et al. "SeaLLMs--Large Language Models for Southeast Asia." *arXiv preprint arXiv:2312.00738* (2023).
56. [SparkAgi] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. Arxiv pre-print 2303.12712.
57. [TaCo] Upadhayay, Bibek, and Vahid Behzadan. "TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes." *arXiv preprint arXiv:2311.10797* (2023).

58. [TIM] Zeng, Jiali, Fandong Meng, Yongjing Yin, and Jie Zhou. "Tim: Teaching large language models to translate with comparison." *AAAI* (2024).
59. [Tower] Alves, Duarte M., José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters et al. "Tower: An Open Multilingual Large Language Model for Translation-Related Tasks." *arXiv preprint arXiv:2402.17733* (2024).
60. [Versatilists] Ye, Jiacheng, Xijia Tao, and Lingpeng Kong. "Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability." *arXiv preprint arXiv:2306.06688* (2023).
61. [WESCHEL] Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.
62. [XCOT] Chai, Linzheng, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang et al. "XCOT: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning." *arXiv preprint arXiv:2401.07037* (2024).
63. [XFactr] Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.
64. [XInstruction] Li, Chong, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. "X-Instruction: Aligning Language Model in Low-resource Languages with Self-curated Cross-lingual Instructions." *arXiv preprint arXiv:2405.19744* (2024).
65. [xLlama] Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, Lei Li. 2023. Extrapolating Large Language Models to Non-English by Aligning Languages. Arxiv eprint 2308.04948.
66. [XLT] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.
67. [ZSTT] Minixhofer, Benjamin, Edoardo Maria Ponti, and Ivan Vulić. "Zero-Shot Tokenizer Transfer." *arXiv preprint arXiv:2405.07883* (2024).