randomly selected documents for each language, extracted from the Wikipedia corpus (Foundation). Figure 4 shows the results of enhancement on high-resource languages (De, Fr, Zh, Es, Ru). The numbers represent the sizes of the training corpus when fine-tuning language-specific neurons, while "Random" represents the fine-tuning of an equivalent number of randomly chosen neurons using a corpus of 400. Our findings reveal that fine-tuning with a few hundred documents yields significant performance improvements on multilingual tasks: 3.4% on MGSM, 4.4% on XQuAD, 4.3% on X-CSQA, and 2.3% on XLSum. Moreover, English performance is enhanced by an average of 3.7% across all tasks. These results further confirm the effectiveness of `MWork` in interpreting structure functionality for LLM's multilingual query handling, offering precise and independent methods for multilingual enhancement. When fine-tuning with 800 documents, the performance deteriorates compared to using 400 documents. This drop can be attributed to the incorporation of additional knowledge, which disrupts the original knowledge distribution and leads to overfitting of the model to Wikipedia. This can be addressed by mixing data from more sources such as textbooks or websites.

In addition, we verify the effectiveness of such enhancement method on low-resource languages, given that low-resource performance is relatively low with the original model. We select four languages including Vietnamese (Vi), Thai (Th), Arabic (Ar), and Swahili (Sw), covering languages with both latin and non-latin scripts and having corresponding testing set in our considered benchmarks. The model was then evaluated on four benchmarks, and the result shown in Table 6 is the average scores among tasks. It is evident that the fine-tuning method using language-specific neurons enhances the model's multilingual performance in low-resource languages by an average of 2.2%. Notably, the improvement of 3.5% in English performance is observed even without an English training corpus, indicating the effectiveness of the distinct language responsibilities assigned to neurons.

5 Related Work

In the era of LLMs, numerous studies have been conducted to develop multilingual benchmarks (Zhang et al., 2023a), enhance multilingual performance without parameter adjustments through translation (Liang et al., 2023; Huang et al., 2023), aligning representations (Nguyen et al., 2023a; Salesky et al., 2023), prompting (Li et al., 2023b; Tanwar et al., 2023). Furthermore, certain works focus on improving multilingual abilities for a single task via cross-lingual transfer (Kim et al., 2017; Lin et al., 2019; Pfeiffer et al., 2020; Zhao et al., 2024b), while others aim to enhance multilingual proficiency by continuous training in one language to obtain mono-lingual LLMs (Cui et al., 2023), or in multiple domain languages to obtain domain-lingual LLMs (Nguyen et al., 2023b). Additionally, some works achieve multilingual LLMs by training from scratch (Muennighoff et al., 2023). However, these studies are limited to specific task types or require substantial training corpora due to a lack of comprehensive understanding of the multilingual mechanisms of LLMs.

Conventional interpretability research investigates the significance of input features with their corresponding outputs (Vig, 2019; Hewitt and Liang, 2019; Qiu et al., 2020). In the era of LLMs, one brunch of work includes efforts to understand knowledge storage, with (Geva et al., 2021) initiating the study of the feed-forward layer as a knowledge base. Subsequent work has furthered this by altering neuron values (Dai et al., 2022), mapping embeddings to words (Geva et al., 2022), modifying inputs to recover embeddings (Meng et al., 2022), and analyzing attention heads (Li et al., 2023a). Another line of research centers on the self-attention layer, examining its connection to reasoning capability (Hou et al., 2023; Stolfo et al., 2023; Friedman et al., 2023) by contrasting the reasoning tree based on attention weights.

6 Conclusion

In this work, we examine how LLMs handle multilingualism. The proposed multilingual workflow (`MWork`) suggests that LLMs initially understand queries by converting multilingual inputs into English, reason in English in intermediate layers while incorporating multilingual knowledge, and generate responses aligned with the original language in the final layers. The validity of `MWork` is verified using Parallel Language-specific Neuron Detection (`PLND`), which identifies activated neurons for different languages without labeled data. By detecting language-specific neurons and fine-tuning them with a small training corpus, `MWork` enhances multilingual abilities in specific languages without compromising others, resulting in significant improvements across tasks.

Acknowledgement

This work was substantially supported by DAMO Academy through DAMO Academy Research Intern Program. This research is partially supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-010-SGIL) and the Singapore Ministry of Education Academic Research Fund Tier 1 (Award No: T1 251RES2207).

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Wikimedia Foundation. Wikimedia downloads.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. 2023. Interpretability illusions in the generalization of simplified models.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919, Singapore. Association for Computational Linguistics.

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.

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

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Shuang Li, Xuming Hu, Aiwei Liu, Yawen Yang, Fukun Ma, Philip S Yu, and Lijie Wen. 2023b. Enhancing cross-lingual natural language inference by soft prompting with multilingual verbalizer. *arXiv preprint arXiv:2305.12761*.

Yaobo Liang, Quanzhi Zhu, Junhe Zhao, and Nan Duan. 2023. Machine-created universal language for cross-lingual transfer. *arXiv preprint arXiv:2305.13071*.

Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, Jie Zhou, et al. 2024. Multilingual knowledge editing with language-agnostic factual neurons. *arXiv preprint arXiv:2406.16416*.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 57.

Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024. Unraveling babel: Exploring multilingual activation patterns within large language models. *arXiv preprint arXiv:2402.16367*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Hoang H Nguyen, Chenwei Zhang, Tao Zhang, Eugene Rohrbaugh, and Philip S Yu. 2023a. Enhancing cross-lingual transfer via phonemic transcription integration. *arXiv preprint arXiv:2307.04361*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023b. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

OpenAI. 2023. Gpt-4 technical report.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.