On the Efficacy of Language Adapters for Cross-lingual Transfer in English-centric LLMs

Julian Schlenker
Saarland University | Linköping University
jusc00031@stud.uni-saarland.de

Abstract

Preliminary findings of an ongoing work examining the efficacy of language adapters for cross-lingual transfer in English-centric LLMs are presented. Using Llama 2 7B as base LLM, language adapters are trained for 13 languages. Their efficacy is assessed by training task adapters on two datasets in various source languages, with a zero-shot evaluation in the target languages. Current results demonstrate that language adapters exhibit inconsistent performance across languages and tasks, frequently harming performance. Some languages perform better with language adapters when a non-English source language is utilized suggesting that English may not be the optimal language for transfer.

1 Introduction

Most state-of-the-art LLMs are English-centric (Touvron et al., 2023; Jiang et al., 2023). To illustrate, in Llama 2 (Touvron et al., 2023), English constitutes 90% of the pre-training data. Despite this data imbalance, recent English-centric LLMs exhibit some multilingual capabilities (Kew et al., 2023; Ye et al., 2023). However, these capabilities are inconsistent across languages and tasks, with low-resource languages being particularly affected (Razumovskaia et al., 2024).

To endow LLMs with more profound multilingual capabilities, cross-lingual transfer (XLT) has emerged as a prevalent paradigm aiming to transfer task-specific knowledge from a high-resource source language to a lower-resource target language, thereby alleviating the constraint of having supervised task data (Philippy et al., 2023). One common setup for enhancing XLT abilities is to combine language and task adapters, parameter-efficient modules that are trained on top of a frozen base LLM and capture language- and task-specific representations, respectively (Pfeiffer et al., 2024). While this setup has been extensively evaluated for small-scale multilingual LLMs (Pfeiffer et al., 2020b; Parović et al., 2022; Rathore et al., 2023; Yong et al., 2023), there is little work that assesses its applicability to large-scale English-centric LLMs (Lin et al., 2024; Razumovskaia et al., 2024). Therefore, this work seeks to address the following RQs:

RQ1: Can adapter-based setups help enhance XLT abilities of English-centric LLMs?

RQ2: What patterns can be observed in terms of source language choice, typological relatedness, and downstream task?

2 Related Work

Language Adapters. Language adapters (LA) represent a parameter-efficient and modular method for language adaptation (Poth et al., 2023). They are added to a frozen base LLM and trained on monolingual, unsupervised data via language modeling in order to learn language-specific representations (Pfeiffer et al., 2020a). In general, any adapter architecture can be utilized for LA training: Prior work on small-scale, multilingual base LLMs has primarily employed *bottleneck adapters* (Houlsby et al., 2019) for LA training (Pfeiffer et al., 2020b; Parović et al., 2022; Faisal and Anastasopoulos, 2022; Yong et al., 2023). They observed enhanced XLT, particularly for lower-resource languages. However, Kunz and Holmström (2024) find that the effect of LAs varies considerably across target languages and omitting LAs is beneficial in some cases. More recent work that employs large-scale, English-centric base LLMs prefers *LoRA adapters* (Hu et al., 2021) for LA training (Lin et al., 2024; Razumovskaia et al., 2024). This may be due to the inference latency that *bottleneck adapters* introduce, which *LoRA adapters* help mitigate by merging their weights with the base LLM's weights (Hu et al., 2021).

Cross-lingual transfer in English-centric LLMs. Previous work evaluating XLT in English-centric LLMs can be roughly divided into four approaches: **LA + ICL** trains LAs for a base LLM followed by in-context learning (ICL)[1] at inference. Lin et al. (2024) report performance gains for languages with low-resource scripts, Razumovskaia et al. (2024) for NLG tasks only. **TA + ICL** directly trains single-task task adapters (TA) followed by ICL. Ye et al. (2023) show that minimal pre-training data for a given target language is conducive to XLT. **IT + ICL** uses multi-task instruction tuning (IT) to fine-tune a base LLM, followed by ICL. Previous work finds that multilingual IT with only a few languages (Aggarwal et al., 2024; Kew et al., 2023), or even monolingual IT in English (Chirkova and Nikoulina, 2024), suffices to elicit robust XLT abilities. **ICL** uses ICL only. Asai et al. (2024) and Ahuja et al. (2024) introduce XLT ICL benchmarks, revealing that English-centric LLMs perform well in high-resource languages but struggle with low-resource languages.

3 Experimental Setup

Unlike most previous work that assessed the XLT abilities of English-centric LLMs, this work begins by adapting the XLT setup as it is commonly employed for multilingual LLMs. The subsequent section details the current experimental setup. The experiments are still in progress.

3.1 Model

The open-source Llama 2 7B (Touvron et al., 2023) is selected as the base LLM. Despite the limited non-English pre-training data (2%), Llama 2 has demonstrated certain XLT abilities when fine-tuned for specific tasks (Ye et al., 2023) or evaluated using ICL (Asai et al., 2024; Ahuja et al., 2024). Refer to Appendix C for a breakdown of the language distribution in Llama 2's pre-training data.

3.2 Adapter Method

At present, this work utilizes *bottleneck adapters* as proposed by Pfeiffer et al. (2020b) to train LAs and TAs. This method injects trainable adapter layers into the frozen base LLM, comprising a down- and an up-projection, situated after the feed-forward block of each transformer layer. Crucially, this architecture allows composition; multiple bottleneck adapters can be easily stacked on top of each other.

[1]Following Li (2023), ICL encompasses any learning without parameter updates including zero-shot evaluation.

3.3 Data

Language Data. Following previous work (Pfeiffer et al., 2020b), this work trains LAs on monolingual, unsupervised data extracted from CC-100 (Conneau et al., 2020).

Task Data. Currently, one NLG task and one NLU task are evaluated: For NLG, MLQA-en (T) - an extractive QA dataset from the Aya Collection (Singh et al., 2024) - extends the English subset of MLQA (Lewis et al., 2020) with translations into 100 languages. For NLU, SIB-200 (Adelani et al., 2024) is selected, a topic classification dataset with seven labels. These datasets were chosen primarily for their extensive language coverage and availability of parallel data. Given the use of autoregressive LLMs, both tasks are framed as generative (see Appendix F for task templates). Exact Match and F1 are used as evaluation metrics for both tasks.

3.4 Languages

The current set includes 13 languages from three language groups: Seven Germanic languages, four Romance languages and two Finno-Ugric languages (see Appendix B for an overview on all languages). In each XLT setup, one language is designated as the source language, with the remaining ones as target languages. At present, English, German, and Spanish are selected as source languages. English serves as a reference, given its frequent use as a source language (e.g., Pfeiffer et al., 2020b; Parović et al., 2022). Based on the assumption that higher-resource languages generally transfer more effectively than lower-resource languages (Senel et al., 2024), German and Spanish are chosen as non-English source languages. Finno-Ugric languages are excluded as source languages due to their limited resources and typological distance from other languages.

3.5 Training & Evaluation Setups

The present work trains and evaluates two simple XLT setups to gain initial insights into the efficacy of LAs for XLT in English-centric LLMs (see Appendix A for training details and Appendix G for a detailed walk-through example):

(1) LA, adapted from Pfeiffer et al. (2020b), first trains language-specific LAs for all relevant languages, then trains a TA in the selected source language on top of the frozen source LA, and finally evaluates XLT zero-shot by replacing the source LA

Setup	af	gl	is	da	hu	fi	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.42	0.46	0.2	0.3	0.28	0.22	0.41	0.44	0.45	0.4	0.34	0.45	**0.78**	0.4
LA_{de}	**0.47**	0.51	**0.29**	0.45	0.4	**0.35**	0.51	0.5	**0.5**	0.45	0.45	0.52	0.45	**0.45**
LA_{es}	0.44	**0.52**	**0.29**	0.45	0.38	0.33	**0.53**	0.51	0.48	**0.53**	0.44	0.46	0.52	**0.45**
$noLA_{en}$	0.41	0.43	0.16	0.42	0.31	0.26	0.51	0.49	**0.5**	0.41	0.43	0.46	**0.78**	0.43
$noLA_{de}$	0.41	0.44	0.2	**0.49**	**0.41**	**0.35**	**0.53**	**0.52**	0.44	0.46	**0.46**	**0.53**	0.38	0.43
$noLA_{es}$	0.38	0.4	0.18	0.44	0.35	0.3	0.47	0.5	0.46	**0.53**	0.42	0.43	0.39	0.4

Table 1: MLQA-en F1 scores for LA and $noLA$ setup using different source languages. Underlined marks the best score within setting (LA or $noLA$), bold marks the best score between settings.

Setup	af	gl	is	da	hu	fi	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.51	0.74	0.31	0.65	0.48	0.48	0.62	0.77	0.77	0.8	0.7	0.79	**0.86**	0.65
LA_{de}	0.72	0.76	0.54	0.77	0.74	0.68	0.75	0.78	0.82	0.81	0.77	0.85	0.78	0.75
LA_{es}	0.7	0.79	**0.56**	0.79	0.69	0.64	0.76	0.83	0.82	0.82	0.81	0.82	0.74	0.75
$noLA_{en}$	0.66	0.76	0.35	0.72	0.63	0.55	0.79	0.83	0.77	0.83	0.74	0.8	0.85	0.71
$noLA_{de}$	**0.78**	**0.81**	0.52	**0.83**	**0.8**	**0.76**	0.84	0.85	**0.86**	0.82	**0.83**	**0.87**	0.85	**0.8**
$noLA_{es}$	0.75	**0.81**	0.45	0.79	0.76	0.68	**0.86**	**0.86**	0.85	**0.84**	0.81	0.82	0.83	0.78

Table 2: SIB-200 F1 scores for LA and $noLA$ setup using different source languages. Underlined marks the best score within setting (LA or $noLA$), bold marks the best score between settings.

with the target LA while retaining the source TA.

(2) $noLA$ omits LAs entirely. Only a TA is trained in the source language, then evaluated zero-shot in the target languages.

If LAs are beneficial, LA should outperform $noLA$. Besides their parameter-efficiency, LAs are motivated by their modularity. To retain modularity - particularly LA replacement at inference - the TA needs to be trained on top of the source LA. Omitting the source LA results in nonsensical outputs, as the model has not been exposed to an adapter stack during training. Alternatively, TAs, and thus the source LA, can be bypassed by using in-context learning, which does not involve task-specific fine-tuning. In-context learning is currently under evaluation.

4 Results & Analysis

Current findings are presented in Table 1 and 2. For each TA, the mean F1 scores over five random seeds are reported (see Appendix E for further results). In Table 1 and 2, the languages are ordered in ascending order according to the amount of pre-training data in Llama 2. The first vertical bar splits into unseen (left) and seen (right) languages.

4.1 Main Results

LAs do not consistently enhance XLT across target languages and tasks and often degrade performance. The average scores in Table 1 and 2 show that in only 2 out of 6 setups - LA_{de} and LA_{es} on MLQA-en - LA outperforms its $noLA$ counterpart. Even for the source languages themselves, LAs are unable to boost performance across tasks. These initial findings are in line with Kunz and Holmström (2024), who also observe inconsistencies across target languages and tasks for multilingual LLMs, as well as performance degradation with LAs in some cases. Moreover, the current findings align with previous work (Yong et al., 2023; Pfeiffer et al., 2020b) indicating that LAs are most beneficial for languages unseen during pre-training suggesting that LAs are able to capture target-language-specific representations without being susceptible to pre-training biases for these languages.

In this limited experimental setup, it cannot be concluded that LAs are a universal XLT booster in English-centric LLMs, as they entail increased computational cost while not consistently boosting performance across target languages and tasks. To test this tentative conclusion in greater detail, potential key variables are discussed below.

4.2 Impact of Task Type and Data

With the exception of Icelandic, the positive effect observed with LAs is limited to the NLG dataset MLQA-en. Kew et al. (2023) and Razumovskaia et al. (2024) also reported more pronounced XLT improvements for tasks requiring input/output language agreement (mostly NLG tasks). However, a more extensive evaluation on more tasks is needed to support this hypothesis since the MLQA-en targets often consist of a named entity that is uniform across several languages. Moreover, in contrast to

SIB-200, MLQA-en is machine translated, which may render it susceptible to translation errors. Considering that the LAs are only beneficial for LA_{de} and LA_{es} on MLQA-en, this may indicate that translated data contain similar noise, thereby facilitating generalization across non-English languages while hindering generalization from English to non-English target languages.

4.3 Impact of Source Language

Employing English as a high-resource source language, does not seem to be optimal. LA_{en} is only able to outperform $noLA_{en}$ for the unseen languages Afrikaans, Galician and Icelandic on MLQA-en. Most target languages exhibit a considerable deficit in performance relative to their $noLA_{en}$ counterparts. LA_{de} and LA_{es} are more effective across target languages on MLQA-en. Again, LAs are most beneficial for unseen languages. LA_{es} even performs on par with or better than $noLA_{es}$ across all target languages, yet is often outperformed by $noLA$ setups with other source languages. Notably, performance drops disproportionately for English as target language, suggesting that a non-English source language disrupts pre-trained English-centric representations. In the case of SIB-200, LA is not superior for any of the source languages tested. However, the performance of LA_{en} again exhibits a significant deficit relative to the other source languages, with gaps of up to 0.26 (Hungarian). Moreover, the impact of a non-English source language on English as target language is less pronounced than on MLQA-en. It is postulated that the use of English, the predominant language in Llama 2, as source language, engenders a further bias towards English and thus, impedes XLT. In addition, current observations indicate that, despite the limited pre-training data (German: 0.17%, Spanish: 0.13%), these languages can be leveraged for XLT. A factor that is believed to contribute to the task differences is the data formatting: Unlike SIB-200, which employs English instructions and labels for all languages, MLQA-en provides instructions and labels in the respective target language.

4.4 Impact of Typological Relatedness

Current results show that XLT is impeded for more distant target languages (here: non-Indo-European Hungarian and Finnish, as well as Icelandic without a close Germanic relative). These languages perform the worst across setup, source language and task. It is hypothesized that the observed deficiencies are due to a small vocabulary overlap, as indicated by the higher fertility in Table 6. Since the LAs employed in the present work do not operate on embedding level they are not expected to mitigate this issue.

Regarding potential benefits of typological relatedness for XLT, current results do not yield a discernible pattern. Comparing LA_{de} and LA_{es}, Table 1 reveals that Romance and Germanic target languages perform slightly, perhaps negligibly better when transferring from Spanish and German, respectively. For Catalan and Dutch, relatedness to their source language may be a crucial factor, as both languages show superior performance in the LA setup when transferring from the related source language and superior performance in the $noLA$ setup when transferring from the more distant source language. However, a comparison within a source language shows for LA_{es} that Romance languages do not consistently exhibit a greater benefit than other target languages. These observations suggest that XLT may benefit from some shallow typological relatedness.

5 Conclusion & Outlook

This work presents initial findings on the efficacy of LAs for XLT in English-centric LLMs. Regarding RQ1, the current results indicate that LAs' effect is largely inconsistent across target languages and tasks as $noLA$ often outperforms LA. As for the language set examined, LAs are most beneficial for target languages unseen during pre-training. Regarding RQ2, non-English source languages seem more suitable for XLT in English-centric LLMs than English. Furthermore, while an increased typological distance appears to adversely affect XLT, a higher typological relatedness does not consistently entail enhanced XLT.

However, given the limitations of the experimental setup, further investigation is required to substantiate the tentative conclusions. Accordingly, as this work progresses, the following variables will be assessed: Further base LLMs encompassing more multilingual pre-training data (Llama 3 and 3.1), further adapter methods (*LoRA* and *Prompt Tuning* are considered, as they differ in architecture and required parametric cost), a cleaner NLG dataset to assess the impact of potentially noisy task data, and multilingual LAs (and TAs) similar to Parović et al. (2022).

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Divyanshu Aggarwal, Ashutosh Sathe, Ishaan Watts, and Sunayana Sitaram. 2024. MAPLE: Multilingual evaluation of parameter efficient finetuning of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14824–14867, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. *Preprint*, arXiv:2402.14778.

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.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *ArXiv*, abs/2312.12683.

Jenny Kunz and Oskar Holmström. 2024. The impact of language adapters in cross-lingual transfer for NLU. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 24–43, St Julians, Malta. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2024. Modular deep learning. *Preprint*, arXiv:2302.11529.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. 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, Online. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: Zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987, Singapore. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet? *Preprint*, arXiv:2403.01929.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *ArXiv*, abs/2306.06688.

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.

A Training Details

Hyperparameter	Value
LAs	
Reduction factor	16
Batch size	4
Training steps	50k
Context (in tokens)	1024
MLQA-en TAs	
Reduction factor	16
Batch size	4
Training epochs	3
SIB-200 TAs	
Reduction factor	16
Batch size	4
Training epochs	20

Table 3: Details for training LAs and TAs. These values apply to all languages. I.e., LAs are trained on 200k samples per language à 1024 tokens. Unspecified hyperparameters were set to the default values as provided in the adapters and transformers library.

B Languages

Germanic	
English	en
German	de
Dutch	nl
Swedish	sv
Danish	da
Icelandic	is
Afrikaans	af
Romance	
Spanish	es
Portuguese	pt
Catalan	ca
Galician	gl
Finno-Ugric	
Finnish	fi
Hungarian	hu

Table 4: Languages used for LA training and evaluation.

C Llama 2 Language Distribution

Language	Data (in %)
en	90.00
de	0.17
sv	0.15
es	0.13
nl	0.12
pt	0.09
ca	0.04
fi	0.03
hu	0.03
da	0.02
is	0.00
gl	0.00
af	0.00

Table 5: Amounts of pre-training data in Llama 2 for languages relevant to this work.

D Fertility

Language	Fertility
en	1.45
de	2.04
sv	2.21
es	1.77
nl	2.00
pt	1.92
ca	1.96
fi	3.75
hu	3.00
da	2.22
is	3.03
gl	1.97
af	2.11

Table 6: Fertility (token/word ratio) as measured on the dev split of Flores-200 (Team et al., 2022) using the English-centric tokenizer of Llama 2.

E Further Results

E.1 F1 Scores with Standard Deviation

Setup	af	gl	is	da	fi	hu	ca
LA_{en}	0.42 (±0.01)	0.46 (±0.03)	0.2 (±0.04)	0.3 (±0.06)	0.22 (±0.01)	0.28 (±0.03)	0.41 (±0.05)
LA_{de}	**0.47** (±0.01)	0.51 (±0.01)	**0.29** (±0.02)	0.45 (±0.01)	**0.35** (±0.01)	0.4 (±0.01)	0.51 (±0.02)
LA_{es}	0.44 (±0.02)	**0.52** (±0.01)	**0.29** (±0.02)	0.45 (±0.02)	0.33 (±0.01)	0.38 (±0.02)	**0.53** (±0.01)
$noLA_{en}$	0.41 (±0.03)	0.43 (±0.05)	0.16 (±0.02)	0.42 (±0.04)	0.26 (±0.02)	0.31 (±0.03)	0.51 (±0.02)
$noLA_{de}$	0.41 (±0.01)	0.44 (±0.0)	0.2 (±0.01)	**0.49** (±0.01)	**0.35** (±0.0)	**0.41** (±0.01)	**0.53** (±0.01)
$noLA_{es}$	0.38 (±0.01)	0.4 (±0.01)	0.18 (±0.01)	0.44 (±0.01)	0.3 (±0.01)	0.35 (±0.01)	0.47 (±0.02)

Setup	pt	nl	es	sv	de	en	avg.
LA_{en}	0.44 (±0.03)	0.45 (±0.02)	0.4 (±0.03)	0.34 (±0.08)	0.45 (±0.01)	**0.78** (±0.0)	0.40
LA_{de}	0.5 (±0.01)	**0.5** (±0.02)	0.45 (±0.0)	0.45 (±0.01)	0.52 (±0.01)	0.45 (±0.11)	0.45
LA_{es}	0.51 (±0.01)	0.48 (±0.01)	**0.53** (±0.01)	0.44 (±0.01)	0.46 (±0.01)	0.52 (±0.04)	0.45
$noLA_{en}$	0.49 (±0.03)	**0.5** (±0.02)	0.41 (±0.04)	0.43 (±0.02)	0.46 (±0.02)	**0.78** (±0.0)	0.43
$noLA_{de}$	**0.52** (±0.01)	0.44 (±0.02)	0.46 (±0.0)	**0.46** (±0.01)	**0.53** (±0.0)	0.38 (±0.01)	0.43
$noLA_{es}$	0.5 (±0.01)	0.46 (±0.01)	**0.53** (±0.01)	0.42 (±0.01)	0.43 (±0.01)	0.39 (±0.08)	0.40

Table 7: MLQA-en F1 avg. scores over five random seeds. Standard deviation in parentheses. Underlined marks the best score within setting (LA or $noLA$), bold marks the best score between settings.

Setup	af	gl	is	da	fi	hu	ca
LA_{en}	0.51 (±0.15)	0.74 (±0.07)	0.31 (±0.09)	0.65 (±0.09)	0.48 (±0.1)	0.48 (±0.1)	0.62 (±0.13)
LA_{de}	0.72 (±0.04)	0.76 (±0.07)	0.54 (±0.09)	0.77 (±0.02)	0.68 (±0.06)	0.74 (±0.03)	0.75 (±0.07)
LA_{es}	0.7 (±0.05)	0.79 (±0.02)	**0.56** (±0.07)	0.79 (±0.07)	0.64 (±0.06)	0.69 (±0.13)	0.76 (±0.11)
$noLA_{en}$	0.66 (±0.04)	0.76 (±0.04)	0.35 (±0.05)	0.72 (±0.04)	0.55 (±0.1)	0.63 (±0.06)	0.79 (±0.06)
$noLA_{de}$	**0.78** (±0.03)	**0.81** (±0.04)	0.52 (±0.05)	**0.83** (±0.01)	**0.76** (±0.04)	**0.8** (±0.04)	0.84 (±0.02)
$noLA_{es}$	0.75 (±0.03)	**0.81** (±0.03)	0.45 (±0.04)	0.79 (±0.03)	0.68 (±0.07)	0.76 (±0.04)	**0.86** (±0.03)

Setup	pt	nl	es	sv	de	en	avg.
LA_{en}	0.77 (±0.04)	0.77 (±0.05)	0.8 (±0.02)	0.7 (±0.05)	0.79 (±0.05)	**0.86** (±0.02)	0.65
LA_{de}	0.78 (±0.07)	0.82 (±0.02)	0.81 (±0.02)	0.77 (±0.07)	0.85 (±0.02)	0.78 (±0.14)	0.75
LA_{es}	0.83 (±0.03)	0.82 (±0.02)	0.82 (±0.03)	0.81 (±0.03)	0.82 (±0.03)	0.74 (±0.14)	0.75
$noLA_{en}$	0.83 (±0.03)	0.77 (±0.02)	0.83 (±0.04)	0.74 (±0.05)	0.8 (±0.03)	0.85 (±0.03)	0.71
$noLA_{de}$	0.85 (±0.03)	**0.86** (±0.02)	0.82 (±0.01)	**0.83** (±0.01)	**0.87** (±0.03)	0.85 (±0.02)	0.80
$noLA_{es}$	**0.86** (±0.03)	0.85 (±0.03)	**0.84** (±0.01)	0.81 (±0.03)	0.82 (±0.02)	0.83 (±0.04)	0.78

Table 8: SIB-200 F1 avg. scores over five random seeds. Standard deviation in parentheses. Underlined marks the best score within setting (LA or $noLA$), bold marks the best score between settings.

E.2 Exact Match Scores with Standard Deviation

Setup	af	gl	is	da	fi	hu	ca
LA_{en}	0.21 (±0.02)	0.26 (±0.03)	0.07 (±0.02)	0.13 (±0.04)	0.08 (±0.01)	0.14 (±0.03)	0.2 (±0.04)
LA_{de}	**0.24** (±0.02)	**0.28** (±0.02)	0.13 (±0.02)	0.25 (±0.02)	**0.17** (±0.01)	0.25 (±0.02)	0.27 (±0.02)
LA_{es}	0.19 (±0.03)	0.25 (±0.01)	**0.14** (±0.02)	0.23 (±0.02)	0.16 (±0.01)	0.23 (±0.01)	0.25 (±0.02)
$noLA_{en}$	0.23 (±0.02)	0.25 (±0.03)	0.07 (±0.02)	0.24 (±0.04)	0.11 (±0.02)	0.18 (±0.02)	**0.3** (±0.01)
$noLA_{de}$	0.22 (±0.01)	0.24 (±0.01)	0.09 (±0.01)	**0.3** (±0.01)	**0.17** (±0.0)	**0.26** (±0.01)	**0.3** (±0.01)
$noLA_{es}$	0.16 (±0.02)	0.14 (±0.01)	0.06 (±0.01)	0.22 (±0.02)	0.13 (±0.01)	0.19 (±0.02)	0.19 (±0.04)

Setup	pt	nl	es	sv	de	en	avg.
LA_{en}	0.22 (±0.03)	0.29 (±0.07)	0.16 (±0.05)	0.15 (±0.05)	0.28 (±0.06)	**0.67** (±0.06)	0.22
LA_{de}	0.26 (±0.01)	0.29 (±0.01)	0.16 (±0.0)	0.23 (±0.01)	**0.33** (±0.01)	0.25 (±0.11)	0.24
LA_{es}	0.24 (±0.01)	0.23 (±0.02)	**0.26** (±0.01)	0.22 (±0.02)	0.23 (±0.02)	0.24 (±0.06)	0.22
$noLA_{en}$	0.27 (±0.02)	**0.32** (±0.02)	0.14 (±0.02)	0.24 (±0.02)	0.27 (±0.02)	0.65 (±0.01)	0.25
$noLA_{de}$	**0.28** (±0.01)	0.26 (±0.01)	0.17 (±0.01)	**0.26** (±0.01)	**0.33** (±0.01)	0.25 (±0.02)	0.24
$noLA_{es}$	0.24 (±0.02)	0.23 (±0.02)	0.25 (±0.01)	0.2 (±0.02)	0.2 (±0.03)	0.15 (±0.05)	0.18

Table 9: MLQA-en Exact Match scores over five random seeds. Standard deviation in parentheses. Underlined marks the best score within setting (LA or $noLA$), bold marks the best score between settings.

Setup	af	gl	is	da	fi	hu	ca
LA_{en}	0.49 (±0.17)	0.73 (±0.08)	0.3 (±0.1)	0.65 (±0.09)	0.47 (±0.11)	0.48 (±0.1)	0.61 (±0.15)
LA_{de}	0.72 (±0.04)	0.76 (±0.07)	0.52 (±0.11)	0.77 (±0.02)	0.67 (±0.07)	0.73 (±0.03)	0.75 (±0.07)
LA_{es}	0.69 (±0.05)	0.79 (±0.03)	**0.55** (±0.08)	0.79 (±0.07)	0.64 (±0.06)	0.69 (±0.14)	0.75 (±0.11)
$noLA_{en}$	0.66 (±0.04)	0.76 (±0.04)	0.35 (±0.05)	0.72 (±0.04)	0.55 (±0.1)	0.63 (±0.06)	0.79 (±0.06)
$noLA_{de}$	**0.78** (±0.03)	**0.81** (±0.04)	0.52 (±0.05)	**0.83** (±0.01)	**0.76** (±0.04)	**0.8** (±0.04)	0.84 (±0.02)
$noLA_{es}$	0.75 (±0.03)	**0.81** (±0.03)	0.45 (±0.04)	0.79 (±0.03)	0.68 (±0.07)	0.76 (±0.04)	**0.86** (±0.03)

Setup	pt	nl	es	sv	de	en	avg.
LA_{en}	0.77 (±0.04)	0.77 (±0.05)	0.79 (±0.02)	0.69 (±0.05)	0.79 (±0.05)	**0.86** (±0.02)	0.65
LA_{de}	0.78 (±0.07)	0.82 (±0.03)	0.81 (±0.02)	0.76 (±0.08)	0.85 (±0.02)	0.77 (±0.15)	0.75
LA_{es}	0.83 (±0.03)	0.82 (±0.02)	0.82 (±0.03)	0.81 (±0.02)	0.82 (±0.03)	0.73 (±0.15)	0.75
$noLA_{en}$	0.83 (±0.03)	0.77 (±0.02)	0.83 (±0.04)	0.74 (±0.05)	0.8 (±0.03)	0.85 (±0.03)	0.71
$noLA_{de}$	0.85 (±0.03)	**0.86** (±0.02)	0.82 (±0.01)	**0.83** (±0.01)	**0.87** (±0.03)	0.85 (±0.02)	0.80
$noLA_{es}$	**0.86** (±0.03)	0.85 (±0.03)	**0.84** (±0.01)	0.81 (±0.03)	0.82 (±0.02)	0.83 (±0.04)	0.78

Table 10: SIB-200 Exact Match scores over five random seeds. Standard deviation in parentheses. Underlined marks the best score within setting (LA or $noLA$), bold marks the best score between settings.

F Task Templates

MLQA-en

```
### Human: Refer to the passage below and then answer the question afterwards
in the same language as the passage:

Passage: {passage}

Question: {question}

### Assistant: {answer}
```

Figure 1: Task template used for training MLQA-en TAs. Instructions and labels are provided in the respective language.

SIB-200

```
Classify the following sentence into one of the following topics:
1. science/technology
2. travel
3. politics
4. sports
5. health
6. entertainment
7. geography

Sentence: {sentence}

Topic: {topic}
```

Figure 2: Task template used for training SIB-200 TAs. Instructions and labels are provided in English only.

G Training & Evaluation Setups

G.1 *LA* Setup

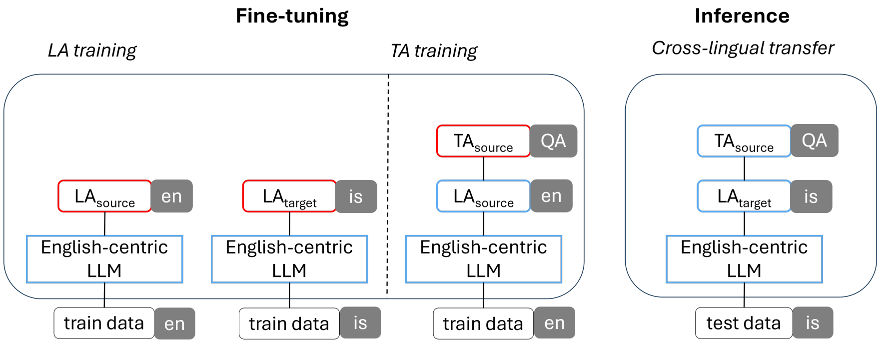

Figure 3: LA setup (blue and red edges indicate frozen and trainable parameters, respectively): (1) Language adapters are trained for each language of interest (here: English and Icelandic) on a frozen English-centric LLM (e.g., Llama 2 7B, as used in this work). (2) A task adapter (in this case, for a QA task) is trained in the source language (here: English) by stacking it on top of the frozen language adapter in the respective source language. (3) During inference, the source language adapter is replaced by the target language adapter (here: Icelandic) while retaining the task adapter in the source language. This setup is then evaluated zero-shot in the target language.

G.2 *noLA* Setup

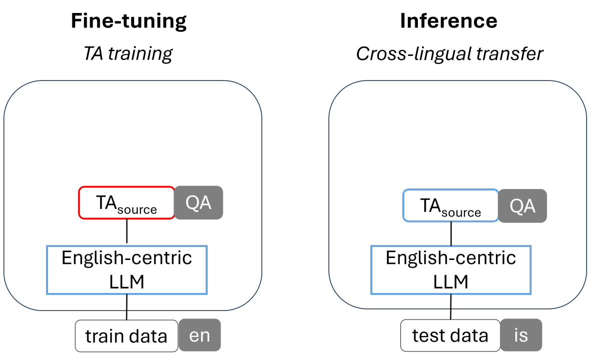

Figure 4: $noLA$ setup (blue and red edges indicate frozen and trainable parameters, respectively): (1) A task adapter (in this case, for a QA task) is trained in the source language (here: English) on top of the frozen English-centric LLM. (2) During inference, the task adapter in the source language is retained and evaluated zero-shot in the target language (here: Icelandic).