Only for the Unseen Languages, Say the Llamas: On the Efficacy of Language Adapters for Cross-lingual Transfer in English-centric LLMs

Julian Schlenker[1], Jenny Kunz[2], Tatiana Anikina[3], Günter Neumann[3], Simon Ostermann[3]

[1]Data and Web Science Group, University of Mannheim, Germany
[2]Dept. of Computer and Information Science, Linköping University, Sweden
[3]German Research Center for Artificial Intelligence, Saarland Informatics Campus, Germany
julian.schlenker@uni-mannheim.de

Abstract

Most state-of-the-art large language models (LLMs) are trained mainly on English data, limiting their effectiveness on non-English, especially low-resource, languages. This study investigates whether language adapters can facilitate cross-lingual transfer in English-centric LLMs. We train language adapters for 13 languages using Llama 2 (7B) and Llama 3.1 (8B) as base models, and evaluate their effectiveness on two downstream tasks (MLQA and SIB-200) using either task adapters or in-context learning. Our results reveal that language adapters improve performance for languages not seen during pre-training, but provide negligible benefit for seen languages. These findings highlight the limitations of language adapters as a general solution for multilingual adaptation in English-centric LLMs.

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., 2024; 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).

As LLMs grow larger and full fine-tuning becomes less feasible, parameter-efficient fine-tuning (PEFT) methods have been explored for XLT (Houlsby et al., 2019; Hu et al., 2021). One common setup for enhancing XLT abilities is to combine small language and task adaptation modules, as introduced by Pfeiffer et al. (2020b). The authors propose language adapters (LAs) and task adapters (TAs), parameter-efficient modules that are trained on top of a frozen base LLM and capture language- and task-specific representations, respectively.

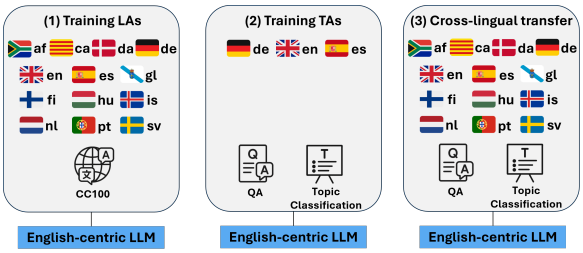

Figure 1: To evaluate cross-lingual transfer, language adapters (for 13 languages) and task adapters (for 3 high-resource source languages) are trained on top of a frozen English-centric LLM. Task adapters are evaluated on all languages of interest on two selected tasks.

While LAs have 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 only a paucity of work that assesses its applicability to large-scale English-centric LLMs (Lin et al., 2024; Razumovskaia et al., 2024). Our work closes this gap by making the following contributions:

1. We evaluate in a systematic manner whether LAs help enhance XLT abilities of English-centric LLMs across 13 linguistically diverse languages and two tasks (one QA and one NLU task) to inspect the impact of typological relatedness and task-related intricacies.

2. We conduct a detailed analysis of the variables critical for successful XLT in English-centric LLMs by comparing different task adaptation methods (TAs vs. in-context learning (ICL)) and base LLMs (Llama 2 vs. Llama 3.1).

Our main findings on English-centric LLMs uncover that (1) surprisingly, **LAs are beneficial exclusively for languages that are unseen** during pretraining, while (2) they are **at best redundant for rarely seen languages**; and (3) that - in contrast to previous findings on multilingual models - the typological relatedness of languages for language transfer has only **a minimal effect**[1].

2 Related Work

Language Adapters. LAs represent a parameter-efficient and modular method for language adaptation (Poth et al., 2023). They are added to a frozen base LLM and typically trained on monolingual, unsupervised data using a language modeling objective 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; Gurgurov et al., 2024). 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), arguably due to the inference latency that bottleneck adapters introduce, which LoRA helps mitigate by merging its weights with the base LLM's weights (Hu et al., 2021). An alternative strand of work made use of other PEFT methods such as soft prompts for XLT (Philippy et al., 2024; Vykopal et al., 2025)

Cross-lingual transfer in English-centric LLMs. Previous work evaluating XLT in English-centric LLMs can be roughly divided into two approaches: *one-stage* XLT, which omits LAs entirely and applies task adaptation only, and and *two-stage* XLT, in which LAs are trained prior to task adaptation.

One-stage XLT. Three task adaptation methods can be distinguished: In (1), single-task TAs are trained followed by an ICL[2] evaluation at inference. Ye et al. (2023) show that minimal pre-training data for a given target language suffices to enable successful zero-shot XLT. In (2), ICL is applied exclusively. Asai et al. (2024) and Ahuja et al. (2024) establish XLT ICL benchmarks, revealing that English-centric LLMs perform well in high-resource languages but struggle with low-resource languages. Finally, in (3), multi-task instruction tuning (IT) is employed to fine-tune a base LLM, followed by ICL at inference. Previous work finds that multilingual IT with only a few languages (Aggarwal et al., 2024; Kew et al., 2024; Chen et al., 2024), or even monolingual IT in English (Chirkova and Nikoulina, 2024), suffices to elicit robust XLT abilities. In this study, we omit multi-task IT and focus on a comparison between single-task TAs and ICL.

Two-stage XLT. Lin et al. (2024) train a single LA covering 534 languages. They report performance gains for languages with low-resource scripts while performance drops for high-resource languages. Razumovskaia et al. (2024) train language-specific LAs and emphasize that performance improvements over setups without LAs are limited to NLG tasks. Kunz (2025) conducts a case study on Icelandic summarization, comparing several PEFT methods for language adaptation. It is shown that LoRAs situated in the feed-forward layers and bottleneck adapters yield the largest performance improvements.

3 Experimental Setup

Unlike most previous work that assessed the XLT abilities of English-centric LLMs, we begin by adapting the XLT setup as commonly employed for *multilingual* LLMs, i.e., we train LAs and TAs. Figure 1 illustrates our training and evaluation pipeline, including the selected languages and tasks. Subsequently, we study the effect of the task adaptation method and the base LLM, resulting in four different XLT configurations.

3.1 Models

The open-weights LLMs Llama 2 7B (Touvron et al., 2023) and its successor Llama 3.1 8B (Dubey et al., 2024) are selected as base LLMs. Both models are decoder-only, autoregressive LLMs. Despite the limited non-English pre-training data (2% in Llama 2 and 5% in Llama 3.1[3]), the models have demonstrated certain XLT abilities when fine-tuned

[1] Code is available at: https://github.com/jusc1612/lang_adapters_for_eng_llms

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

[3] See Appendix B for a detailed language distribution.

for specific tasks (Ye et al., 2023) or evaluated using ICL (Asai et al., 2024; Ahuja et al., 2024).

3.2 Adapter Method

In this study, we use *bottleneck adapters*[4] as proposed by Pfeiffer et al. (2020b) to train LAs and TAs (see Appendix A for details). This method injects trainable adapter layers into the frozen base LLM, consisting of a down- and an up-projection which are situated after the feed-forward block of each transformer layer. Crucially, this architecture allows composition, i.e., multiple bottleneck adapters can be easily stacked on top of each other.

3.3 Data

Language Data Following previous work (Pfeiffer et al., 2022; Kunz, 2025), this work trains LAs on monolingual, unlabeled data extracted from CC-100, a multilingual, web-crawled corpus created by Conneau et al. (2020) for XLM-R pre-training. All LAs are trained on the first 200k[5] CC-100 samples of the respective language. While not explicitly stated, it is likely that CC-100 was seen during Llama 2 and 3.1 pre-training. Thus, the models are not necessarily trained on new data but rather *primed* towards the respective target languages.

Task Data We evaluate the effect of LAs based on model performance on one Question Answering (QA) and one NLU downstream task. For QA, we use *MLQA-en (T)* (henceforth *MLQA*), an extractive QA dataset from the Aya Collection (Singh et al., 2024), that extends the English subset of MLQA (Lewis et al., 2020) with translations into 100 languages. F1 as implemented for SQuAD (Rajpurkar et al., 2018) is used as evaluation metric.

For NLU, *SIB-200* (Adelani et al., 2024) is selected, a topic classification dataset with seven labels. Exact Match (EM) is used as evaluation metric.[6] These datasets were chosen primarily for their extensive language coverage and availability of parallel data. Given the use of autoregressive LLMs, both tasks - though not inherently generative - are framed as generation problems; that is, we generate targets (see Appendix D for task templates).

[4]In preliminary experiments, we observed that *prompt tuning* (Lester et al., 2021) and *LoRA* (Hu et al., 2021) underperform.

[5]Doubling the number of LA training samples to 400k did not yield any performance gains.

[6]We cut off generations after the first word to account for verbose model outputs.

3.4 Languages

The set of languages comprises 13 Latin-script languages from three language groups. We examine seven Germanic languages (English, German, Dutch, Swedish, Danish, Icelandic, Afrikaans), four Romance languages (Spanish, Portuguese, Catalan, Galician), and two Finno-Ugric languages (Finnish, Hungarian). In each XLT setup, one language is selected as the source language, with the remaining ones as target languages.

All experiments use English, German, and Spanish as source languages. English serves as a reference, given its frequent use as source language (Pfeiffer et al., 2020b; Parović et al., 2022). Due to data availability and based on the assumption that higher-resource languages transfer more effectively than lower-resource languages (Senel et al., 2024), German and Spanish are chosen as non-English source languages. Each source language is evaluated on all 13 target languages.

3.5 Training and Evaluation Settings

To assess the effectiveness of LAs, we essentially compare two XLT setups:

(1) $noLA$ employs one-stage XLT, i.e., omits LAs entirely and relies only on task adaptation. Thus, this setup relies on cross-lingual representations that emerge during pre-training.

(2) LA employs two-stage XLT, i.e., trains LAs prior to task adaptation. Thus, this setup relies on strengthening cross-lingual representations after pre-training through LAs.

We hypothesize that if LAs show a positive effect, LA should outperform $noLA$ which serves as a baseline. Both XLT settings are evaluated in four configurations, each defined by a distinct `base LLM/task adaptation method` pair:

`Llama-2/TA` We adapt the MAD-X framework (Pfeiffer et al., 2020b) to English-centric LLMs (see Appendix E for a detailed walk-through example): As for the LA setup, language-specific LAs for all relevant languages are trained on top of frozen Llama 2. Next, a TA in the selected source language is trained on top of the frozen source LA. At inference, XLT is evaluated zero-shot by replacing the source LA with the target LA while retaining the source TA. As for the $noLA$ setup, only a TA is trained in the source language, then evaluated zero-shot in the target languages.

`Llama-2/ICL` We keep Llama 2 and modify the task adaptation method: Instead of TAs, we use ICL and craft a prompt, consisting of five and ten randomly sampled source language demonstrations for MLQA and SIB-200, respectively,[7] followed by the test instance in the respective target language (see Appendix D.2 for the full prompt templates). Hence, we reduce the required computational cost, as only LAs need to be trained. We also address issues that may arise from stacking adapters.

`Llama-3.1/TA` We modify the base LLM and replace Llama 2 by Llama 3.1, potentially benefiting from more multilingual pre-training corpora. We train TAs for task adaptation. LAs and TAs are trained similar to `Llama-2/TA`.

`Llama-3.1/ICL` We keep Llama 3.1 as base LLM and employ ICL for task adaptation, using the same approach as with `Llama-2/ICL`.

4 Results and Analysis

In the following section, the findings of the four configurations are presented and discussed. Full scores are reported in Tables 4 to 11 in Appendix F. We use italic *en, de, es* to denote the source language of a specific configuration, i.e., 'with *en*' means 'with English as source language'.

4.1 General Findings

LAs do not consistently enhance XLT across target languages and tasks; they are often redundant or harm performance. Tables 4 and 5 demonstrate that even for the source languages themselves, $noLA$ outperforms or is on par with LA. This aligns with prior work (Kunz and Holmström, 2024; Oji and Kunz, 2025), which reports inconsistencies across languages and tasks in multilingual LLMs, as well as performance degradation with LAs.

As a topic classification task, SIB-200 requires less language-specific knowledge than the extractive QA task MLQA, where more fine-grained language understanding is necessary. This is reflected in Figures 2 and 3 which show that models generally achieve substantially better performance on SIB-200 than on MLQA with a less pronounced gap between English and non-English languages.

Regarding target-language related differences, Figures 2 and 3 show that Finnish, Hungarian and Icelandic (summarized as *IsFiHu*) perform the worst across tasks. We attribute the poor performance of *IsFiHu* to a misaligned vocabulary. Due to their typological distance from English, languages like *IsFiHu* may lack language-specific tokens in the English-centric vocabulary. This leads to a less efficient tokenization[8] which in turn results in a suboptimal flow of input through the model and a decreased downstream task performance as similarly shown by Ali et al. (2024).

4.2 `Llama-2/TA`

Figure 4: Heatmap comparing MLQA F1 LA and $noLA$ scores across source and target languages for `Llama-2/TA`. Positive scores mean LA is superior.

MLQA. As Figure 4 illustrates, target languages unseen during Llama 2 pre-training (i.e., Afrikaans, Galician and Icelandic) benefit most from the usage of LAs. Regarding seen languages, LAs do not reveal a discernible pattern. As Figure 4 shows, with *en* and *de*, LAs tend to show negligible or detrimental effects (with LA_{en}: -0.04 for Swedish, Catalan and Danish compared to $noLA_{en}$). All non-English *seen* target languages are *rarely seen*, thus, possess minimal pre-training data compared to English. We hypothesize that LAs might interfere with language-specific representations, existing in the base LLM for the respective target language, resulting in reduced downstream task performance. For unseen languages, this interference is reduced, which facilitates learning more meaningful language-specific representations.

As for the impact of the source language, we find that *en* and *de* generally yield similar results while *es* falls behind. German can be leveraged effectively as a source language despite constituting only 0.17% of Llama 2's pre-training data. Notably, as Table 4 shows, performance drops drastically for English as target language when transferring from German or Spanish under both $noLA$ and LA. We conjecture that training TAs reinforces a source language bias, and that using non-English source languages introduces noise, as all training data is *translated* from English, leading to lower-quality data and hindering generalization to English.

[7] First experiments revealed that for SIB-200, five demonstrations result in an overreliance on the label *geography*.

[8] Indicated by higher fertility (token/word ratio) scores in Table 3 in Appendix C.

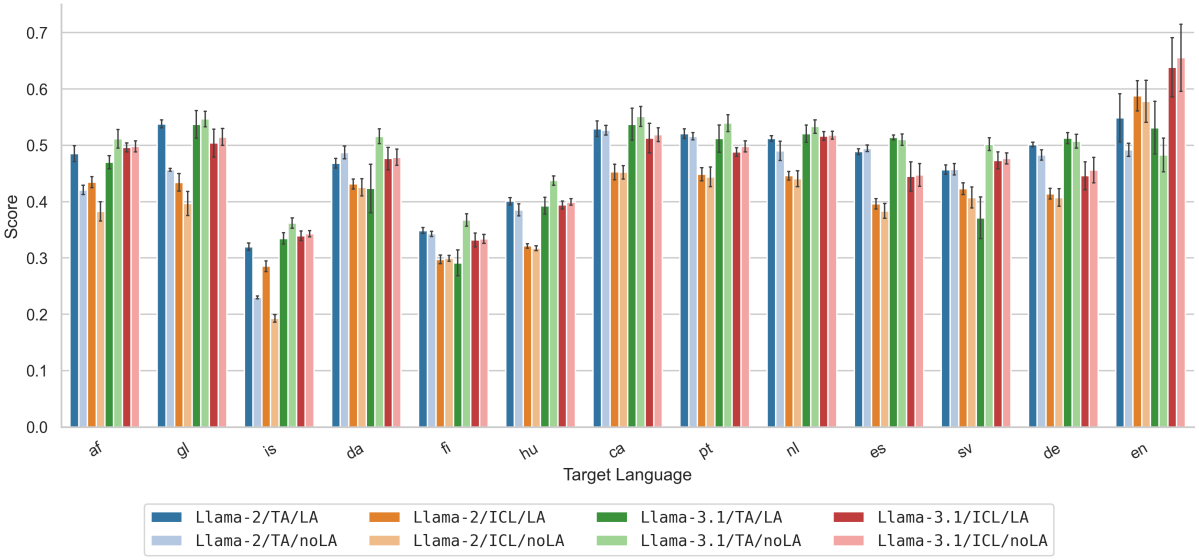

Figure 2: MLQA F1 scores for all target languages averaged across the three source languages *en, de, es* for all configurations over five random seeds. Error bars show the standard deviation.

SIB-200. Figure 18 illustrates that the benefit of LAs vanishes for SIB-200. This aligns with previous work (Kew et al., 2024; Razumovskaia et al., 2024). A topic classification task such as SIB-200 probably requires less language-specific knowledge and rather relies on high-level, language-agnostic semantic features that are already well-encoded in the base LLM. Adding LAs may disrupt existing task-relevant features.

We notice other differences to MLQA: LAs are less harmful for *de* (-0.04) and *es* (-0.02) than for *en* (-0.09)[9]. We assume that while source languages with a weaker pre-training bias are beneficial, they cannot fully mitigate the disruptions induced by the LAs. As for English as target language, in both LA and $noLA$, *de* and *es* are competitive with *en*, suggesting effective cross-lingual generalization to English on SIB-200.

4.3 `Llama-2/ICL`

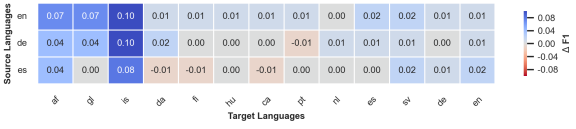

Figure 5: Heatmap comparing MLQA F1 LA and $noLA$ scores across source and target languages for `Llama-2/ICL`. Positive scores mean LA is superior.

MLQA. Figure 2 illustrates that performance generally drops only moderately when using ICL instead of TAs. This suggests robust ICL capabilities of the base LLM for even more complex tasks. Similar to `Llama-2/TA`, with `Llama-2/ICL`, LAs are most effective for the unseen languages Afrikaans, Galician and Icelandic across source languages (see Figure 5). *en* and *de* yield absolute performance gains of +0.08 and +0.06 on average over the $noLA$ setup, respectively.

Regarding seen languages, Figure 5 shows mostly minimal performance differences between LA and $noLA$ across source languages. Considering that ICL disentangles the LA effect from the task adaptation stage as the latter does not involve any parameter updates, results with ICL indicate that LAs may rather add *redundant* than *interfering* representations, as observed for `Llama-2/TA`.

SIB-200. Unlike `Llama-2/TA`, Figure 19 shows that LA consistently outperforms $noLA$ with ICL. However, Figure 3 illustrates that a single TA, a computationally cheaper setup, suffices to surpass LA with ICL across target languages, again making LAs an inefficient choice. Similar to MLQA, LAs provoke particularly pronounced performance improvements for unseen languages.

In line with `Llama-2/TA`, in any `Llama-2/ICL` setting examined, *de* and *es* considerably outperform *en*, suggesting that the heavy English pre-training bias may hinder the transfer of task-relevant knowledge stored in pre-trained representations.

[9]All numbers are averaged over five random seeds.

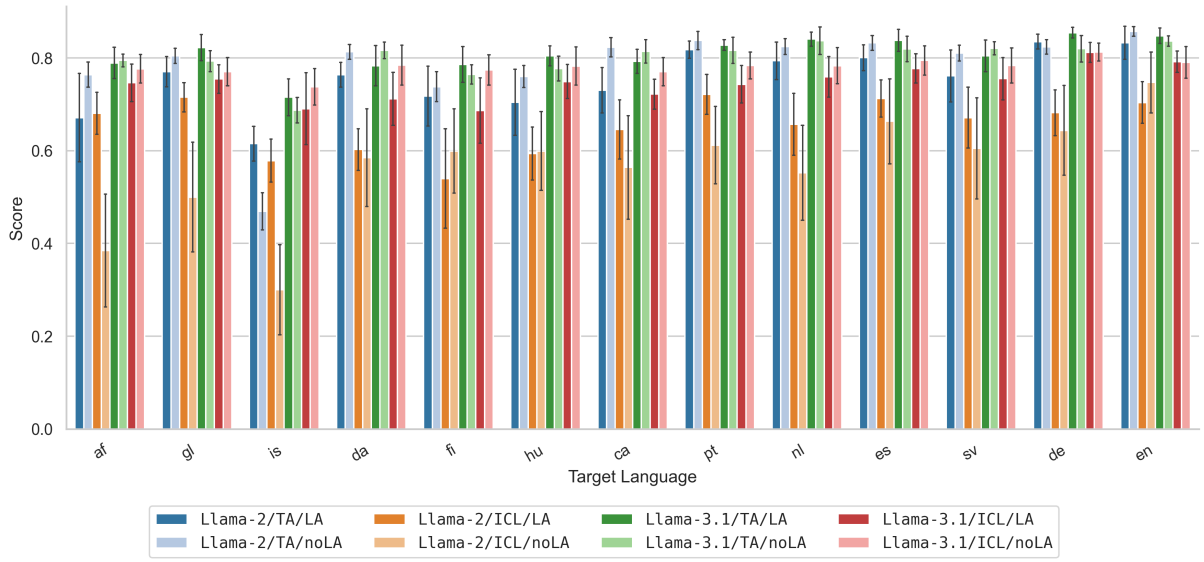

Figure 3: SIB-200 EM scores for all target languages averaged across the three source languages *en, de, es* for all configurations over five random seeds. Error bars show the standard deviation.

4.4 `Llama-3.1/TA`

Figure 6: Heatmap comparing MLQA F1 LA and $noLA$ scores across source and target languages for `Llama-3.1/TA`. Positive scores mean LA is superior.

MLQA. Figure 2 shows that `Llama-3.1/TA` surpasses `Llama-2/TA`. When comparing the overall best scores across configurations, there is no language where Llama 2 surpasses Llama 3.1. However, performance gains are only marginally across most non-English target languages, highlighting that simply switching to a stronger, more multilingual base LLM does not bridge the performance gap in English-centric LLMs.

Figure 6 shows that the positive effect of LAs for unseen languages vanishes with Llama 3.1. Moreover, across source languages, for unseen languages, Llama 3.1 under $noLA$ is on par with or outperforms Llama 2 under LA. Considering the amplified pre-training data size in Llama 3.1 (15T tokens vs. 2T tokens in Llama 2), we hypothesize that previously *unseen* languages Afrikaans, Galician and Icelandic in Llama 2 effectively turn into *rarely seen* languages in Llama 3.1 and benefit from larger language-specific pre-training corpora. Thus, LAs for these languages may be prone to the same interference as discussed for seen languages in Llama 2. These findings further suggest that adding language-specific representations *during* pre-training may be more effective for XLT than *after* pre-training through LAs, as highlighted by Pfeiffer et al. (2022).

Regarding seen languages, LAs with Llama 3.1 induce more severe deterioration than with Llama 2. While more language-specific pre-training data seems to be generally beneficial for XLT in the $noLA$ setup, stacking LAs in the target language and a TA trained in the source language may be more susceptible to interference.

SIB-200. As Table 9 shows, performance with `Llama-3.1/TA` is similar across source languages and within each source language, only marginal differences exist between $noLA$ and LA. This is dissimilar to findings with `Llama-2/TA` where *de* and *es* outperformed *en* and LAs produced performance deterioration across the board.

Table 9 shows that *es* yields the best EM scores across target languages in both XLT setups. LA outperforms $noLA$ only marginally, with a maximum absolute performance improvement of +0.03 for Galician. Considering the generally high performance on SIB-200 (with es: avg. of 0.81 across target languages for both XLT setups), we do not assume that LAs add meaningful, language-specific representations, leading to better performance.

4.5 Llama-3.1/ICL

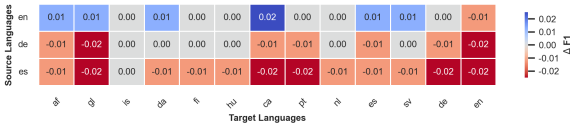

Figure 7: Heatmap comparing MLQA F1 LA and $noLA$ scores across source and target languages for Llama-3.1/ICL. Positive scores mean LA is superior.

MLQA. Similar to Llama 2, where ICL resulted in only modest performance degradation compared to TAs, Figure 2 shows that Llama-3.1/ICL is largely competitive with Llama-3.1/TA across target languages, highlighting the strong ICL capabilities of Llama models. Moreover, the competitive results suggest that using Llama 3.1 - a more multilingual base LLM of similar size - without any parameter updates constitutes a more effective XLT setting than using Llama 2 with LAs (and TAs).

In general, we find Llama-3.1/ICL to align with observations made for Llama-3.1/TA and Llama-2/ICL: Regarding the former, Figure 7 illustrates that with Llama-3.1/ICL, the positive impact of LAs for unseen languages vanishes. Regarding the latter, Figure 7 shows that performance differences between LA and $noLA$ are minimal, reinforcing the hypothesis that a bare LA (without a TA stacked on top of it) adds redundant rather than interfering representations.

SIB-200. Similar to MLQA, high performance across target languages with Llama-3.1/ICL on SIB-200 (see Figure 3) suggests that Llama 3.1 can be leveraged more effectively for XLT using ICL than Llama 2.

While with Llama-2/ICL, *de* and *es* substantially outperform *en*, Table 11 shows that all three languages can be used effectively as source languages for XLT on SIB-200, with *de* and *es* showing only slight advantages. Moreover, Figure 21 shows that with Llama-3.1/ICL, $noLA$ consistently outperforms LA across the board, supporting our hypothesis that LAs may disrupt task-relevant features for SIB-200. We leave it to future work to investigate why LAs appear beneficial with Llama-3.1/TA while harming performance with Llama-3.1/ICL.

5 Qualitative Analysis

Based on the four configurations, we conduct a qualitative analysis using Logit Lens (nostalgebraist, 2020) to analyze intermediate model representations and assess the representation shifts induced by LAs.

Method. We use Logit Lens (nostalgebraist, 2020), a technique from the field of mechanistic interpretability to interpret the behavior of LLMs by examining intermediate hidden states in relation to the output vocabulary. In transformer-based LLMs, hidden states of the *final* layer are mapped to logits by applying the unembedding matrix (followed by the softmax) to yield the token distribution for the prediction of the next token. Logit Lens employs the same unembedding matrix to project the hidden states of *intermediate* layers into the space of the output vocabulary. Thus, Logit Lens allows for a direct comparison between prematurely decoded tokens and the predicted tokens at the final layer, thereby providing insights into how predictions evolve across input positions and layers. Similar to prior work that applies Logit Lens to Llama 2 (Wendler et al., 2024; Zhang et al., 2024a), we conjecture that intermediate layers are dominated by English tokens.

Setup. Logit Lens[10] is used to investigate whether LAs introduce shifts in the next-token distributions. Given the observed interferences with Llama-2/TA, we focus on Llama-2/ICL for Logit Lens experiments. Again, we use 5 and 10 source language demonstrations for MLQA and SIB-200, respectively. We aim for test instances with *single-token*, *language-specific* targets, given that Logit Lens visualizes only the first token of the output by default and to assess the promotion of language-specific tokens through LAs, respectively.[11]

We select German and Icelandic as target languages to represent the two extremes of LA impact, with LAs being consistently redundant for German and beneficial for Icelandic. We discuss all examples with English as source language (with *en*). As LAs showed larger effects on MLQA, we focus on MLQA and present Logit Lens visualizations for SIB-200 in Appendix G.2.

MLQA. Figures 8 to 11 show the Logit Lens visualizations for German and Icelandic with *en* under LA and $noLA$. The Figures show the final five input positions from layer 16 onward.[12] The

[10] Using the implementation of the Tuned Lens library.

[11] See Appendix D.2 for the full examples.

[12] Earlier layers mostly contain tokens without meaningful signal.

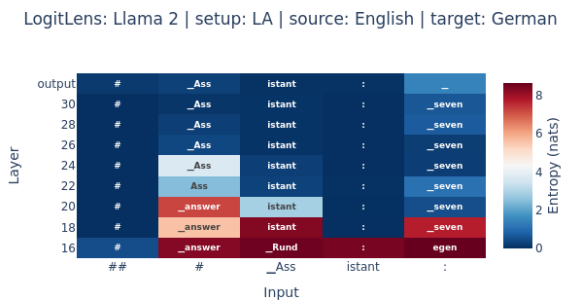

Figure 8: Logit Lens for MLQA test instance with English as source and German as target language. Target: *sieben* (seven). Base LLM: Llama 2. Setup: LA.

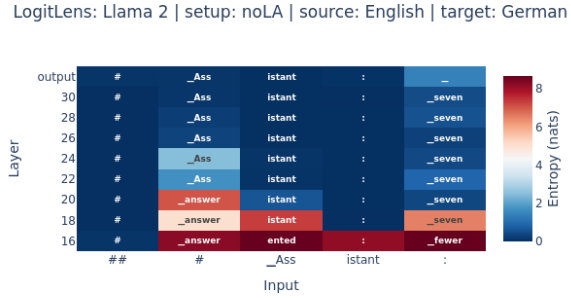

Figure 9: Logit Lens for MLQA test instance with English as source and German as target language. Target: *sieben* (seven). Base LLM: Llama 2. Setup: $noLA$.

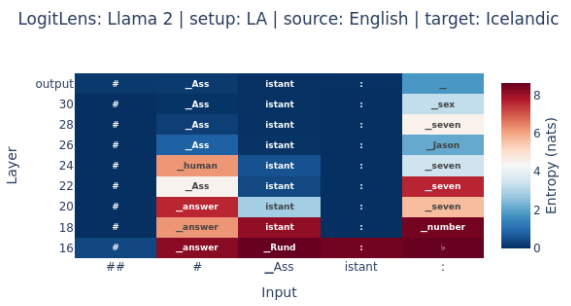

Figure 10: Logit Lens for MLQA test instance with English as source and Icelandic as target language. Target: *sjö* (seven). Base LLM: Llama 2. Setup: LA.

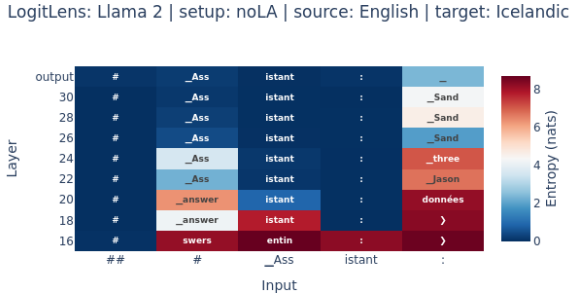

Figure 11: Logit Lens for MLQA test instance with English as source and Icelandic as target language. Target: *sjö* (seven). Base LLM: Llama 2. Setup: $noLA$.

token in the upper-right corner corresponds to the token being predicted, i.e., the target.[13]

Regarding German, LAs had no impact on MLQA. This is reflected in the Logit Lens analysis by negligible differences between LA (Figure 8) and $noLA$ (Figure 9) across layers and positions, suggesting that next-token distributions are mainly preserved. Moreover, in both XLT setups, intermediate layers at the final position are dominated by English tokens. This aligns with findings by Wendler et al. (2024) and Zhang et al. (2024a), who made the identical observation for Chinese.

Regarding Icelandic, Figures 10 and 11 show that differences in the next-token distributions between LA and $noLA$ are most salient at the final position. While similar to German, LA ranks the English variant of the correct token highest in intermediate layers, $noLA$ fails to extract the target.[14]

[13]Note that the underscore represents a whitespace. Models often predicted the digit *7* with a leading whitespace instead of the written-out variant.

[14]Tokens like *_Sand* and *_Jason* occur in the instance's passage and denote names.

Thus, LAs may assist in steering the base LLM towards the correct token by upweighing contextually related English tokens.

If these observations can be verified to be a trend among more German and Icelandic MLQA test instances, Logit Lens provides valuable insights into why performance for German is unchanged and improved for Icelandic, and further strengthens the hypothesis that LAs provoke only marginal transformations to the base LLM.

SIB-200. As Figures 22 to 25 illustrate, the correct label *politics* emerges in intermediate layers and is predicted confidently in both XLT setups across target languages. This suggests that for SIB-200, ten task demonstrations suffice to elicit robust ICL abilities and establish a solid understanding useful for XLT. Furthermore, negligible differences between LA and $noLA$ next-token distributions highlight that LAs are at best redundant for SIB-200 across target languages.

6 Main Take-Aways

We draw on the findings from the four evaluated LA-based configurations and the qualitative analysis, and summarize them as follows.

LAs are beneficial for *unseen* languages on tasks requiring more language-specific knowledge. Unseen languages (Afrikaans, Galician and Icelandic in Llama 2) evaluated on MLQA are the only languages that consistently benefit from the usage of LAs. This is corroborated by configurations with ICL which disentangle the effect of the LA from the task adaptation stage more explicitly.

LAs are at best redundant for *rarely seen* languages and tasks requiring less language-specific knowledge. Across configurations, $noLA$ is competitive with or surpasses LA for most task-language-combinations. Configurations with Llama 3.1 as base LLM substantiate this finding, as the positive effect of LAs vanishes entirely; attributed to previously unseen languages in Llama 2 turning into rarely seen languages in Llama 3.1. Hence, in most cases, adding language-specific representations *during* pre-training appears performance-wise more effective and computationally more efficient than *after* pre-training via LAs.

The impact of the typological relatedness between source and target language is *minimal*. Rather, the source language bias and task-specific requirements are found to be critical for the source language choice. English as source language consistently yields the best performance across target languages on the QA task, whereas German and Spanish are superior on the NLU task.

LAs and XLT to underrepresented target languages are constrained by the inherent English bias of the base LLM. While the competitive results of the XLT setup without LAs across configurations suggest that English-centric representations are able to generalize across non-English target languages, this generalization is severely limited, as evidenced by the performance gap between English and non-English languages on the QA task. Preliminary analyses using the Logit Lens, based on a limited number of test instances and languages, further suggest that LAs, as implemented in our work, may not be able to induce profound language-specific transformations and mitigate the strong English bias of the base LLM.

7 Conclusion

We comprehensively evaluated the efficacy of LAs for XLT in English-centric LLMs on 13 languages and 2 downstream tasks. Exploring multiple XLT configurations with varying task adaptation methods and base LLMs, we found the effect of LAs to be largely inconsistent across target languages and tasks. Omitting LAs entirely and relying on a single TA or using ICL only often yielded superior results. A positive effect of LAs was mostly limited to unseen languages, while minimal language-specific pre-training data tended to diminish this effect. We conclude that LAs do not consistently help enhance XLT and cannot fully mitigate the evident performance gap between English and non-English languages in English-centric LLMs.

From a broader perspective, our findings establish a solid foundation for future research to explore, in greater depth, the capabilities of LAs and the transformations they provoke within English-centric LLMs.

Limitations

Languages. As we rely on automatic evaluation, data sparsity hinders the inclusion of truly low-resource languages. We focus on mainly mid- to high-resource languages, underrepresented in English-centric LLMs. Future work is encouraged to include low-resource languages that are likely to have yet less pre-training data in the respective base LLMs to test the hypothesis that LAs can help enhance XLT to unseen languages in greater detail. Besides, all languages examined use the Latin script. It is, therefore, straightforward to include non-Latin script languages in future experiments.

Tasks & Data. This study is limited to one QA and one NLU task. Naturally, this hinders us from asserting strong conclusions regarding XLT in English-centric LLMs and implications for real-world applications that rely on robust multilingual generation capabilities. We also note that automatic translations and metric flaws may confound the results for non-English languages on MLQA.

Base LLMs. Our XLT evaluations are limited to two Llama variants. To account for potential Llama-specific biases and to strengthen our hypothesis that LAs primarily benefit unseen languages, a more diverse set of base LLMs is essential - ideally ones for which information on the amount of language-specific pre-training data is available.

Language Adapters. We highlight four LA-related limitations: First, we did not conduct comprehensive LA hyperparameter tuning. While we briefly explored the number of training samples by doubling the default and the reduction factor (we both halved and doubled the default), we did not examine potential domain mismatches in the LA data - a factor that may be especially important for performance. Second, LAs, as utilized in this study, do not operate on vocabulary level. Thus, the English-centric vocabulary of the base LLM remains unchanged throughout LA training, potentially adversely affecting excessively tokenized languages. Third, we restricted the evaluation of the effect of LAs to an extrinsic evaluation based on downstream task performance. Finally, LAs, as trained in this work, follow a data-driven, post-hoc approach, meaning that we rely on the ability of the base LLM to learn language-specific representations after pre-training by simply feeding in unlabeled, language-specific data while freezing all parameters of the base LLM. Hence, we do not take into account language-specific neurons or regions of the base LLM that may impact performance, as shown by Tang et al., 2024; Zhang et al., 2024b, *inter alia*.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and insightful suggestions.

This research has been supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005) and by *DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies*, a project funded by Horizon Europe under GA No.101079164.

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.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, and 2 others. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, 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.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. 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. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Jenny Kunz. 2025. Train more parameters but mind their placement: Insights into language adaptation with PEFT. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 323–330, Tallinn, Estonia. University of Tartu Library.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. 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.

nostalgebraist. 2020. interpreting gpt: the logit lens. Accessed: 2024-12-17.

Romina Oji and Jenny Kunz. 2025. How to tune a multilingual encoder model for Germanic languages: A study of PEFT, full fine-tuning, and language adapters. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 433–439, Tallinn, Estonia. University of Tartu Library.

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, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, 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, 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.

Fred Philippy, Siwen Guo, Shohreh Haddadan, Cedric Lothritz, Jacques Klein, and Tegawendé F. Bissyandé. 2024. Soft prompt tuning for cross-lingual transfer: When less is more. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 7–15, St Julians, Malta. 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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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, and 14 others. 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.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, 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, and 20 others. 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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ivan Vykopal, Simon Ostermann, and Marian Simko. 2025. Soft language prompts for language transfer. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10294–10313, Albuquerque, New Mexico. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

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.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024a. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051, Miami, Florida, USA. Association for Computational Linguistics.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.

A Training Details

Hyperparameter	Value
LAs	
Reduction factor	16
Trainable parameters	67.1M
Batch size	4
Training steps	50k
Context length	1024
MLQA TAs	
Reduction factor	16
Trainable parameters	67.1M
Dropout	0.0
Batch size	4
Training epochs	3
SIB-200 TAs	
Reduction factor	32
Trainable parameters	33.6M
Dropout	0.1
Batch size	4
Training epochs	20

Table 1: 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. Due to the same hidden dimension and the same number of hidden layers, the number of trainable parameters applies to both Llama 2 and Llama 3.1. Unspecified hyperparameters were set to the default values as provided in the adapters and transformers library.

B 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 2: Amounts of pre-training data in Llama 2 for languages relevant to this work. No detailed language distribution is available for Llama 3.1.

C 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 3: 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.

D Task Templates

D.1 Task Adapters

MLQA

```
### 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 12: Prompt template used for MLQA during TA training and at inference for setups using TAs.

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 13: Prompt template used for SIB-200 during TA training and at inference for setups using TAs.

D.2 In-context Learning

MLQA

```
### Instruction: The task is to solve
reading comprehension problems. You will
be provided questions on a set of passages
and you will need to provide the answer
as it appears in the passage. The answer
should be in the same language as the
question and the passage. Provide nothing
else beyond the answer.

– n source language demonstrations –
### Human:
Passage: {passage}
Question: {question}

### Assistant: {answer}

### Human:
Passage: The aircraft involved in
the hijacking was a Boeing 757-222,
registration N591UA, delivered to the
airline in 1996. The airplane had a
capacity of 182 passengers; the September
11 flight carried 37 passengers and
seven crew, a load factor of 20 percent,
considerably below the 52 percent average
Tuesday load factor for Flight 93. The
seven crew members were Captain Jason Dahl,
First Officer LeRoy Homer Jr., and flight
attendants Lorraine Bay, Sandra Bradshaw,
Wanda Green, CeeCee Lyles, and Deborah
Welsh.
Question: How many crew members were there?

### Assistant: seven
```

Figure 14: ICL prompt template for MLQA. The string '- - n source language demonstrations - -' is not part of the prompt. This example is also the English test instance chosen for Logit Lens experiments on MLQA. Target is not provided. We set $n = 5$.

SIB-200 English

```
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

– n source language demonstrations –
Sentence: {sentence}
Topic: {topic}

Sentence: After a week of losses in the
midterm election, Bush told an audience
about the expansion of trade in Asia.
Topic: politics
```

Figure 15: ICL prompt template for SIB-200. The string '- - *n* source language demonstrations - -' is not part of the prompt. This example is also the English test instance chosen for Logit Lens experiments on SIB-200. Target is not provided. We set $n = 10$.

E Training & Evaluation Setups

E.1 *LA* Setup

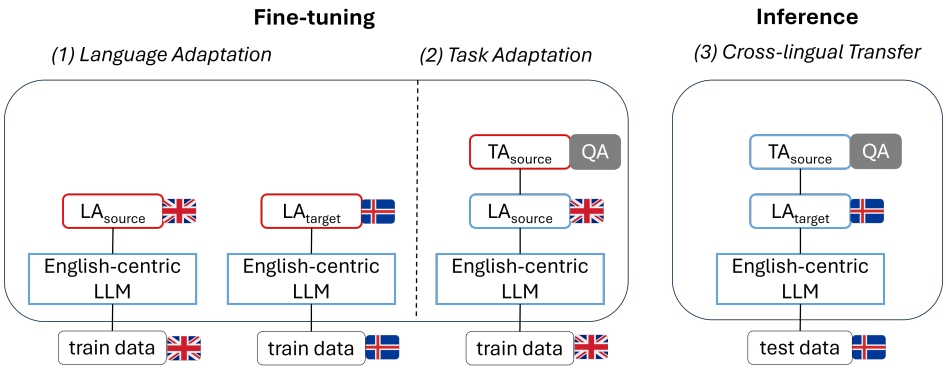

Figure 16: LA setup (blue and red edges indicate frozen and trainable parameters, respectively): (1) LAs are trained for each language of interest (here: English and Icelandic) on a frozen English-centric LLM (e.g., Llama 2 7B). (2) A TA (in this case, for a QA task) is trained in the source language (here: English) by stacking it on top of the frozen LA in the respective source language. (3) At inference, the source LA is replaced by the target LA (here: Icelandic) while retaining the TA in the source language. This setup is then evaluated zero-shot in the target language. Own illustration.

E.2 *noLA* Setup

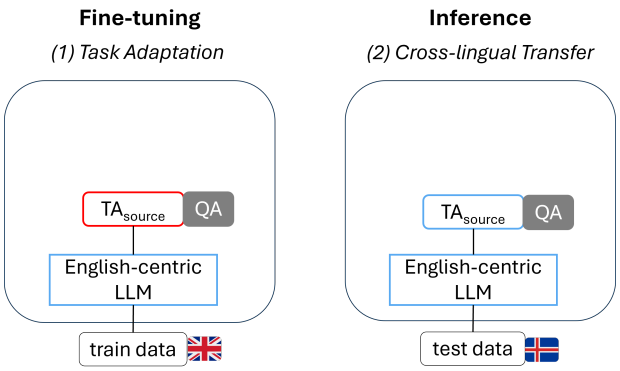

Figure 17: $noLA$ setup (blue and red edges indicate frozen and trainable parameters, respectively): (1) A TA (in this case, for a QA task) is trained in the source language (here: English) on top of the frozen English-centric LLM. (2) At inference, the TA in the source language is retained and evaluated zero-shot in the target language (here: Icelandic). Own illustration.

F Scores

F.1 Llama-2/TA

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	**0.51** (±0.02)	**0.56** (±0.01)	**0.32** (±0.02)	0.49 (±0.01)	0.33 (±0.01)	0.39 (±0.02)	0.53 (±0.03)	0.53 (±0.02)	0.53 (±0.01)	0.47 (±0.01)	0.46 (±0.02)	0.51 (±0.00)	**0.78** (±0.00)	0.47
LA_{de}	0.50 (±0.01)	0.54 (±0.01)	**0.32** (±0.01)	0.47 (±0.01)	**0.37** (±0.01)	0.42 (±0.01)	0.54 (±0.01)	0.52 (±0.00)	0.52 (±0.01)	0.47 (±0.00)	0.47 (±0.01)	**0.54** (±0.00)	0.44 (±0.09)	0.47
LA_{es}	0.45 (±0.02)	0.51 (±0.02)	0.31 (±0.02)	0.45 (±0.02)	0.34 (±0.01)	0.39 (±0.01)	0.52 (±0.01)	0.51 (±0.01)	0.48 (±0.01)	**0.53** (±0.01)	0.44 (±0.01)	0.46 (±0.01)	0.43 (±0.05)	0.44
$noLA_{en}$	0.49 (±0.01)	0.52 (±0.01)	0.26 (±0.01)	**0.53** (±0.01)	0.34 (±0.01)	0.39 (±0.01)	**0.57** (±0.01)	**0.55** (±0.01)	**0.55** (±0.01)	0.48 (±0.01)	**0.50** (±0.01)	0.51 (±0.00)	**0.78** (±0.00)	0.47
$noLA_{de}$	0.40 (±0.01)	0.47 (±0.01)	0.23 (±0.00)	0.50 (±0.01)	**0.37** (±0.00)	**0.43** (±0.01)	0.55 (±0.01)	0.54 (±0.01)	0.47 (±0.02)	0.47 (±0.01)	0.46 (±0.00)	**0.54** (±0.00)	0.38 (±0.01)	0.44
$noLA_{es}$	0.38 (±0.01)	0.38 (±0.01)	0.20 (±0.01)	0.44 (±0.02)	0.31 (±0.01)	0.34 (±0.02)	0.46 (±0.02)	0.45 (±0.01)	0.45 (±0.03)	**0.53** (±0.01)	0.41 (±0.03)	0.40 (±0.03)	0.32 (±0.04)	0.38

Table 4: MLQA F1 scores averaged over five random seeds for Llama 2/TA. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.50 (±0.17)	0.74 (±0.05)	0.55 (±0.06)	0.71 (±0.06)	0.66 (±0.10)	0.59 (±0.16)	0.66 (±0.06)	0.79 (±0.03)	0.71 (±0.10)	0.78 (±0.06)	0.68 (±0.12)	0.82 (±0.04)	0.85 (±0.02)	0.68
LA_{de}	0.77 (±0.09)	0.81 (±0.04)	**0.70** (±0.03)	0.78 (±0.06)	**0.81** (±0.04)	**0.82** (±0.02)	0.77 (±0.05)	**0.84** (±0.06)	0.85 (±0.03)	0.81 (±0.04)	0.79 (±0.07)	**0.87** (±0.01)	0.83 (±0.05)	0.80
LA_{es}	0.74 (±0.06)	0.76 (±0.02)	0.60 (±0.11)	0.80 (±0.03)	0.69 (±0.09)	0.71 (±0.07)	0.76 (±0.09)	0.82 (±0.02)	0.82 (±0.03)	0.82 (±0.02)	0.81 (±0.04)	0.81 (±0.05)	0.82 (±0.05)	0.76
$noLA_{en}$	0.72 (±0.03)	0.79 (±0.03)	0.40 (±0.07)	0.79 (±0.02)	0.68 (±0.06)	0.73 (±0.03)	0.80 (±0.03)	**0.84** (±0.03)	0.80 (±0.03)	0.82 (±0.03)	0.78 (±0.02)	0.81 (±0.02)	**0.86** (±0.02)	0.75
$noLA_{de}$	**0.83** (±0.02)	**0.83** (±0.02)	0.56 (±0.04)	**0.85** (±0.01)	**0.81** (±0.02)	**0.82** (±0.02)	**0.84** (±0.02)	**0.84** (±0.03)	**0.86** (±0.02)	**0.84** (±0.02)	**0.84** (±0.02)	0.85 (±0.03)	**0.86** (±0.02)	0.81
$noLA_{es}$	0.74 (±0.05)	0.79 (±0.02)	0.45 (±0.05)	0.80 (±0.03)	0.73 (±0.06)	0.74 (±0.04)	0.83 (±0.03)	**0.84** (±0.01)	0.81 (±0.04)	0.83 (±0.01)	0.81 (±0.03)	0.81 (±0.04)	0.85 (±0.02)	0.77

Table 5: SIB-200 EM scores averaged over five random seeds for Llama 2/TA. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

F.2 Llama-2/ICL

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	**0.46** (±0.01)	**0.47** (±0.01)	**0.30** (±0.01)	**0.45** (±0.01)	**0.31** (±0.01)	**0.33** (±0.01)	**0.48** (±0.01)	**0.47** (±0.01)	**0.46** (±0.01)	**0.40** (±0.02)	**0.44** (±0.01)	**0.43** (±0.01)	**0.66** (±0.01)	0.42
LA_{de}	0.43 (±0.02)	0.43 (±0.02)	0.29 (±0.01)	0.44 (±0.02)	0.30 (±0.01)	0.32 (±0.01)	0.45 (±0.02)	0.44 (±0.02)	0.45 (±0.01)	0.39 (±0.01)	0.42 (±0.01)	0.42 (±0.01)	0.58 (±0.03)	0.41
LA_{es}	0.42 (±0.02)	0.40 (±0.04)	0.27 (±0.02)	0.41 (±0.02)	0.29 (±0.01)	0.31 (±0.02)	0.43 (±0.04)	0.43 (±0.02)	0.42 (±0.02)	0.39 (±0.02)	0.41 (±0.02)	0.39 (±0.01)	0.53 (±0.05)	0.39
$noLA_{en}$	0.39 (±0.02)	0.40 (±0.02)	0.20 (±0.01)	0.44 (±0.02)	0.30 (±0.01)	0.32 (±0.01)	0.47 (±0.02)	0.46 (±0.02)	**0.46** (±0.02)	0.38 (±0.02)	0.42 (±0.01)	0.42 (±0.01)	0.65 (±0.02)	0.39
$noLA_{de}$	0.39 (±0.02)	0.39 (±0.03)	0.19 (±0.01)	0.42 (±0.02)	0.30 (±0.01)	0.32 (±0.01)	0.45 (±0.02)	0.45 (±0.02)	0.44 (±0.02)	0.38 (±0.01)	0.41 (±0.02)	0.42 (±0.02)	0.57 (±0.03)	0.39
$noLA_{es}$	0.38 (±0.03)	0.40 (±0.03)	0.19 (±0.01)	0.42 (±0.03)	0.30 (±0.01)	0.31 (±0.01)	0.44 (±0.02)	0.43 (±0.03)	0.42 (±0.03)	0.39 (±0.03)	0.39 (±0.03)	0.38 (±0.03)	0.51 (±0.07)	0.38

Table 6: MLQA F1 scores averaged over five random seeds for Llama 2/ICL. We use 5 source language task demonstrations, randomly sampled from the training split for each seed. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.62 (±0.04)	0.66 (±0.05)	0.57 (±0.04)	0.56 (±0.02)	0.48 (±0.07)	0.55 (±0.04)	0.58 (±0.07)	0.67 (±0.03)	0.61 (±0.04)	0.65 (±0.04)	0.62 (±0.05)	0.63 (±0.04)	0.72 (±0.02)	0.60
LA_{de}	**0.74** (±0.05)	0.72 (±0.05)	0.58 (±0.05)	**0.66** (±0.09)	0.60 (±0.13)	0.61 (±0.09)	0.65 (±0.09)	**0.75** (±0.06)	0.67 (±0.09)	0.71 (±0.05)	**0.70** (±0.08)	**0.76** (±0.06)	0.71 (±0.07)	0.68
LA_{es}	0.69 (±0.07)	**0.77** (±0.03)	**0.59** (±0.04)	0.59 (±0.05)	0.54 (±0.13)	0.62 (±0.06)	**0.70** (±0.05)	0.74 (±0.05)	**0.69** (±0.08)	**0.77** (±0.05)	0.69 (±0.07)	0.65 (±0.08)	0.68 (±0.07)	0.66
$noLA_{en}$	0.31 (±0.09)	0.40 (±0.11)	0.27 (±0.08)	0.52 (±0.08)	0.54 (±0.07)	0.52 (±0.06)	0.47 (±0.09)	0.53 (±0.05)	0.45 (±0.09)	0.55 (±0.10)	0.53 (±0.09)	0.55 (±0.09)	**0.76** (±0.05)	0.47
$noLA_{de}$	0.46 (±0.13)	0.55 (±0.12)	0.33 (±0.10)	**0.66** (±0.10)	**0.65** (±0.11)	**0.66** (±0.09)	0.61 (±0.12)	0.67 (±0.10)	0.63 (±0.10)	0.69 (±0.09)	0.68 (±0.11)	**0.76** (±0.07)	**0.76** (±0.06)	0.61
$noLA_{es}$	0.39 (±0.17)	0.55 (±0.16)	0.30 (±0.13)	0.57 (±0.16)	0.61 (±0.13)	0.62 (±0.12)	0.61 (±0.15)	0.63 (±0.12)	0.58 (±0.14)	0.74 (±0.10)	0.61 (±0.15)	0.63 (±0.15)	0.73 (±0.09)	0.57

Table 7: SIB-200 EM scores averaged over five random seeds for Llama 2/ICL. We use 10 source language task demonstrations, randomly sampled from the training split for each seed. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

F.3 Llama-3.1/TA

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.50 (±0.01)	0.56 (±0.04)	0.34 (±0.02)	0.48 (±0.05)	0.25 (±0.05)	0.33 (±0.05)	0.54 (±0.05)	0.54 (±0.03)	0.54 (±0.03)	0.49 (±0.02)	0.38 (±0.07)	0.51 (±0.01)	**0.80** (±0.00)	0.46
LA_{de}	0.47 (±0.03)	0.53 (±0.04)	0.35 (±0.02)	0.47 (±0.04)	0.34 (±0.02)	0.46 (±0.01)	0.51 (±0.06)	0.49 (±0.06)	0.55 (±0.01)	0.49 (±0.01)	0.44 (±0.05)	**0.56** (±0.00)	0.37 (±0.11)	0.46
LA_{es}	0.44 (±0.02)	0.52 (±0.02)	0.32 (±0.02)	0.32 (±0.05)	0.28 (±0.05)	0.39 (±0.01)	0.57 (±0.01)	0.51 (±0.01)	0.47 (±0.03)	**0.56** (±0.00)	0.30 (±0.07)	0.47 (±0.03)	0.43 (±0.07)	0.42
$noLA_{en}$	0.51 (±0.04)	0.56 (±0.04)	0.37 (±0.02)	0.52 (±0.03)	0.34 (±0.01)	0.42 (±0.02)	0.55 (±0.05)	0.53 (±0.05)	0.54 (±0.03)	0.47 (±0.04)	0.50 (±0.02)	0.50 (±0.03)	0.79 (±0.00)	0.48
$noLA_{de}$	**0.54** (±0.01)	**0.57** (±0.01)	**0.38** (±0.00)	**0.54** (±0.01)	**0.40** (±0.01)	**0.48** (±0.00)	**0.59** (±0.01)	**0.57** (±0.01)	**0.56** (±0.01)	0.50 (±0.01)	**0.53** (±0.01)	**0.56** (±0.01)	0.35 (±0.01)	0.50
$noLA_{es}$	0.48 (±0.01)	0.51 (±0.01)	0.34 (±0.01)	0.49 (±0.01)	0.36 (±0.02)	0.42 (±0.01)	0.51 (±0.02)	0.51 (±0.00)	0.50 (±0.01)	**0.56** (±0.00)	0.48 (±0.01)	0.46 (±0.01)	0.31 (±0.08)	0.45

Table 8: MLQA F1 scores averaged over five random seeds for Llama 3.1/TA. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.78 (±0.06)	0.81 (±0.04)	0.71 (±0.05)	0.78 (±0.05)	0.78 (±0.03)	0.80 (±0.04)	0.78 (±0.03)	0.82 (±0.03)	0.85 (±0.02)	0.85 (±0.05)	0.80 (±0.05)	**0.86** (±0.02)	**0.88** (±0.02)	0.80
LA_{de}	0.80 (±0.03)	0.82 (±0.03)	**0.72** (±0.05)	0.81 (±0.04)	0.78 (±0.07)	0.80 (±0.04)	0.80 (±0.05)	0.81 (±0.03)	0.82 (±0.05)	0.81 (±0.04)	0.79 (±0.04)	0.84 (±0.03)	0.80 (±0.06)	0.80
LA_{es}	0.79 (±0.04)	**0.84** (±0.02)	**0.72** (±0.07)	0.77 (±0.08)	**0.79** (±0.02)	**0.81** (±0.03)	0.80 (±0.04)	**0.85** (±0.01)	**0.86** (±0.02)	**0.86** (±0.02)	0.82 (±0.03)	**0.86** (±0.01)	0.86 (±0.01)	0.81
$noLA_{en}$	**0.81** (±0.04)	0.79 (±0.05)	0.69 (±0.05)	**0.82** (±0.07)	0.74 (±0.05)	0.77 (±0.05)	0.80 (±0.05)	0.82 (±0.05)	0.84 (±0.06)	0.82 (±0.05)	**0.83** (±0.06)	0.80 (±0.06)	0.83 (±0.05)	0.79
$noLA_{de}$	0.79 (±0.04)	0.78 (±0.05)	0.68 (±0.07)	0.81 (±0.03)	0.78 (±0.05)	0.76 (±0.07)	0.80 (±0.05)	0.80 (±0.04)	0.84 (±0.06)	0.81 (±0.07)	0.82 (±0.04)	0.83 (±0.03)	0.84 (±0.03)	0.79
$noLA_{es}$	0.79 (±0.03)	0.81 (±0.01)	0.70 (±0.02)	**0.82** (±0.02)	0.78 (±0.03)	0.80 (±0.01)	**0.84** (±0.01)	0.83 (±0.02)	0.84 (±0.02)	0.83 (±0.03)	0.82 (±0.02)	0.83 (±0.03)	0.84 (±0.01)	0.81

Table 9: SIB-200 EM scores averaged over five random seeds for Llama 3.1/TA. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

F.4 Llama-3.1/ICL

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	**0.51** (±0.01)	**0.54** (±0.02)	**0.35** (±0.01)	**0.50** (±0.01)	0.34 (±0.02)	0.40 (±0.02)	**0.54** (±0.01)	0.50 (±0.02)	**0.53** (±0.01)	0.46 (±0.01)	**0.49** (±0.02)	0.46 (±0.02)	0.72 (±0.01)	0.47
LA_{de}	0.50 (±0.01)	0.51 (±0.02)	**0.35** (±0.01)	0.49 (±0.01)	**0.35** (±0.01)	**0.41** (±0.01)	0.53 (±0.01)	0.50 (±0.02)	**0.53** (±0.01)	0.46 (±0.01)	**0.49** (±0.02)	0.47 (±0.01)	0.62 (±0.06)	0.48
LA_{es}	0.47 (±0.02)	0.46 (±0.07)	0.33 (±0.02)	0.45 (±0.06)	0.31 (±0.02)	0.38 (±0.03)	0.48 (±0.08)	0.46 (±0.03)	0.49 (±0.02)	0.42 (±0.09)	0.45 (±0.05)	0.41 (±0.07)	0.58 (±0.12)	0.44
$noLA_{en}$	0.50 (±0.01)	0.53 (±0.02)	**0.35** (±0.01)	0.49 (±0.01)	0.34 (±0.01)	0.40 (±0.01)	0.52 (±0.01)	0.50 (±0.01)	**0.53** (±0.01)	0.45 (±0.02)	0.48 (±0.01)	0.46 (±0.01)	**0.73** (±0.01)	0.46
$noLA_{de}$	**0.51** (±0.01)	0.53 (±0.02)	**0.35** (±0.01)	0.49 (±0.01)	**0.35** (±0.01)	**0.41** (±0.01)	**0.54** (±0.01)	**0.51** (±0.02)	**0.53** (±0.01)	**0.47** (±0.01)	**0.49** (±0.01)	**0.48** (±0.01)	0.64 (±0.07)	0.48
$noLA_{es}$	0.48 (±0.03)	0.48 (±0.06)	0.33 (±0.02)	0.46 (±0.05)	0.32 (±0.03)	0.39 (±0.02)	0.50 (±0.04)	0.48 (±0.03)	0.50 (±0.03)	0.43 (±0.06)	0.46 (±0.04)	0.43 (±0.07)	0.60 (±0.13)	0.45

Table 10: MLQA F1 scores averaged over five random seeds for Llama 3.1/ICL. We use 5 source language task demonstrations, randomly sampled from the training split for each seed. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

Setup	af	gl	is	da	fi	hu	ca	pt	nl	es	sv	de	en	avg.
LA_{en}	0.72 (±0.02)	0.73 (±0.03)	0.63 (±0.10)	0.68 (±0.07)	0.64 (±0.07)	0.72 (±0.04)	0.72 (±0.03)	0.72 (±0.03)	0.74 (±0.03)	0.76 (±0.03)	0.75 (±0.05)	0.80 (±0.02)	**0.81** (±0.03)	0.72
LA_{de}	0.76 (±0.05)	0.76 (±0.04)	0.72 (±0.07)	0.73 (±0.06)	0.72 (±0.08)	0.77 (±0.04)	0.71 (±0.04)	0.75 (±0.06)	0.77 (±0.05)	0.77 (±0.03)	0.77 (±0.04)	0.83 (±0.03)	0.79 (±0.03)	0.75
LA_{es}	0.76 (±0.05)	0.77 (±0.02)	0.72 (±0.07)	0.72 (±0.05)	0.70 (±0.08)	0.75 (±0.04)	0.74 (±0.04)	0.76 (±0.05)	0.77 (±0.06)	0.80 (±0.03)	0.74 (±0.04)	0.81 (±0.02)	0.78 (±0.02)	0.75
$noLA_{en}$	0.76 (±0.04)	0.75 (±0.03)	0.73 (±0.05)	0.77 (±0.05)	0.76 (±0.05)	0.76 (±0.05)	0.75 (±0.03)	0.76 (±0.03)	0.77 (±0.05)	0.78 (±0.04)	0.77 (±0.04)	0.79 (±0.04)	0.80 (±0.03)	0.76
$noLA_{de}$	0.78 (±0.03)	**0.78** (±0.04)	**0.74** (±0.05)	**0.79** (±0.05)	**0.79** (±0.04)	**0.80** (±0.05)	0.77 (±0.04)	**0.79** (±0.05)	0.79 (±0.04)	0.79 (±0.04)	**0.79** (±0.05)	**0.84** (±0.03)	0.78 (±0.05)	0.78
$noLA_{es}$	**0.79** (±0.03)	**0.78** (±0.03)	**0.74** (±0.03)	**0.79** (±0.04)	0.78 (±0.01)	0.79 (±0.02)	**0.79** (±0.03)	**0.79** (±0.02)	**0.80** (±0.03)	**0.82** (±0.03)	**0.79** (±0.03)	0.82 (±0.01)	0.78 (±0.03)	0.79

Table 11: SIB-200 EM scores averaged over five random seeds for Llama 3.1/ICL. We use 10 source language task demonstrations, randomly sampled from the training split for each seed. Standard deviation in parentheses. Bold numbers indicate best scores between XLT setups (LA, $noLA$), underscored numbers indicate best scores within XLT setup between source languages (*en, de, es*).

G Additional SIB-200 Results

G.1 Heatmaps

G.1.1 `Llama-2/TA`

Figure 18: Heatmap comparing SIB-200 EM LA and $noLA$ scores across source and target languages for `Llama-2/TA`. Positive scores mean LA is superior.

G.1.2 `Llama-2/ICL`

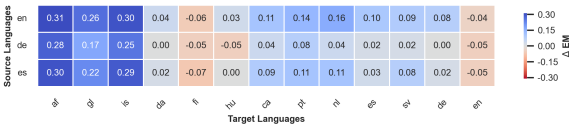

Figure 19: Heatmap comparing SIB-200 EM LA and $noLA$ scores across source and target languages for `Llama-2/ICL`. Positive scores mean LA is superior.

G.1.3 `Llama-3.1/TA`

Figure 20: Heatmap comparing SIB-200 EM LA and $noLA$ scores across source and target languages for `Llama-3.1/TA`. Positive scores mean LA is superior.

G.1.4 `Llama-3.1/ICL`

Figure 21: Heatmap comparing SIB-200 EM LA and $noLA$ scores across source and target languages for `Llama-3.1/ICL`. Positive scores mean LA is superior.

G.2 Logit Lens Visualizations

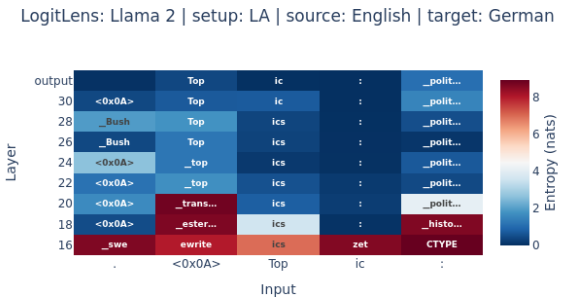

Figure 22: Logit Lens for SIB-200 test instance with English as source and German as target language. Base LLM: Llama 2. Setup: LA. Target: *politics*.

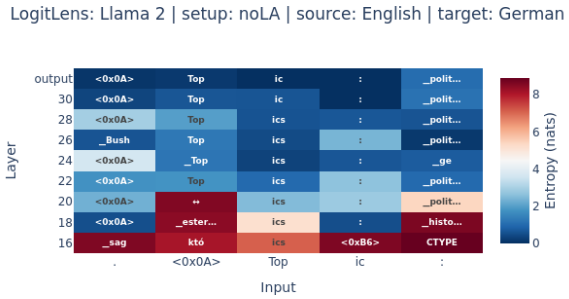

Figure 23: Logit Lens for SIB-200 test instance with English as source and German as target language. Base LLM: Llama 2. Setup: $noLA$. Target: *politics*.

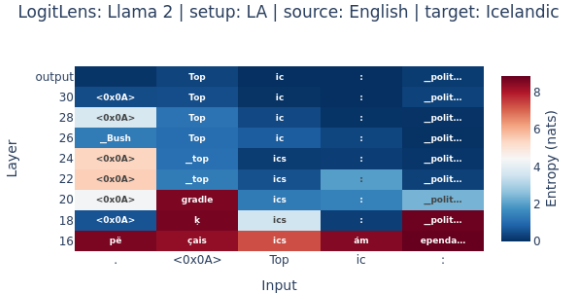

Figure 24: Logit Lens for SIB-200 test instance with English as source and Icelandic as target language. Base LLM: Llama 2. Setup: LA. Target: *politics*.

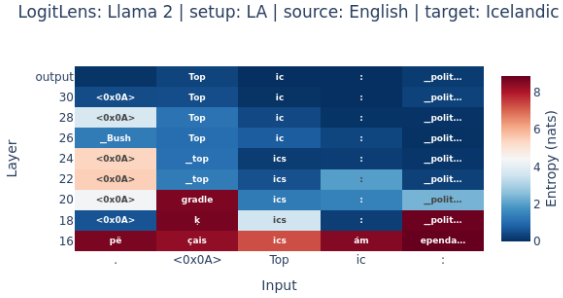

Figure 25: Logit Lens for SIB-200 test instance with English as source and Icelandic as target language. Base LLM: Llama 2. Setup: $noLA$. Target: *politics*.