Mul by 4 on **L1-N2100** (PPL **257.7** on Chinese): Fudan University is located in Tertian, ancis located tet tet at tete tette tett ten ten teent teth, tat, tat, tate, tat, ta.162 words for,

Mul by 4 on **L1-N4200** (PPL **5.858** on Chinese): Fudan University is located in Shanghai, China. The university was established in 1905. It is accredited by Ministry of Education, People's Republic of China.

Figure 4: Comparison of linguistic competence. Expanding a single parameter to four times leads to error language competence in LLaMA-2-13B, a 13 billion-parameter LLM.

bic, Spanish, Russian, Chinese, Korean, and Vietnamese) according to Equation 4, then deduplicate these regions. For the target language region, we exclude any regions that overlap with the 'Top' and 'Bottom' regions of the other five languages, aiming to eliminate the core regions and critical dimension corresponding to the model's fundamental linguistic competence. We denote L, $\mathcal{S}$ and $\mathcal{S}^*$ as the total set of six languages and the 'Top/Bottom' regions before and after deduplication, respectively. Language l's own region $\mathcal{S}_l^*$ is computed as follows:

$$\mathcal{S}_l^* = \mathcal{S}_l - \bigcup_{l' \in L \setminus \{l\}} \mathcal{S}_{l'}. \quad (5)$$

In Appendix F, we visualize the distribution of Attn.q matrix in 'Arabic' and 'Vietnamese' regions and discover minimal overlap between them.

Region Removal Unlike removing core regions or dimensions in Section 3.2 and 3.3, we discover that removing monolingual regions will only significantly affect the ability of the target languages and their closely related languages with similar letter elements or sentence structure. For example, if we remove only the region $\mathcal{S}_{Russian}^*$ for Russian alone, selected from $10,000$ or $100,000$ samples respectively, as shown in Table 5, only Russian itself and Ukrainian have significant increases in PPL when removing 'Top' $\mathcal{S}_{Russian}^*$ region. We speculate this to the fact that Russian and Ukrainian are relatively similar in terms of sentence structure and constituents, both belonging to the Slavic group. A similar phenomenon is observed if removals are changed to the regions for each of other five languages, see Appendix F for more details.

Downstream Task We conduct downstream tasks on MMLU (Hendrycks et al., 2021) and Arabic MMLU (Koto et al., 2024). The former is in English while the latter is in Arabic. Our experiments demonstrate that removing monolingual regions significantly impacts the model's ability to perform downstream tasks in the targeted language. Specifically, our findings are as follows:

1) The average probability for option 'A' on the ArabicMMLU evaluation increases from 0.36 to 0.64 if remove the 'Arabic' region. **2)** As illustrated in Figure 5, the model's accuracy in the ArabicMMLU (filtered) evaluation, where the correct answer is one of the options 'B/C/D/E', drops significantly from 25.6% to merely 1.5% when the 'Arabic' region is excluded. Conversely, the removal of the 'Vietnamese' region has a negligible impact on accuracy, which remains at 26.7%. **3)** For the MMLU evaluation, the model's accuracy is minimally impacted by the removal of monolingual regions. Compared to a baseline accuracy of 42.46%, the removals of the 'Arabic' and 'Vietnamese' regions result in accuracies of 39.27% and 39.68%, respectively.

Furthermore, using "*There are 365 days in a year and 12*" as a prompt for generation, we test outputs on the removal of the 'Arabic' and 'Vietnamese' regions. More details are provided in Appendix F.

Languages		Russian (**10K**)		Russian (**100K**)	
	Base	Top	Bottom	Top	Bottom
Arabic	6.771	7.105	6.785	7.071	6.787
Chinese	8.562	8.927	8.593	8.878	8.599
Italian	14.859	16.155	14.931	16.274	14.935
Japanese	10.888	11.212	10.931	11.119	10.951
Korean	4.965	5.19	4.972	5.149	4.974
Persian	6.509	6.93	6.506	6.894	6.515
Portuguese	15.318	16.51	15.247	16.421	15.247
Russian	12.062	**28.93**	12.141	**41.381**	12.137
Spanish	17.079	18.07	17.224	17.894	17.211
Ukrainian	9.409	**18.147**	9.43	**22.622**	9.435
Vietnamese	5.824	6.086	5.872	6.079	5.873

Table 5: LLaMA-2-7B perplexity on 11 languages with a Russian region removal. Here, 'Russian' and 'Ukrainian' are gray-filled while others are unfilled, 'Top' and 'Bottom' are deduplicated, and 'Base' is unchanged. Values with greater changes compared to the other regions' removals are in bold.

3.5 Further Pre-training Optimization

In this section, we demonstrate that stabilizing the core linguistic regions (identified in Section 3.2) during further pre-training mitigates the catastrophic forgetting (CF) issue (McCloskey and Cohen, 1989; Kemker et al., 2018) in LLMs, while

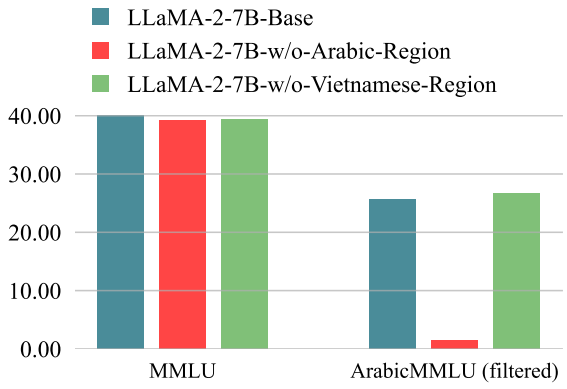

Figure 5: Model's accuracy on MMLU and ArabicMMLU (filtered) test. Here, 'filtered' denotes removing questions whose correct answer is 'A'.

maintaining learning proficiency comparable to full-scale fine-tuning in target language. Our experimental setup involves further pre-training LLaMA-2-7B on 100,000 Arabic sentences, with a batch size of 256, a maximum token length of 512, and learning rates (lr) of $5e{-}5$ or $5e{-}6$, employing perplexity (PPL) as evaluation criterion.

Full-scale Model Fine-tuning Traditional full-scale fine-tuning, when increasing the learning rate or the amount of corpus data, enhances learning in the target language but aggravates forgetting in non-target languages. As shown in Table 2, since LLaMA-2 is primarily trained on English corpora, conducting a second stage of pre-training solely on large-scale Chinese corpora can lead to the forgetting of English competence. Additionally, as depicted on the left side of Figure 6 in blue line, increasing lr from $5e{-}6$ (dotted line) to $5e{-}5$ (solid line) under full-scale fine-tuning boosts the acquisition of the target language (Arabic), while simultaneously accelerates the forgetting rate of the non-target languages (English and Chinese), as shown in the middle and right side.

Freeze Core Regions Fine-tuning We hypothesize that CF problem occurs due to the amplification of parameter adjustments when increasing the learning rate, which leads to significant shifts in the core linguistic region, adversely affecting language alignment. To mitigate this, we protect the core linguistic region and key dimensions by freezing the 'Top 5%' core language area for fine-tuning, as shown by the red line in Figure 6. At a lr of $5e{-}6$ (dotted line), the difference between freezing fine-tuning and full-scale fine-tuning is minimal. However, when the lr increases to $5e{-}5$ (solid line), freezing fine-tuning not only similarly facilitates faster learning in the target language (preserving comparable performance in Arabic PPL: 3.557 vs. 3.566), but also significantly reduces the forgetting of non-target languages (showing improvements in English and Chinese PPL: 18.796 vs. 20.557 and 90.84 vs. 563.423, respectively).

The potential reason for this phenomenon may lie in the preservation of the core regions within the cross-lingual alignment competence. Restricting the magnitude of updates in the core region's parameters is a future strategy that we intend to employ. Notably, unlike regularization methods (Srivastava et al., 2014; Goodfellow et al., 2014), such approaches restricts to a minimal core region in LLMs, and can be implemented alongside blending previous data, retraining the entire network, or possibly only the final layers, without adding additional components to the model.

4 Related Work

Intrinsic Regions Prior works aimed to extract a sub-network capable of executing specific downstream tasks (Frankle and Carbin, 2019; Zhang et al., 2022) or task-specific subspaces to limit parameter fine-tuning within it (Aghajanyan et al., 2021; Zhang et al., 2023). For parameter importance estimation, an effective metric is to use parameter magnitude (Zhu and Gupta, 2018; Renda et al., 2020; Zafrir et al., 2021). Another metric involves estimating the sensitivity of parameters (Molchanov et al., 2019; Sanh et al., 2020; Liang et al., 2021; Sapkota and Bhattarai, 2023). In this work, we employ the latter method to select the most crucial parameters to unveil linguistic regions. Unlike (Frankle and Carbin, 2019), which extracted a *complete* sub-network for downstream tasks, as exemplified by the optimal scale of the lottery network at 21.2%, our findings indicate that the linguistic region is a much smaller (1%) *functional* region.

Cross-lingual Transfer Multilingual language models exhibit significant zero-shot and few-shot cross-lingual transferability across diverse tasks (Pires et al., 2019; Xu et al., 2023). Fine-tuned on one language enables model to obtain comparable capabilities in another language (Muennighoff et al., 2023; Ye et al., 2023), often displaying code-switching behavior in context generation (Khanuja

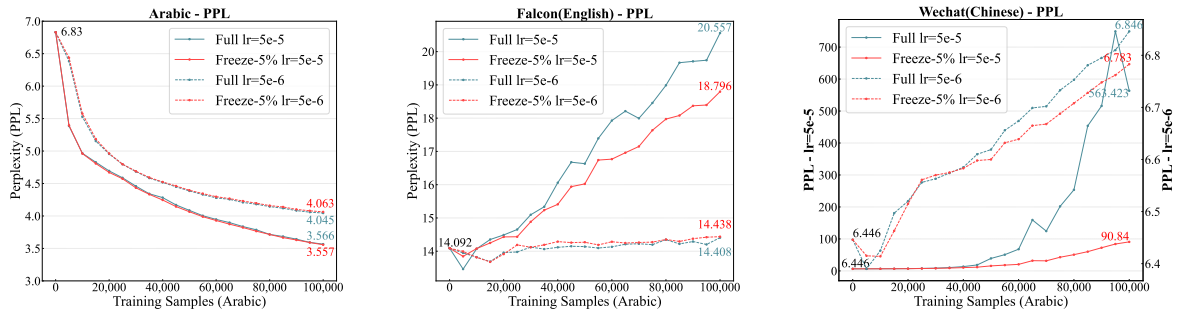

Figure 6: Perplexity of LLaMA-2 across Arabic, English, and Chinese when training on 100,000 Arabic sentences. Blue represents full-scale fine-tuning, and red denotes fine-tuning with the 'Top 5%' of the model parameters frozen. Dashed lines indicate a learning rate (lr) of $5e{-}6$, and solid lines represent lr of $5e{-}5$. We find fine-tuning with the 'Top 5%' region frozen during further pre-training effectively mitigates forgetting of non-target languages while maintaining target language acquisition.

et al., 2020; Zhao et al., 2024). While enhancements in cross-lingual generalization through parameter and information transfer learning (Üstün et al., 2020; Choenni et al., 2023), compulsory language alignment (Sherborne and Lapata, 2022; Shaham et al., 2024) and in-context learning techniques (Winata et al., 2021; Tanwar et al., 2023) have been effective, a comprehensive understanding of the internal mechanisms enabling cross-linguistic alignment in LLMs is still lacking.

Linguistic Abilities Probing Prior works have shown that multilingual LMs rely on a shared subword vocabulary and joint pre-training across multiple languages (Wu and Dredze, 2019; Pires et al., 2019; Cahyawijaya et al., 2023). However, new insights highlight these models' capacity for learning universal semantic abstractions (Artetxe et al., 2020; Chi et al., 2020) and demonstrate that embeddings of similar words in similar sentences across languages are approximately aligned already (Cao et al., 2020; Conneau et al., 2020; Xu et al., 2022). Analysis from a hierarchical perspective reveals that classifiers linked to different BERT (Devlin et al., 2019) layers assess semantic features through varied probe tasks (Lin et al., 2019; Jawahar et al., 2019). In this work, we introduce a parameter partitioning perspective within LLMs, identifying core linguistic and monolingual regions, which underpin cross-lingual alignment and language-specific characteristics, respectively.

Outlier Dimensions Multiple studies have identified outlier dimensions in Transformer-based LMs (Kovaleva et al., 2021; Dettmers et al., 2022). Researches have found that certain outlier dimensions in pre-trained LMs are highly sensitive to the fine-tuning of downstream tasks (Kovaleva et al., 2021; Puccetti et al., 2022). Furthermore, (Puccetti et al., 2022; Rudman et al., 2023) discovered that these outlier dimensions encode task-specific knowledge, and disabling these dimensions significantly degrades model performance. (Luo et al., 2021) attributed positional embeddings to the emergence of outliers, while (Rajaee and Pilehvar, 2022) reported inconsistent results. Additionally, (Sun et al., 2024) found that these outlier dimensions exhibit significantly larger activation values than others in LLMs. Our findings demonstrate that beyond the outlier dimensions, other non-outlier dimensions within the core linguistic region also play an indispensable role in the model's core linguistic competence.

5 Conclusion

This paper explores the pivotal role of certain parameters in Large Language Models (LLMs), identifying a core region essential for multilingual alignment and generalization. Removing this region causes a complete loss of linguistic competence in LLMs. Furthermore, we discover that this core region is concentrated in specific dimensions, perturbing only one dimension can cause a significant decrease in linguistic ability. Moreover, beyond the core linguistic regions, we observe that monolingual regions exist within LLMs that affect specific languages. Importantly, we note that the catastrophic forgetting phenomenon during further pre-training may be related to drastic changes in core linguistic regions, as freezing this part during further pre-training alleviates the issue substantially. Our analysis and findings provide new perspectives and explanations for LLMs' linguistic competence.