Unveiling Linguistic Regions in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated considerable cross-lingual alignment and generalization ability. Current research primarily focuses on improving LLMs' cross-lingual generalization capabilities. However, there is still lacks of research on the intrinsic mechanisms of how LLMs achieve cross-lingual alignment. From the perspective of region partitioning, this paper conducts several investigations on the linguistic competence of LLMs. We discover a core region in LLMs that corresponds to linguistic competence, accounting for approximately 1% of the total model parameters. Removing this core region by setting parameters to zero results in a significant performance decrease across 30 different languages. Furthermore, this core region exhibits significant dimensional dependency, perturbations to even a single parameter on specific dimensions leading to a loss of linguistic competence. Moreover, we discover that distinct regions exist for different monolingual families, and disruption to these specific regions substantially reduces the LLMs' proficiency in those corresponding languages. Our research also indicates that freezing the core linguistic region during further pre-training can mitigate the issue of catastrophic forgetting (CF), a common occurrence observed during further pre-training of LLMs. Overall, exploring the LLMs' functional regions provides insights into the foundation of their intelligence.

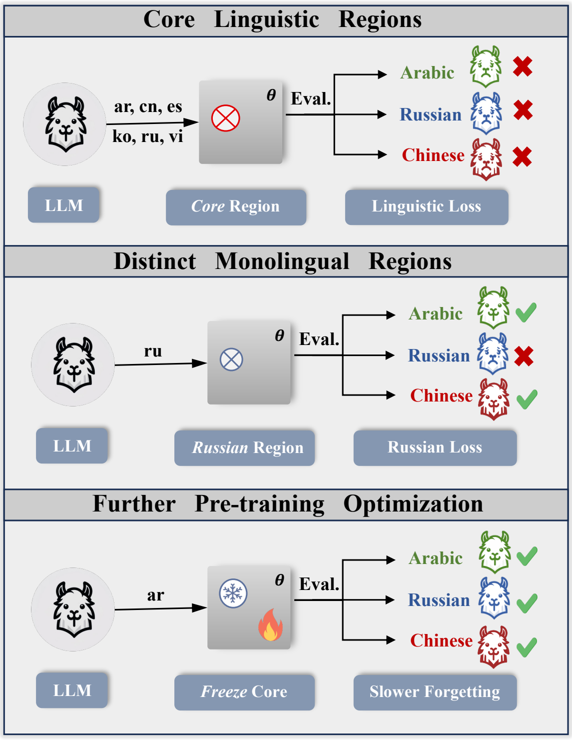

Figure 1: Three main findings of our experiments: (1) Identification of core language regions within the LLMs, where removals lead to linguistic competence loss; (2) Discovery of monolingual regions, where removals cause significant proficiency loss in specific languages; (3) Optimization of freezing core regions during further pre-training decelerates language forgetting.

1 Introduction

Over the years, the field of Natural Language Processing (NLP) has been at the forefront of understanding the core principles of intelligence. The emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), PaLM 2(Anil et al., 2023), and LLaMA 2(Touvron et al., 2023), showcases a significant breakthrough. Thanks to unparalleled scales of model architecture and the vastness of training data, these LLMs now exhibit exceptional linguistic competence and can execute complex tasks requiring abstract knowledge (Dong et al., 2023) and reasoning (Cobbe et al., 2021).

Previous research has revealed that LLMs naturally capture cross-linguistic similarities in their representation space, facilitating zero-shot cross-lingual transfer(Pires et al., 2019; Wu and Dredze, 2019; Xu et al., 2023). The model is fine-tuned on one language, enabling the acquisition of comparable capabilities in another language(Muennighoff et al., 2023; Ye et al., 2023), and exhibits the phenomenon of code-switching when generating context, switching between languages within a single utterance(Khanuja et al., 2020; Zhao et al., 2024). Attempts to improve LLMs' cross-lingual generalization abilities have been successful through parameter and information transfer learning(Üstün et al., 2020; Choenni et al., 2023), aligning languages compulsorily(Sherborne and Lapata, 2022; Shaham et al., 2024) and utilizing in-context learning techniques(Winata et al., 2021; Tanwar et al., 2023). However, a detailed research of underlying the internal mechanisms of how LLMs possess cross-linguistic alignment capability remains elusive.

To delve deeper into the intrinsic mechanisms of LLMs' linguistic competence, this paper focus on the LLMs' parameter importance and investigate the linguistic regions of LLMs based on 30 distinct languages' performance, with the purpose of figuring out the following questions:

Q1: Does a core linguistic region exist within LLMs that facilitates cross-lingual generalization? Through pre-training across six languages and measurement of parameter importance (Section 2.2), we discover a core region in LLMs corresponding to linguistic competence, which accounts for approximately 1% of the model's total parameters. As shown at the top of Figure 1, removing this region (setting parameters to *zero*) consistently lead to a significant decline in performance across 30 test languages (Section 3.2).

Furthermore, by visualizing the core linguistic region (Figure 2), we observe that the linguistic core region of LLMs exhibits significant dimensional dependency. In certain dimensions, only perturbing a single parameter could lead to the model losing its linguistic competence (Section 3.3).

Q2: Beyond the core multilingual region within LLMs , do distinct monolingual (or monolingual family) regions exist that specifically influence individual languages? While LLMs possess strong multilingual capabilities, we discover that each individual language (or language family) encompasses independent regions within the LLMs. As shown in the middle of Figure 1, the analysis of the Russian sentences identifies a particular linguistic region that likewise exerts influence both on the Russian and Ukrainian language, both of which belong to the Slavic group (Section 3.4).

Q3: If and how core linguistic regions affect further pre-training, how to utilize it to optimize further pre-training? After pre-training, core linguistic parameter regions of the LLMs are established for multilingual alignment. Notable shifts in these regions potentially lead to a decline in model lingual capabilities. Our findings reveal that freezing this core region can mitigate the issue of catastrophic forgetting(McCloskey and Cohen, 1989; Kemker et al., 2018), a common occurrence observed during further pre-training of LLMs. As shown at the bottom of Figure 1, we investigate the impact of selectively freezing 5% key parameters of all parameters fine-tuning during further pre-training, compared to the full-scale fine-tuning technique. Findings indicate that this method facilitated comparable or even more efficient learning of the target language, while concurrently decelerating the rate of language attrition for previously learned languages (Section 3.5). Significantly, our methodology is compatible with the data-replay techniques(Robins, 1995), with no necessity for integrating extra components into the model. Unlike regularization methods(Srivastava et al., 2014; Goodfellow et al., 2014), our approach restricts to a minimal core region in LLMs.

The main contributions of our work are summarized as follows:

- We discover that LLMs possess a core linguistic region, and removing this region (setting parameters to *zero*) results in a significant loss of the model's linguistic capabilities. Furthermore, perturbations to specific dimensions or even a single parameter can lead to a substantial decline in the model's linguistic abilities.
- We observe that distinct monolingual regions exist in LLMs for different languages (or language families). Removing a specific monolingual region causes a significant deterioration in the linguistic capabilities within that language (or language family).
- We perform further pre-training for specific languages within the core linguistic region of LLMs frozen, achieving comparable or even superior performance in the target language while mitigating catastrophic forgetting in non-target languages.

2 Background and Metric

2.1 Model Pre-training

Pre-training is a crucial process by which LLMs acquire linguistic competence and gain general knowledge about the real world. Formally, given a large corpus $\mathcal{D}$, the training objective is to find the optimal θ, minimize the following loss $\mathcal{L}$:

$$\mathcal{L}(\mathcal{D}, \theta) = \sum_{x \in \mathcal{D}} \sum_{i} \log p_{\theta}(x_i | x_1, ..., x_{i-1}), \quad (1)$$

where $x = \{x_1, ..., x_n\}$ denotes an input token sequence and θ denotes parameters of the model.

2.2 Parameter Importance

Drawing upon the observations of linguistic alignments, we propose that particular parameters regions within the model exert significant influence on its inherent language alignment capabilities. Evaluating parameter sensitivity is a crucial metric for determining the significance of parameters in model pruning(Sanh et al., 2020; Liang et al., 2021; Zhang et al., 2022). If removing a parameter (zero-out) significantly affects the loss, the model is sensitive to it. More specifically, given a large corpus $\mathcal{D}$ and $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ as the parameters of a model, with each $\theta_j \in \mathbb{R}$ denoting the j-th parameter, the training objective is to minimize loss $\mathcal{L}(\mathcal{D}, \theta)$ (defined in 2.1). The importance of each θ is denoted as $\mathcal{I}(\theta) \in \mathbb{R}^d$, where its j-th index $\mathcal{I}_j(\theta)$ signifies the importance for θ_j.

Under an independent and identically distributed data (i.i.d.) assumption, the importance of a parameter $\mathcal{I}_j(\theta)$ is measured by the increase in prediction loss when it is removed, calculated as the absolute difference between prediction losses with and without the parameter(θ_j):

$$\mathcal{I}_j(\theta) = |\mathcal{L}(\mathcal{D}, \theta) - \mathcal{L}(\mathcal{D}, \theta | \theta_j = 0)| . \quad (2)$$

Calculating $\mathcal{I}_j(\theta)$ for each parameter, as outlined in 2, is computationally expensive because it involves d distinct versions of the network computing, for each removed parameter. This becomes particularly challenging as the number of model parameters, d, grows to hundreds of billions. However, according to the Taylor expansion formula for $\mathcal{L}$ at $\theta_j = 0$:

$$\begin{aligned} &\mathcal{L}(\mathcal{D}, \theta) = \mathcal{L}(\mathcal{D}, \theta | \theta_j = 0) \\ &+ \frac{\partial \mathcal{L}}{\partial \theta_j}(\theta_j - 0) + \frac{1}{2!}\frac{\partial^2 \mathcal{L}}{\partial \theta_j^2}(\theta_j - 0)^2 + \cdots , \end{aligned} \quad (3)$$

we can estimate $I_j(\theta)$ with its first-order Taylor expansion, eliminating the requirement for d distinct networks computation:

$$\mathcal{I}_j(\theta) \approx |g_j \theta_j| , \quad (4)$$

where $g_j = \frac{\partial \mathcal{L}}{\partial \theta_j}$ are elements of the parameter gradient g, and the importance is easily calculated since the gradient g can be obtained from back-propagation.

3 Experiments

3.1 Experimental Setup

To localize the functional regions corresponding to linguistic competence within LLMs and analyze their nature, we perform language further pre-training (next token prediction) on various languages and observe the relationship between internal parameter removal and external output quality. We utilize LLaMA-2-7B/13B (Touvron et al., 2023) as our model instance, as it stands out as one of the most notable state-of-the-art open-source LLMs in current academia.

Our experimental dataset comprises materials from Chinese platforms like Zhihu and Wechat, English sources from Arxiv and Falcon, and a corpus including books from 28 languages, totaling 30 languages in all. Six languages, namely Arabic, Spanish, Russian, Chinese, Korean, and Vietnamese, are chosen for language further pre-training and region localization, with $100,000$ samples for each (distinct from the samples in the test set). All 30 languages are employed for model testing and functional region analysis, with the specific languages and token count detailed in A. We use perplexity (PPL) as the criterion for evaluating the linguistic competence of a language model.

3.2 Core Linguistic Competence Region

In this section, we conduct further pre-training experiments on LLaMA-2 across six languages, aiming to explore and identify core parameter regions associated with linguistic competence. Specifically, according to Equation 4, we cumulatively compute $\mathcal{I}^*(\theta) = \Sigma \mathcal{I}(\theta)$ values across six different languages' training, positing that the set of parameters exhibiting maximal importance score $\mathcal{I}^*(\theta)$ during the language further pre-training may have a strong correlation with the model's linguistic competence, and we provide both logical and empirical evidence to support this hypothesis.

Logical Evidence The phenomenon of code-switching suggests that the LLMs can align languages and may possess core linguistic regions. As discussed in Section 2.2, if a parameter θ_j is crucial for the LLMs' core linguistic competence, the model should be sensitive to θ_j, shown by a

Languages	Base	LLaMA-2 **3%** Removal Top	Bottom	Random
Arabic	6.771	127208.250	6.772	7.895
	6.261	102254.758	6.316	7.112
Chinese	8.652	295355.5	8.565	9.837
	7.838	84086.906	7.806	8.619
Italian	14.859	58908.879	14.860	17.341
	13.694	47375.844	13.730	15.207
Japanese	10.888	322031.406	10.896	12.535
	10.072	75236.031	10.137	11.661
Korean	4.965	125345.359	4.967	5.649
	4.724	90768.844	4.743	5.241
Persian	6.509	81959.719	6.511	7.628
	6.205	92201.812	6.229	7.009
Portuguese	15.318	47763.059	15.319	17.297
	13.667	51498.402	13.982	15.376
Russian	12.062	170776.750	12.064	13.728
	11.048	112574.609	10.948	11.757
Spanish	17.079	51940.859	17.082	18.98
	16.351	54005.891	16.138	17.292
Ukrainian	9.409	120719.938	9.409	10.875
	8.295	116287.305	8.297	9.076
Vietnamese	5.824	40126.527	5.824	6.614
	5.471	42336.426	5.437	5.995

Table 1: LLaMA-2 perplexity on 11 languages with 3% removal ratio. The 13B model is gray-filled while the 7B model is unfilled. 'Top' and 'Bottom' respectively indicate the N parameters with the highest and lowest cumulative $\mathcal{I}_j^*(\theta)$ during the further pre-training across the six languages. 'Random' denotes the randomly selecting N while 'Base' represents no removal. Here, N equals 3% of the total number in each parameter matrix.

significant increase on the loss $\mathcal{L}$ when θ_j is removed, severely impairing the LLMs' linguistic performance. Conversely, other parameters impact rarely on core linguistic capabilities.

Empirical Evidence 1 Table 1 illustrates that even a 3% removal on the 'Top' region leads to a substantial increase in perplexity (PPL), reaching over $40,000$ across 11 languages, indicating a complete loss of linguistic competence. In contrast, removing the 'Bottom' region is comparable to non-removal 'Base' in model PPL, and a 'Random' removal of equal magnitude has no significant impact on the model's linguistic competence. Moreover, refer to Appendix B, additional experiments with reducing training samples to $10,000$ or adjusted the region selection ratio to 1% and 5% yield consistent findings: removing the 'Top' region deprives LLaMA-2 of its capability across all 30 languages. This suggests the model's linguistic competence is directly influenced by the 'Top' region, while removing the 'Bottom' and 'Random' region don't have a significant direct impact on language capabilities. See Appendix B for evaluations on 30 languages and further experiments.

Testing Dataset (Language)	# Training Samples (Chinese)	Removal Ratio = 1% Top & Freeze	Bottom & Freeze	Top & Unfreeze
Wechat (Chinese)	0K	254772480	6.452	254772480
	2K	674.076	6.052	6.05
	5K	292.499	6.053	6.058
	10K	116.859	6.305	6.303
	20K	20.722	6.556	6.559
	50K	9.129	6.18	6.175
	200K	6.246	5.581	5.604
Falcon (English)	0K	4244070	14.02	4244070
	2K	158431.282	14.507	14.445
	5K	343498	15.732	15.415
	10K	175567.219	15.878	15.875
	20K	32505.828	18.689	18.952
	50K	12455.038	29.029	31.583
	200K	5301.527	488.429	448.804

Table 2: Removing-freezing analysis at 1% removal ratio in different regions of LLaMA-2-7B. 'Top/Bottom' denotes the removal region, while 'Freeze/Unfreeze' indicates whether the corresponding region is frozen after removal.

Empirical Evidence 2 In the experiment corresponding to Table 2, we initially zero out various regions within LLaMA. Consistent with the findings from Table 1, removing the 'Top' region leads to a loss of linguistic competence, whereas the 'Bottom' region don't. However, in this experiment, we sought to ascertain if LLaMA could reacquire its lost cross-lingual generalization competence. Thus, we train on different amounts of Chinese Zhihu corpus and evaluate on Chinese Wechat and English Falcon corpora. The results indicate that unlike the 'Bottom' region, if the 'Top' region is removed and frozen, the model have to relearn basic language rules in other regions based on the provided Chinese Zhihu corpus, but these rules are inherently biased towards Chinese. Consequently, while its proficiency in Chinese is restored, the English perplexity remains high (5301.527). If the 'Top' region is removed but not frozen, the model can rebuild its linguistic competence in-place. As its proficiency in Chinese is restored, so is its proficiency in English. This implies that the 'Top' region encodes generalizable fundamental linguistic competence. When 'Top' region is zeroed-out and frozen, other regions significantly adapt to regain core linguistic competence. Similar conclusions can be obtained with an expanded removal ratio 5%. For further details, see Appendix C.

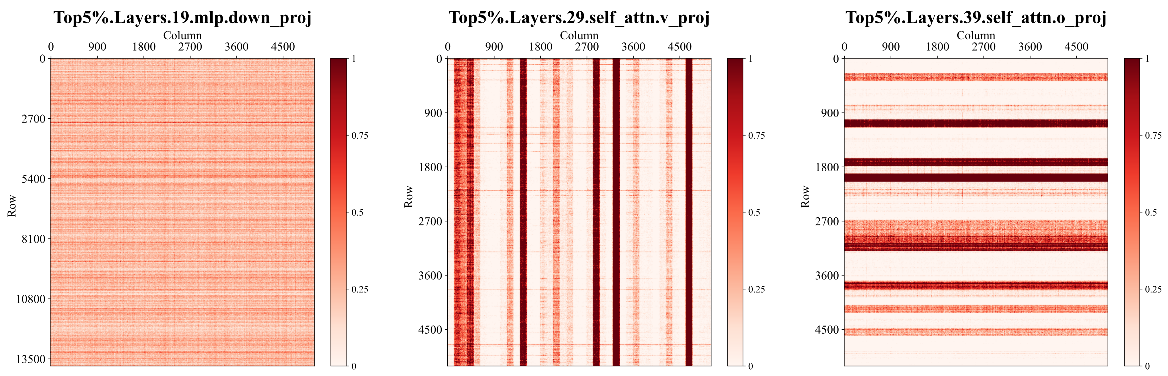

Figure 2: Visualization of the linguistic competence region (the 'Top' 5% region). The scale from 0 to 1 (after normalization) represent the proportion of parameters within a 3 × 3 vicinity that belong to the Top region.

Model Size	# Training Samples	N_d	Attn.o(row), Attn.k/q/v+FFN.down(column)			
			Top	Middle	Bottom	Random
7B	100K	1	848.326	6.447	6.447	6.48
	100K	3	72594.445	6.455	6.458	6.487
	100K	5	48001.992	6.461	6.463	6.495
	100K	10	62759.516	6.478	6.48	6.529
13B	100K	1	5218.1	5.857	5.857	5.856
	100K	3	37344.078	5.863	5.858	5.985
	100K	5	41840.613	5.867	5.86	5.89
	100K	10	465740.125	5.879	5.869	6.992
13B	10K	1	23120.977	5.859	5.856	5.865
	10K	3	28816.867	5.862	5.86	5.875
	10K	5	73268.289	5.866	5.862	5.878
	10K	10	592922.25	5.879	5.871	5.993

Table 3: Perplexity of LLaMA-2 after removing certain dimensions in the Attention and Feedforward layers. Here, N_d denotes the number of dimensions to remove, 'Top', 'Middle', and 'Bottom' refer to the dimensions with the most, moderate, and least cumulated $\mathcal{I}_\theta$ during further pre-training. 'Random' denotes an equivalent number of dimensions chosen randomly for comparison.

3.3 Dimensional Dependence

To provide a more intuitive revelation of the spatial distribution characteristics of the linguistic competence region within the model, we visualize the 'Top' region. As shown in Figure 2, whether in the attention mechanism layer or the feed-forward layer, the linguistic region displays a distinct concentration in both the rows and columns of the matrices. Such distribution features seem to imply that the model's linguistic competence is concentrated in specific dimensions.

Structured Removal Instead of discretely removing different unstructured parameters, we selectively remove structured certain rows or columns for each matrix, especially those dimensions encompassing a significant number of 'Top' region parameters, termed as 'Top' dimensions. As illustrated in Table 3, we attempt to remove the columns of FFN.down and Attn.k/q/v, as well as the rows of Attn.o. The results indicate that removing just these 'Top' dimensions leads to a substantial decline in the model's linguistic competence. However, disturbances to the 'Middle' ,'Bottom' and 'Random' dimensions do not yield noticeable effects. Selecting the dimensional region only from the Attention matrix or inverting rows and columns removals lead to similar findings, as described in D.

Single Dimension Perturbation Here, we explore whether a specific dimension significantly impact the model's linguistic competence. As illustrated in Figure 3, we iterate through the key dimensions mentioned in Section 3.3, attempting to perturb the same dimension (random initialization) across all Transformer layers. The results indicate that the impact of dimensions 2100 and 4743 on the LLaMA-2-13B substantially surpassed other dimensions, even when compared to the other three in the 'Top5' dimensions. In contrast, perturbing two randomly selected dimensions, 2800 and 4200, yield linguistic performance almost indistinguishable from the unperturbed state.

Single Parameter Perturbation We discover that even a slight modification to a single parameter in a model with over 13 billion parameters can lead to a significant decline in its output quality. In Table 4, merely resetting the 2100-th parameter in the 'Input_LayerNorm' module of the 1-st layer to its initial value causes LLaMA-2-13B's PPL value to skyrocket from 5.865 to 83224.078. If this weight parameter is multiplied by 10, the PPL value also rises to 4363.462. This suggests that even minor changes to a single parameter can cause the model

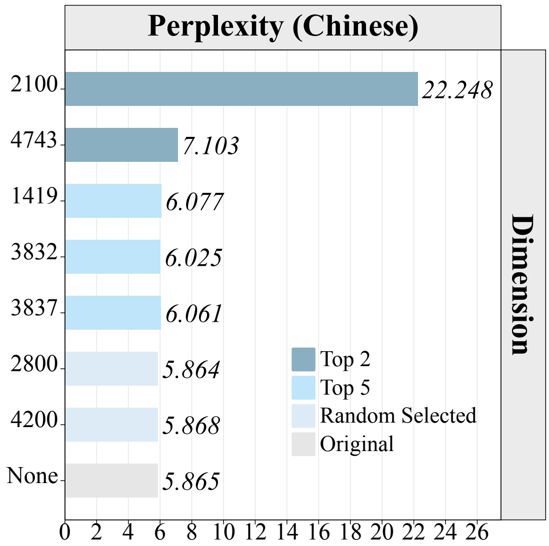

Figure 3: Perplexity of the LLaMA-2-13B when perturbing the same single dimension (Att.O and FFN.down matrices) across all layers. 'Topk' represents the top k dimensions that disrupt the model the most. 'Random selected' refers to a randomly chosen dimension. 'Original' indicates that no dimensions are disrupted.

to lose nearly all of its linguistic competence. However, randomly altering the parameters at dimensions 2800 and 4200 doesn't noticeably impact the model. For more details, refer to Appendix E.

Output Under Perturbations To visually illustrate the impact of the linguistic competence region on the model's output quality, we use "Fudan University is located in" as a premise and observe the model's outputs under different parameter perturbations. The results are shown in Figure 4. Compared to randomly selected dimension 4200-th, perturbing model on 2100-th dimension significantly leads to model loses its linguistic competence, producing error or even nonsensical strings.

3.4 Monolingual Family Region

In this section, we wonder if LLMs possess distinct regions within an individual language (or language family). Unlike the core linguistic regions described in Section 3.2, a monolingual family region only has a strong correlation with certain languages, and removing it will only cause significant influence on LLMs' proficiency in those corresponding languages.

Region Localization Different from Section 3.2, we initially identify and select the 1% 'Top' and 'Bottom' regions for each of the six languages (Arabic, 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 abilities. 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:

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

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 $10,000$ (10K) or $100,000$ (100K) samples respectively, as shown in Table 5, only Russian itself and Ukrainian have significant increases in PPL when removing 'Top' 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.

Perturbation	Parameter	Perplexity
-	-	5.865
Reset 1	L1-N2100	83224.078
Reset 1	L1-N2800	5.860
Reset 1	L1-N4200	5.858
Mul 10	L1-N2100	4363.462
Mul 10	L1-N2800	5.859
Mul 10	L1-N4200	5.864

Table 4: Perplexity of LLaMA-2-13B on Chinese when perturbing a single weight parameter. Here, 'Reset 1' represents resetting the parameter to 1 (the initial value before pre-training), 'Mul 10' represents multiplying the parameter by 10. 'L1' represents 1-st layers. 'N' represents the 'Input_LayerNorm' module, followed by the perturbed dimension.

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.

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, ‘Arabic’ and ‘Persian’ 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 mitigates the catastrophic forgetting (CF) issue(McCloskey and Cohen, 1989; Kemker et al., 2018) in LLMs, while preserving performance comparable to full-scale fine-tuning in target language proficiency. Our experimental setup involves further pre-training LLaMA-2-7B on 100,000 Arabic sentences, with a batch size of 256, a maximum token count of 512, and learning rates (lr) and $5e{-}5$ or $5e{-}6$, employing perplexity (PPL) as the evaluation criterion.

Full-scale Model Fine-tuning Traditional full-scale fine-tuning, when increasing the learning rate, enhances learning in the target language but aggravates forgetting in non-target languages. To counteract this forgetting, it is often essential to incorporate a portion of data from these other languages. As depicted on the left side of Figure 5 in line Blue, increasing lr from $5e{-}6$(dotted blue line) to $5e{-}5$ (solid blue line) under full-scale fine-tuning boosts the acquisition of the target language (Arabic), while accelerates the forgetting rate of the non-target languages (English and Chinese) simultaneously, 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 5. 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 facilitates faster learning in the target language (achieving better 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 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 Realated Work

Neuron Importance Estimating Several works have empirically estimated neuron importance. An effective importance metric is to utilize parameter magnitude(Zhu and Gupta, 2018; Renda et al., 2020; Zafrir et al., 2021). However, such a simple approach may inadequately measure a weight’s impact on model output. Another importance metric is to estimate sensitivity of parameters, essentially

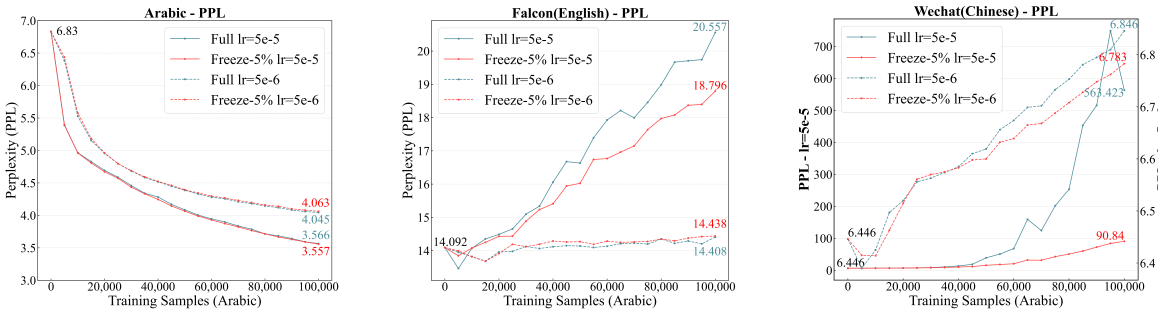

Figure 5: Perplexity of LLaMA-2 across Arabic, English, and Chinese when training on 100,000 Arabic sentences. Blue represents full-scale fine-tuning, and red denotes fine-tuning with the 'Top 5%' of the model parameters frozen. Dashed lines indicate a learning rate (lr) of 5e-6, and solid lines represent 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.

approximates the change in loss when a parameter is zeroed out(Molchanov et al., 2019; Sanh et al., 2020; Liang et al., 2021; Sapkota and Bhattarai, 2023). These studies all focus on pruning unimportant parameters through parameter importance estimation. In this work, we select the most crucial parameters to unveil the core linguistic and monolingual regions.

Cross-lingual Transfer Multilingual language models exhibit significant zero-shot and few-shot cross-lingual transferability across diverse tasks(Pires et al., 2019; Wu and Dredze, 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 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, comprehensive understanding of the internal mechanisms enabling cross-linguistic alignment in large language models (LLMs) is still lacking.

Linguistic Abilities Probing Researchers have investigated the mechanisms underlying strong cross-lingual performance. Prior works have shown that multilingual multilingual language models rely on a shared subword vocabulary and joint pre-training across multiple languages(Pires et al., 2019; Cahyawijaya et al., 2023; Wu and Dredze, 2019). However, new insights highlight these models' capacity for learning universal semantic abstractions(Artetxe et al., 2020; Chi et al., 2020) and demonstrate that mBERT(Devlin et al., 2019) 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.

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 ability in LLMs. Further more, we discover that this core region is concentrated in specific dimensions, perturbing only one dimension can cause a significant decrease in language 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.

Limitations

In this paper, while we discover the core linguistic region and distinct monolingual regions within Large Language Models (LLMs), our work presents two notable limitations. First, our experiments are based on LLaMA-2-7B/13B, and it remains to be further determined whether the same phenomenon are observable in larger or differently architected models. Despite this, our focus on LLaMA-2-7B/13B reveals the existence of linguistic regions within the model, providing an explanation for the model's linguistic capabilities. Secondly, we optimize full-scale fine-tuning through the freezing operation, which is not suited to extensive datasets. A more feasible approach is to limit the magnitude of parameter updates, which is the direction of our future experiments. Nevertheless, it is important to emphasize that slowing down forgetting through freezing core region suggests that in further pre-training, the core region is different from the other regions. Range of variation amplitude in core region should be smaller to maintain the cross-lingual generalization capabilities of the model. Additionally, while our study focuses on linguistic regions, beyond language, knowledge is a higher-level semantic representation, which is a critical direction for us to explore in the future.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4623–4637. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Halim Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Akbarianto Wibowo, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. Nusacrowd: Open source initiative for indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13745–13818. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5564–5577. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Comput. Linguistics*, 49(3):613–641.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6022–6034. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA,* June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3651–3657. Association for Computational Linguistics.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3390–3398. AAAI Press.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3575–3585. Association for Computational Linguistics.

Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2021. Super tickets in pre-trained language models: From model compression to improving generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6524–6538. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 241–253. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. *Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem*, page 109–165. Elsevier.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. 2019. Importance estimation for neural network pruning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11264–11272. Computer Vision Foundation / IEEE.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001. Association for Computational Linguistics.

Alex Renda, Jonathan Frankle, and Michael Carbin. 2020. Comparing rewinding and fine-tuning in neural network pruning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Anthony V. Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.*, 7(2):123–146.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Suman Sapkota and Binod Bhattarai. 2023. Importance estimation with random gradient for neural network pruning. *CoRR*, abs/2310.20203.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *CoRR*, abs/2401.01854.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4134–4153. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6292–6307. Association for Computational Linguistics.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2302–2315. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *CoRR*, abs/2109.07684.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 833–844. Association for Computational Linguistics.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-linguistic syntactic difference in multilingual BERT: how good is it and how does it affect transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8073–8092. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2023. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13951–13976. Association for Computational Linguistics.

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

Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, and Moshe Wasserblat. 2021. Prune once for all: Sparse pre-trained language models. *CoRR*, abs/2111.05754.

Qingru Zhang, Simiao Zuo, Chen Liang, Alexander Bukharin, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2022. PLATON: pruning large transformer models with upper confidence bound of weight importance. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26809–26823. PMLR.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *CoRR*, abs/2401.01055.

Michael Zhu and Suyog Gupta. 2018. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

A Languages in Evaluation Corpus

We use evaluation data composed of 30 languages to assess the model's linguistic competence. The 30 languages and their respective token counts (use LLaMA-2 Tokenizer) are as follows: Arabic (4702998), Chinese (2869208), Czech (1362041), Danish (36467), Dutch (3991305), English (1216599), Finnish (372303), French (6755281), German (2884921), Greek (474622), Hungarian (1229433), Indonesian (19226), Italian (6332560), Japanese (501899), Korean (2730794), Malay (5842), Malayalam (1489244), Norwegian (42289), Persian (1736589), Polish (4948702), Portuguese (7598161), Romanian (1381598), Russian (5205716), Spanish (7163860), Swahili

(630), Swedish (1450236), Tamil (2920808), Turkish (2484186), Ukrainian (455720), Vietnamese (3606202).

B Evaluation on 30 Languages

The regions are localized from six languages: Arabic, Spanish, Russian, Chinese, Korean, and Vietnamese.

Removal 3% ratio (100K) LLaMA-2 perplexity on 30 languages when the removal ratio is 3% ratio, with $100,000$ samples for each language. Refer to Table 14 for more details.

Removal 3% ratio (10K) LLaMA-2 perplexity on 30 languages when the removal ratio is 3% ratio, with reduced $10,000$ samples for each language. Refer to Table 15 for more details.

Removal 1% and 5% ratio (100K) LLaMA-2-7B perplexity on 30 languages when the removal ratio is changed to 1% and 5% ratio, with $100,000$ equivalent samples for each language. Refer to Table 16 for more details.

C Remove and Freeze Fine-tuning

In Table 6, we expand removing-freezing ratio to 5% for LLaMA-2-7B fine-tuning, and also observe that removing the 'Top 5%' regions destroys the original cross-lingual generalization competence.

Testing Dataset (Language)	# Training Samples (Chinese)	Removal Ratio = 5%		
		Top & Freeze	Bottom & Freeze	Top & Unfreeze
Wechat (Chinese)	0K	204630.063	6.454	204630.063
	2K	816.94	6.057	6.05
	5K	302.806	6.039	6.054
	10K	167.891	6.285	6.315
	20K	57.355	6.525	6.56
	50K	18.92	6.154	6.173
	200K	8.8	5.582	5.584
Falcon (English)	0K	66877.711	14.027	66877.711
	2K	248218.656	14.486	14.44
	5K	476895.25	14.831	15.074
	10K	467330.656	16.248	16.334
	20K	131648.75	18.579	18.743
	50K	25494.166	29.576	29.377
	200K	6948.663	321.353	504.645

Table 6: Removing-freezing analysis at 5% removal ratio in different regions of LLaMA-2-7B. 'Top/Bottom' denotes the removal region, while 'Freeze/Unfreeze' indicates whether the corresponding region is frozen after removal.

D Attn. Dimensional Removal Evaluation

Figure 6 (left) illustrates that the columns of the Attn.k/q/v matrices in the attention layer, as well as the rows of the Attn.o matrix, correspond to different attention head parameters. Conversely, the rows of the Attn.k/q/v matrices and the columns of the Attn.o matrix are closely associated with features in the representation space.

We remove the 'Top' dimensions in the attention layer, and the results is displayed in Tables 7 and 8. Table 7 reveals that removing the 'Top' dimensions continues to produce more detrimental effects than other dimensions. The visualizations in Figure 2 show that these dimensions are largely concentrated in a few attention heads, suggesting that some attention heads contribute more significantly to the model's linguistic competence. Table 8 indicates that the removals under the second setting cause more damage than the first. Considering that, in the second setting, the 'Top' dimensions in the matrix directly interact with the corresponding features in the representational space, we can conjecture that these features are tightly linked with the model's linguistic competence.

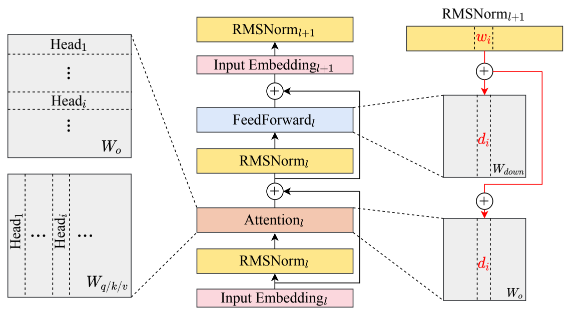

Figure 6: One can see from the left that each row of the Attn.o (W_o) corresponds to a particular attention head, and each column of the Attn.q/k/v ($W_{q/k/v}$) matrix corresponds to one as well. On the right, one can observe the perturbation applied to one weight within RMSNorm, which can be seen as affecting a column of the FFN.down and the Attn.o.

E Single Parameter Perturbation

In a Transformer block, each column in the Attn.o and the MLP.down matrix of the feed-forward layer can be considered as the input weights of a neuron. Thus, perturbing a column can be seen as disturbing the input weights of a neuron. Viewed from another angle, if we disturb the output activation value of this neuron, a similar effect should be observed.

Model Size	# Training Samples	N_d	Attn.o(row), Attn.k/q/v(column)			
			Top	Middle	Bottom	Random
7B	100K	1	9.731	6.448	6.445	6.471
	100K	3	25.82	6.449	6.445	6.474
	100K	5	62.794	6.452	6.446	6.482
	100K	10	875.016	6.456	6.446	6.504
13B	100K	1	10.899	5.857	5.856	5.856
	100K	3	44.384	5.858	5.855	5.98
	100K	5	33.52	5.861	5.856	5.884
	100K	10	118.968	5.863	5.857	5.966
13B	10K	1	8.094	5.856	5.855	5.864
	10K	3	21.561	5.857	5.855	5.866
	10K	5	111.766	5.858	5.856	5.865
	10K	10	108.133	5.861	5.857	5.977

Table 7: Perplexity of LLaMA-2 after removing certain dimensions (zeroed-out) in the attention (Attn) layers. Here, N_d denotes the number of dimensions to remove, 'Top', 'Middle', and 'Bottom' refer to the dimensions with the most, moderate, and least cumulated $\mathcal{I}_\theta$ during further pre-training across six languages, respectively. 'Random' denotes an equivalent number of dimensions chosen at random for comparison.

Model Size	# Training Samples	N_d	Attn.o(column), Attn.k/q/v(row)			
			Top	Middle	Bottom	Random
7B	100K	1	167.804	6.446	6.446	6.446
	100K	3	68554.102	6.446	6.447	6.448
	100K	5	4259.861	6.449	6.447	6.449
	100K	10	68170.25	6.454	6.452	6.449
13B	100K	1	17.609	5.855	5.856	5.856
	100K	3	313.178	5.857	5.856	5.863
	100K	5	526.464	5.858	5.856	5.857
	100K	10	5841.446	5.859	5.858	5.852
13B	10K	1	17.03	5.855	5.856	5.857
	10K	3	206.225	5.856	5.856	5.858
	10K	5	1110.781	5.857	5.856	5.86
	10K	10	9600.097	5.859	5.858	5.874

Table 8: Perplexity of LLaMA-2 after removing certain dimensions in attention (Attn) layers. Different from Table 7, in this table, the columns of the Attn.O and the rows of the Attn.K/Q/V are removed.

Within LLaMA, there is a specific module called RMSNorm, where each dimension is associated with a weight. Perturbations to these weights can be regarded as disturbances to the output activation values of the corresponding neurons (In Figure 6 (right), we visually demonstrate how RMSNorm affects a column of the Attn.o and the FFN.down matrix).

F Individual Language Family Region

Tables 9-13 demonstrate LLaMA-2-7B perplexity after removing Arabic, Spanish, Chinese, Korean, and Vietnamese regions, respectively. The region is obtained by removing the intersections with other languages' respective regions from the 1% 'Top/Bottom' regions, selected from 10,000 or 100,000 sentences during further pre-training according to Equation 4.

Languages		Arabic **(10K)**		Arabic **(100K)**	
	Base	Top	Bottom	Top	Bottom
Arabic	6.771	**81.659**	6.785	**135.02**	6.786
Chinese	8.562	9.309	8.593	9.165	8.588
Italian	14.859	16.61	14.959	16.366	14.919
Japanese	10.888	12.238	10.932	11.956	10.923
Korean	4.965	5.534	4.972	5.442	4.969
Persian	6.509	**34.142**	6.52	**43.414**	6.508
Portuguese	15.318	16.909	15.262	16.86	15.239
Russian	12.062	13.708	12.145	13.781	12.141
Spanish	17.079	18.543	17.24	18.314	17.2
Ukrainian	9.409	11.243	9.433	11.225	9.439
Vietnamese	5.824	6.412	5.874	6.335	5.871

Table 9: LLaMA-2-7B perplexity on 11 languages with an Arabic region removal. Here, 'Arabic' and 'Persian' 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.

Languages		Spanish **(10K)**		Spanish **(100K)**	
	Base	Top	Bottom	Top	Bottom
Arabic	6.771	7.158	6.788	7.15	6.789
Chinese	8.562	8.984	8.594	8.971	8.596
Italian	14.859	**21.292**	14.933	**27.004**	14.95
Japanese	10.888	11.376	10.913	11.426	10.933
Korean	4.965	5.169	4.967	5.167	4.972
Persian	6.509	6.906	6.484	6.945	6.529
Portuguese	15.318	**21.217**	15.249	**26.877**	15.256
Russian	12.062	13.039	12.133	13.252	12.141
Spanish	17.079	**38.876**	17.224	**64.513**	17.225
Ukrainian	9.409	10.027	9.439	10.082	9.439
Vietnamese	5.824	6.136	5.875	6.145	5.877

Table 10: LLaMA-2-7B perplexity on 11 languages with a Spanish region removal. Here, 'Spanish', 'Italian' and 'Portuguese' are gray-filled while others are unfilled, and values with greater changes compared to the other regions' removals are in bold.

Languages		Chinese **(10K)**		Chinese **(100K)**	
	Base	Top	Bottom	Top	Bottom
Arabic	6.771	7.161	6.79	7.714	6.784
Chinese	8.562	**10.899**	8.592	**12.079**	8.586
Italian	14.859	16.041	14.939	15.881	14.932
Japanese	10.888	**12.265**	10.922	**12.878**	10.904
Korean	4.965	5.343	4.974	5.341	4.960
Persian	6.509	6.92	6.519	6.865	6.516
Portuguese	15.318	16.285	15.27	16.241	15.26
Russian	12.062	12.887	12.136	12.973	12.145
Spanish	17.079	18.068	17.216	17.974	17.219
Ukrainian	9.409	10.144	9.439	10.207	9.447
Vietnamese	5.824	6.261	5.878	6.296	5.870

Table 11: LLaMA-2-7B perplexity on 11 languages with a Chinese region removal. Here, 'Chinese' and 'Japanese' are gray-filled while others are unfilled, and values with greater changes compared to the other regions' removals are in bold.

Languages		Korean **(10K)**		Korean **(100K)**	
	Base	Top	Bottom	Top	Bottom
Arabic	6.771	7.259	6.791	7.316	6.783
Chinese	8.562	9.14	8.594	9.173	8.594
Italian	14.859	15.91	14.941	15.791	14.938
Japanese	10.888	**13.273**	10.919	**15.062**	10.932
Korean	4.965	**8.364**	4.971	**13.128**	4.971
Persian	6.509	7.38	6.522	7.574	6.522
Portuguese	15.318	16.113	15.259	15.984	15.26
Russian	12.062	12.758	12.138	12.827	12.136
Spanish	17.079	17.981	17.214	17.858	17.225
Ukrainian	9.409	10.065	9.434	10.108	9.442
Vietnamese	5.824	6.188	5.874	6.177	5.874

Table 12: LLaMA-2-7B perplexity on 11 languages with a Korean region removal. Here, 'Korean' and 'Japanese' are gray-filled while others are unfilled, and values with greater changes compared to the other regions' removals are in bold.

Languages		Vietnamese **(10K)**		Vietnamese **(100K)**	
	Base	Top	Bottom	Top	Bottom
Arabic	6.771	7.435	6.785	7.341	6.789
Chinese	8.562	9.576	8.589	9.372	8.592
Italian	14.859	16.979	14.952	16.497	14.937
Japanese	10.888	12.027	10.946	11.814	10.941
Korean	4.965	5.44	4.97	5.335	4.979
Persian	6.509	7.315	6.501	7.243	6.521
Portuguese	15.318	17.159	15.249	16.805	15.258
Russian	12.062	13.107	12.141	13.007	12.144
Spanish	17.079	18.801	17.244	18.369	17.233
Ukrainian	9.409	10.316	9.447	10.217	9.433
Vietnamese	5.824	**24.382**	5.872	**27.817**	5.874

Table 13: LLaMA-2-7B perplexity on 11 languages with a Vietnamese region removal. Here, 'Vietnamese' is gray-filled while others are unfilled, and values with greater changes compared to the other regions' removals are in bold.

Languages	LLaMA-2-7B **3% (100K)**				LLaMA-2-13B **3% (100K)**			
	Base	Top	Bottom	Random	Base	Top	Bottom	Random
Arabic	6.771	127208.250	6.772	7.895	6.261	102254.758	6.316	7.112
Chinese	8.652	295355.5	8.565	9.837	7.838	84086.906	7.806	8.619
Czech	19.834	62692.367	19.835	24.005	17.744	56102.227	17.650	20.485
Danish	8.372	47654.156	8.372	9.929	7.402	47213.586	7.401	8.278
Dutch	16.959	48478.594	16.959	20.121	15.64	46303.559	15.572	18.295
English	7.653	16573.422	7.653	8.359	7.447	25212.217	7.234	7.821
Finnish	7.566	45711.992	7.566	8.934	6.887	48811.242	6.861	7.826
French	13.605	48268.211	13.605	15.003	12.765	45674.492	12.573	13.682
German	18.355	64015.117	18.356	15.404	17.29	51692.125	16.973	18.972
Greek	3.832	224595.781	3.833	4.527	3.599	80657.891	3.599	4.146
Hungarian	16.365	52828.691	16.363	20.039	14.756	58107.137	14.834	17.633
Indonesian	44.269	33121.945	44.318	48.175	37.909	51611.625	37.838	38.548
Italian	14.859	58908.879	14.860	17.341	13.694	47375.844	13.730	15.207
Japanese	10.888	322031.406	10.896	12.535	10.072	75236.031	10.137	11.661
Korean	4.965	125345.359	4.967	5.649	4.724	90768.844	4.743	5.241
Malay	66.581	22603.727	66.843	74.167	46.885	40468.750	46.912	58.947
Malayalam	5.133	373710.188	5.134	6.396	4.972	16990.266	4.972	5.654
Norwegian	14.425	31526.176	14.427	17.854	13.142	45820.109	13.139	15.041
Persian	6.509	81959.719	6.511	7.628	6.205	92201.812	6.229	7.009
Polish	12.629	66906.469	12.629	14.843	11.414	55923.156	11.311	12.987
Portuguese	15.318	47763.059	15.319	17.297	13.667	51498.402	13.982	15.376
Romanian	10.893	43498.008	10.895	13.061	9.652	54986.055	9.693	10.969
Russian	12.062	170776.750	12.064	13.728	11.048	112574.609	10.948	11.757
Spanish	17.079	51940.859	17.082	18.98	16.351	54005.891	16.138	17.292
Swahili	75.908	29234.168	75.892	89.380	70.519	48802.227	70.402	81.216
Swedish	14.714	49425.969	14.714	17.258	13.229	48622.266	13.337	14.933
Tamil	4.162	381070.844	4.162	5.04	4.028	111060.516	4.049	4.488
Turkish	11.214	46986.391	11.215	13.765	9.834	50303.562	9.763	11.374
Ukrainian	9.409	120719.938	9.409	10.875	8.295	116287.305	8.297	9.076
Vietnamese	5.824	40126.527	5.824	6.614	5.471	42336.426	5.437	5.995

Table 14: LLaMA-2 perplexity on 30 languages with 3% removal ratio. '100K' means that the region is selected from 100,000 samples. 'Top' and 'Bottom' respectively indicate the N parameters with the highest and lowest cumulative $\mathcal{I}_j^*(\theta)$ during the further pre-training across the six languages. 'Random' denotes the randomly selecting N while 'Base' represents no removal. Here, N equals 3% of the total number in each parameter matrix.

Languages	LLaMA-2-7B **3% (10K)**				LLaMA-2-13B **3% (10K)**			
	Base	Top	Bottom	Random	Base	Top	Bottom	Random
Arabic	6.771	115398.328	6.772	7.895	6.261	88678.016	6.315	7.112
Chinese	8.652	369027.531	8.564	9.837	7.838	70912.242	7.806	8.619
Czech	19.834	78480.219	19.837	24.005	17.744	53699.43	17.655	20.485
Danish	8.372	46503.742	8.373	9.929	7.402	39408.91	7.401	8.278
Dutch	16.959	55704.191	16.961	20.121	15.64	41159.938	15.572	18.295
English	7.653	15738.043	7.654	8.359	7.447	23678.322	7.234	7.821
Finnish	7.566	46616.094	7.568	8.934	6.887	47002.539	6.861	7.826
French	13.605	44385.668	13.609	15.003	12.765	38755.539	12.678	13.682
German	18.355	84497.234	18.361	21.404	17.29	43319.586	17.02	18.972
Greek	3.832	147740.5	3.833	4.527	3.599	70136.242	3.6	4.146
Hungarian	16.365	52652.363	16.367	20.039	14.756	48407.305	14.735	17.633
Indonesian	44.269	39055.945	44.267	48.175	37.909	36912.34	37.929	38.548
Italian	14.859	54297.523	14.865	17.341	13.694	42515.969	13.69	15.207
Japanese	10.888	358722.188	10.891	12.535	10.072	68055.984	10.118	11.661
Korean	4.965	102918.828	4.966	5.649	4.724	65209.328	4.736	5.241
Malay	66.581	23501.082	67.158	74.167	46.885	35517.879	47.191	58.947
Malayalam	5.133	314088.969	5.136	6.396	4.972	131629.438	4.971	5.654
Norwegian	14.425	38111.27	14.431	17.854	13.142	38500.664	13.138	15.041
Persian	6.509	78203.031	6.51	7.628	6.205	98292.281	6.22	7.009
Polish	12.629	81373.273	12.633	14.843	11.414	52403.461	11.393	12.987
Portuguese	15.318	47779.789	15.321	17.297	13.667	41184.457	13.86	15.376
Romanian	10.893	45836.578	10.897	13.061	9.652	51766.957	9.694	10.969
Russian	12.062	227916.828	12.061	13.728	11.048	103490.719	11.004	11.757
Spanish	17.079	57679.461	17.087	18.98	16.351	40338.426	16.265	17.292
Swahili	75.908	42977.977	75.93	89.38	70.519	40400.949	70.443	81.216
Swedish	14.714	55893.812	14.717	17.258	13.229	45396.66	13.301	14.933
Tamil	4.162	447989.969	4.162	5.04	4.028	141214.188	4.052	4.488
Turkish	11.214	57037.605	11.215	13.765	9.834	41566.105	9.791	11.374
Ukrainian	9.409	168085.672	9.408	10.875	8.295	94307.312	8.296	9.076
Vietnamese	5.824	36374.734	5.825	6.614	5.471	31730.328	5.467	5.995

Table 15: LLaMA-2 perplexity on 30 languages with 3% removal ratio. '10K' means that the region is selected from 10,000 samples. Here, we reduce training samples to $10,000$ during further pre-training across six languages.

Languages	LLaMA-2-7B **1% (100K)**				LLaMA-2-7B **5% (100K)**			
	Base	Top	Bottom	Random	Base	Top	Bottom	Random
Arabic	6.771	67579496	6.77	7.021	6.771	112504.609	6.774	10.823
Chinese	8.652	120887480	8.561	8.818	8.652	156026.938	8.565	12.775
Czech	19.834	24343856	19.835	21.176	19.834	96580.281	19.845	24.797
Danish	8.372	1631186.625	8.372	8.775	8.372	82876.266	8.375	13.565
Dutch	16.959	6845146	16.963	18.056	16.959	79497.211	16.961	27.01
English	7.653	512756	7.654	7.851	7.653	46197.477	7.656	9.289
Finnish	7.566	4727027.5	7.567	7.948	7.566	60183.328	7.56	13.005
French	13.605	4768049	13.608	14.198	13.605	87642.109	13.611	19.076
German	18.355	17940508	18.357	19.724	18.355	106160.992	18.364	28.772
Greek	3.832	14242545	3.833	3.972	3.832	141320.578	3.835	6.45
Hungarian	16.365	130584	16.366	17.35	16.365	77265.188	16.369	30.376
Indonesian	44.269	1654245	44.347	49.476	44.269	83353.344	44.298	64.743
Italian	14.859	5265871.5	14.863	15.607	14.859	83076.164	14.865	22.6
Japanese	10.888	28104000	10.88	11.196	10.888	124647.633	10.895	16.619
Korean	4.965	16449047	4.965	5.095	4.965	59954.559	4.967	7.831
Malay	66.581	7875206	66.673	78.545	66.581	51824.859	66.751	90.933
Malayalam	5.133	7151096	5.133	5.359	5.133	182008.484	5.137	7.905
Norwegian	14.425	4223085	14.429	15.35	14.425	79399.109	14.434	23.621
Persian	6.509	2233196	6.507	6.782	6.509	107342.734	6.511	10.236
Polish	12.629	6547834.5	12.631	13.36	12.629	88912.945	12.632	21.372
Portuguese	15.318	6249820	15.319	15.927	15.318	78851.766	15.324	22.608
Romanian	10.893	5251915.5	10.895	11.526	10.893	71228.375	10.899	19.21
Russian	12.062	17596800	12.061	12.067	12.062	102639.602	12.066	18.504
Spanish	17.079	8220832.5	17.084	18.029	17.079	96575.547	17.084	24.007
Swahili	75.908	7875009	75.845	83.963	75.908	77765.133	75.8	131.709
Swedish	14.714	4712167.5	14.716	15.534	14.714	81574.734	14.717	22.628
Tamil	4.162	20660974	4.162	4.265	4.162	173728.312	4.164	5.881
Turkish	11.214	4489915	11.214	11.882	11.214	58347.055	11.218	19.76
Ukrainian	9.409	11689088	9.409	9.811	9.409	90008.312	9.414	14.807
Vietnamese	5.824	2235468	5.825	6.018	5.824	54187.02	5.825	9.015

Table 16: LLaMA-2-7B perplexity on 30 languages with 1% and 5% removal ratio. '100K' means that the region is selected from 100,000 samples. Here, we change the removal ratio from 3% to 1% and 5%.