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.

have a strong correlation with the model's linguistic competence. 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 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

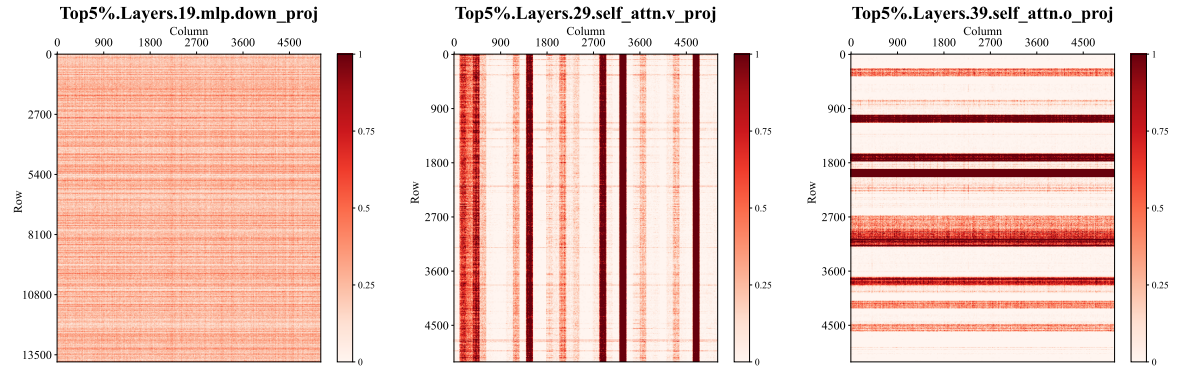

Figure 2: Visualization of the linguistic competence region (the 'Top' 5% region). The scale from 0 to 1 (after normalization) represent the proportion of parameters within a 3×3 vicinity that belong to the 'Top' region.

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.

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 multilingual linguistic competence. When 'Top' region is zeroed-out and frozen, other regions difficultly adapt to regain the core linguistic competence.

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. In Appendix B, we also discover that in various layers, the core linguistic region is concentrated on different heads of the Attn.o matrix. Such distribution features seem to imply that the model's linguistic competence is localized in specific rows and columns.

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, removals 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 C.

Single Dimension Perturbation Here, we explore whether a specific dimension significantly impacts the model's linguistic competence. As illustrated in Figure 3, we iterate through the key dimensions mentioned in Section 3.3, attempting to perturb (random initialization) the same dimension across all Transformer layers. The results indicate that the impact of dimensions 2100 and 4743 on the LLaMA-2-13B substantially surpasses other dimensions, even when compared to the other three in the 'Top 5' dimensions. In contrast, perturbing two randomly selected dimensions, 2800 and 4200, yields linguistic performance almost indistinguishable from the unperturbed state.

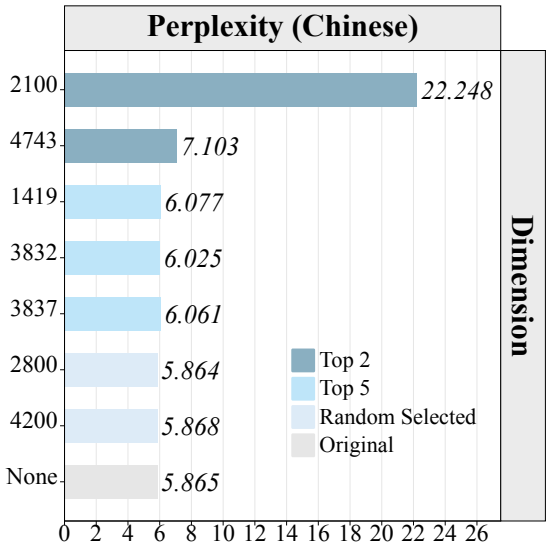

Figure 3: Perplexity of the LLaMA-2-13B when perturbing one single dimension (Att.O and FFN.down columns) across all layers. 'Top k' represents the top k dimensions that disrupt the model the most. 'Random selected' refers to a randomly chosen dimension. 'Original' indicates that no dimensions are disrupted.

Removal Region	N_d	LLaMA-2-**13B Top(100K)**	
		w/ outlier d	**w/o** outlier d
Attn.o(row) Attn.k/q/v(column) FFN.down(column)	1	5218.1	11.079
	3	37344.078	77.519
	5	41840.613	590.895
	10	62579.516	513998.437
Attn.o(row) Attn.k/q/v(column)	1	10.899	10.901
	3	44.384	44.389
	5	33.52	29.793
	10	118.968	120.977
Attn.o(column) Attn.k/q/v(row)	1	17.609	13.666
	3	313.178	63.99
	5	526.464	163.388
	10	5841.446	2675.347
FFN.up/gate(row) FFN.down(column)	1	154.925	6.142
	3	33995.949	6.668
	5	32572.888	8.139
	10	524867.687	45.408

Table 4: Perplexity of LLaMA-2-13B after removing 'Top' certain dimensions w/ or w/o outlier dimensions respectively. Here, N_d denotes the number of dimensions to remove, 'Top' refers to the dimensions with the most cumulated $\mathcal{I}_\theta$ during further pre-training.

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 8, 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. However, randomly altering the parameters at dimensions 2800 and 4200 doesn't noticeably impact the model. For more details, refer to Appendix D.

Ablation Study Considering that the collapse of PPL may be possibly caused solely by the removal of outlier dimensions, rather than the collective effect of the entire linguistic region, we conduct an ablation experiment without removing the parameters of outlier dimensions (1512/2533 for LLaMA-2-7B and 2100/4743 for LLaMA-2-13B). The results for the 13B model, presented in Table 4, indicate that while outlier dimensions are contained within the core linguistic region, non-outlier dimensions in this region are also critical. Not altering any row or column of the matrix parameters for the two outlier dimensions also results in a great increase in PPL, although the collapse of PPL is slowed. This finding reveals that all dimensions in the core linguistic region are tightly interrelated, and disrupting even a small part of it can lead to a PPL collapse. For a more detailed analysis of the experiments and results for the 7B model, please refer to Appendix E.

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

3.4 Monolingual Region

In this section, we wonder if LLMs possess distinct regions within different individual languages. Unlike the core linguistic regions described in Section 3.2, a monolingual 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 (Ara-