Unveiling Linguistic Regions in Large Language Models

Zhihao Zhang[1*], **Jun Zhao**[1*], **Qi Zhang**[13†], **Tao Gui**[2], **Xuanjing Huang**[1]
[1] School of Computer Science, Fudan University
[2] Institute of Modern Languages and Linguistics, Fudan University
[3] Shanghai Collaborative Innovation Center of Intelligent Visual Computing
{zhangzhihao19, zhaoj19, qz, tgui, xjhuang}@fudan.edu.cn

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 a lack 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 dependence, perturbations to even a single parameter on specific dimensions leading to a loss of linguistic competence. Moreover, we discover that distinct monolingual regions exist for different languages, 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 phenomenon observed during further pre-training of LLMs. Overall, exploring the LLMs' functional regions provides insights into the foundation of their intelligence [1].

1 Introduction

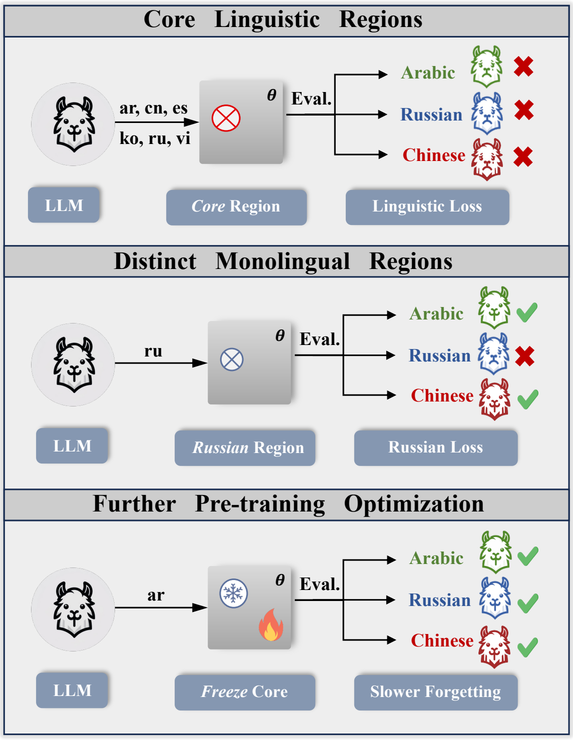

Figure 1: Three main findings of our experiments: (1) Identification of core language regions within the LLMs, where removals lead to linguistic competence loss; (2) Discovery of monolingual regions, where removals cause significant proficiency loss in specific languages; (3) Optimization of freezing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

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), showcase 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,

[*] Equal contributions.
[†] Corresponding authors.
[1] Our code is released in https://github.com/zzhang0179/Unveiling-Linguistic-Regions-in-LLMs.

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 (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 investigation into 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 focuses 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 alignment and generalization? By conducting further pre-training across six languages and evaluating models' parameter importance (Section 2.2), we discover a region in LLMs corresponding to the core 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 leads 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 dependence. In certain dimensions, only perturbing a single parameter could lead to the model losing its linguistic competence (Section 3.3). Additionally, ablation study in 3.3 shows that beyond outlier dimensions, other non-outlier dimensions in this region are also critical.

Q2: Beyond the core linguistic region within LLMs , do distinct monolingual regions exist that specifically influence individual languages? While LLMs possess strong multilingual capabilities, we discover that each individual language (or language with similar compositional elements or grammatical structures) 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 phenomenon 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 during further pre-training, compared to the full-scale fine-tuning technique. Findings indicate that this method facilitates comparable 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; Wei et al., 2023), 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. Removing a specific monolingual region causes a significant deterioration in the linguistic capabilities within corresponding language.

- We perform further pre-training for specific languages within the core linguistic region of LLMs frozen, achieving comparable 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 for auto-regressive language modeling is to find the optimal θ that minimizes 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, similar to several prior works (Molchanov et al., 2019; Zhang et al., 2022), using 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 backpropagation.

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 the core linguistic region in LLMs. Here, we define the region as "***core linguistic***", attributed to its minimal proportion within LLMs, constituting only 1% of the total parameters, and its association with model's linguistic competence modeling.

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