Converging to a Lingua Franca: Evolution of Linguistic Regions and Semantics Alignment in Multilingual Large Language Models

Hongchuan Zeng[1], **Senyu Han**[1], **Lu Chen**[1,2†], **Kai Yu**[1,2† *]
[1]X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence, SJTU AI Institute
Shanghai Jiao Tong University, Shanghai, China
[2]Suzhou Laboratory, Suzhou, China
{charlie68, cnlnpjhsy, chenlusz, kai.yu}@sjtu.edu.cn

Abstract

Large language models (LLMs) have demonstrated remarkable performance, particularly in multilingual contexts. While recent studies suggest that LLMs can transfer skills learned in one language to others, the internal mechanisms behind this ability remain unclear. We observed that the neuron activation patterns of LLMs exhibit similarities when processing the same language, revealing the existence and location of key linguistic regions. Additionally, we found that neuron activation patterns are similar when processing sentences with the same semantic meaning in different languages. This indicates that LLMs map semantically identical inputs from different languages into a "Lingua Franca", a common semantic latent space that allows for consistent processing across languages. This semantic alignment becomes more pronounced with training and increased model size, resulting in a more language-agnostic activation pattern. Moreover, we found that key linguistic neurons are concentrated in the first and last layers of LLMs, becoming denser in the first layers as training progresses. Experiments on BLOOM and LLaMA2 support these findings, highlighting the structural evolution of multilingual LLMs during training and scaling up. This paper provides insights into the internal workings of LLMs, offering a foundation for future improvements in their cross-lingual capabilities.

1 Introduction

In recent years, large language models (LLMs) have gained significant attention for their remarkable performance. The multilingual capabilities of LLMs are a crucial area of research, especially as AI technology spreads to people with diverse backgrounds and different native languages. Interestingly, recent studies have shown that LLMs can develop cross-lingual abilities, transferring skills learned in one language to others they have not been trained on (Chirkova, Nikoulina, 2024; Pires et al., 2019; Wu, Dredze, 2019). However, the internal mechanisms by which multilingual LLMs function and develop cross-lingual abilities remain an understudied topic.

In the field of neuroscience, research has shown some interesting findings. First, when polyglots process different languages, their brains' language networks exhibit distinct response patterns (Malik-Moraleda et al., 2024). It is believed that different language capacities are stored in different compartments of the human brain (Paradis, 1985, 2000). Second, while processing the same task in different languages, the human brain exhibits similar activation patterns (Xu et al., 2021). Third, as

*†Lu Chen and Kai Yu are the corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

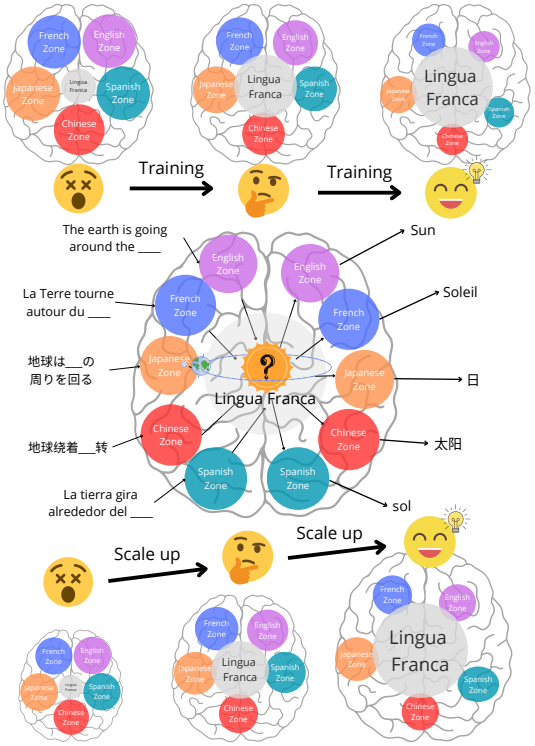

Figure 1: LLMs encode inputs into a *"Lingua Franca"*, a latent semantic space representation shared by all languages, and then decode this *"Lingua Franca"* into the target language. As training progresses and the models scale up, LLMs become better at mapping inputs to this common semantic space.

individuals gain proficiency in a new language, the activation pattern for that language becomes more similar to those of other languages (Nichols et al., 2021; Li et al., 2019). These findings raise the question: Do these phenomena also manifest in LLMs? Our answer is yes. We found that LLMs use different neurons to process inputs in different languages, and map inputs with the same semantic meaning but in different languages into a *"Lingua Franca"*, a common semantic latent space shared by all languages.

First, by examining neuron activation in LLMs, *we observe that neuron activation exhibits similar patterns when processing different inputs in the same language*. Zhang et al. (2024) indicates that certain key parameters in LLMs correspond to linguistic competence, with language-specific parameters existing for different languages. Our research supports this view. By probing the neurons that contribute most to this similarity, we can identify the key linguistic region for a specific language. This key linguistic region consists of neurons responsible for processing specific languages, with each language having its own dedicated region. *When the key linguistic region of a specific language is deactivated, the LLM significantly loses its capacity for the specific language while maintaining its capacity for others.*

Second, we found that *when processing inputs with the same semantic meaning but in different languages, LLM neuron activation shows similar patterns.* This indicates that multilingual LLMs map inputs into a common semantic latent space, allowing them to process information similarly across languages, facilitating cross-lingual ability transfer. We refer to this phenomenon as **semantic alignment**.

Third, we found that as training progresses, the sizes of key linguistic regions become smaller, and the activation pattern becomes more language-agnostic. At the same time, semantic alignment becomes more significant. Similarly, as the model scale increases, the activation pattern of neurons become more language-agnostic, but the semantic alignment becomes more pronounced. We defined metrics to facilitate the comparison of linguistic region distinctions and semantic alignment.

By examining the internal structure of LLMs, we found that *the key linguistic region neurons are generally located in the first and last few layers.* As training progresses and model scale increases, these key regions become denser in the first layers. Based on this information, we hypothesize that LLMs encode inputs into a *"Lingua Franca"*, a latent semantic space representation shared by all languages, and then decode this *"Lingua Franca"* into the target language. As training or model size increases, the model can more efficiently map inputs with the same semantic meaning to a common semantic space and then perform reasoning.

The experiments were primarily conducted on BLOOM (Workshop, 2023), using their released intermediate checkpoints to examine the evolution during the training process. To showcase the extensibility of our results on other models, we equally performed the experiments on LLaMa2 (Touvron et al., 2023), and we obtained similar results.

The following is a summary of our observations:

Neuron Activation Patterns:

- Neuron activation exhibits similar patterns when processing the same language, revealing the existence and location of key linguistic regions in LLMs. Deactivating these key neurons significantly impairs performance in the corresponding language.
- Neuron activation exhibits similar patterns when processing sentences with the same semantic meaning in different languages.

Dynamics with Training and Scaling Up:

- As the training process progresses, linguistic regions become smaller, while semantic alignment becomes more significant, resulting in a more language-agnostic activation pattern.
- As the model's scale grows, the activation becomes more language-agnostic, and semantic alignment becomes more pronounced.
- Important neurons are generally located in the first and last few layers. As training steps increase and model scale grows, key regions become denser in the first layers.

2 Background

Large language models like GPT-4 (OpenAI, 2024), LLaMA (Touvron et al., 2023), and OPT (Zhang et al., 2022) have revolutionized natural language processing with their ability to understand and generate nuanced text. Additionally, multilingual large language models such as BLOOM (Workshop, 2023) and XLM-R (Conneau et al., 2020) overcome language barriers by learning universal representations from texts in multiple languages. Multilingual large language models generally incorporate multilingual data in the pretraining stage for better alignment (Qin et al., 2024). These models typically use the transformer architecture (Vaswani et al., 2023) and are decoder-only, with each layer composed of an attention module and a multilayer perceptron module.

The attention module is crucial in transformer models, allowing the model to focus on different parts of the input sequence, assigning varying importance to each token. It enables the model to dynamically prioritize important information in the input, enhancing its ability to capture dependencies and relationships, thereby improving performance.

A multilayer perceptron (MLP) consists of fully connected neurons with a nonlinear activation function, organized in at least three layers, and is notable for its ability to distinguish data that is not linearly separable. In BLOOM, the operations can be expressed as follows:

$$\mathbf{z} = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1, \quad \mathbf{h} = \sigma(\mathbf{z}), \quad \mathbf{y} = \mathbf{W}_2\mathbf{h} + \mathbf{b}_2, \tag{1}$$

where $\mathbf{x}$ is the input vector, $\mathbf{W}_i$ are the weight matrices, $\mathbf{b}_i$ are the bias vectors, $\mathbf{h}$ is the hidden layer output, $\mathbf{y}$ is the final output, and $\sigma(\cdot)$ is an activation function. In BLOOM, the activation is GeLU (Hendrycks, Gimpel, 2023).

In LLaMA, the MLP module can be expressed as:

$$\mathbf{h}_1 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1, \quad \mathbf{h}_2 = \mathbf{W}_2\mathbf{x} + \mathbf{b}_2, \tag{2}$$

$$\text{SwiGLU}(\mathbf{x}) = \text{Swish}(\mathbf{h}_1) \odot \mathbf{h}_2, \tag{3}$$

$$\mathbf{y} = \mathbf{W}_3 \cdot \text{SwiGLU}(\mathbf{x}) + \mathbf{b}_3, \tag{4}$$

where Swish function is defined in Ramachandran et al. (2017).

As pointed out by Dai et al. (2022), MLP weights can store complex syntactical and semantic patterns, which are the building blocks of language. Therefore, *in the following parts of our paper, we refer to neurons of BLOOM as* $\mathbf{h}$*, representing the hidden layer output, and neurons of LLaMA as the output of SwiGLU(*$\mathbf{x}$*).*

3 Methods

3.1 Measuring Activation Similarity of Neurons

In order to decouple the key linguistic regions and the *"Lingua Franca"*, which is the latent semantic space representation shared by all languages, we use a **parallel corpus** to activate the neurons of the LLM. We record the neuron activation results (e.g. the hidden layer output $\mathbf{h}$ of MLP layer in BLOOM or the output of SwiGLU$(\mathbf{x})$ in LLaMA) as the LLM processes each token. We then average these results across all tokens in each sample. By concatenating the averaged results from all layers, we create an activation vector for each sample and normalize it. Finally, we calculate the cosine similarity between each pair of samples to generate the similarity map.

Neuron Activation Extraction:

$$\overline{\mathbf{h}}_m^{s_i} = \frac{1}{T_{s_i}} \sum_{t=1}^{T} \mathbf{h}_{m,t}^{s_i}, \tag{5}$$

where $\mathbf{h}_{m,t}^{s_i}$ represents the activation of token t in layer m for sample s_i, and T_{s_i} is the number of tokens in the sample s_i. Motivated by Sentence-BERT Reimers, Gurevych (2019), we apply a mean pooling strategy across all tokens to obtain the representation of a sample sentence.

Concatenation of Layer Activations:

$$\mathbf{a}^{s_i} = \frac{1}{|| \left[\overline{\mathbf{h}}_1^{s_i} \,\middle|\, \overline{\mathbf{h}}_2^{s_i} \,\middle|\, \cdots \,\middle|\, \overline{\mathbf{h}}_M^{s_i} \right] ||} \left[\overline{\mathbf{h}}_1^{s_i} \,\middle|\, \overline{\mathbf{h}}_2^{s_i} \,\middle|\, \cdots \,\middle|\, \overline{\mathbf{h}}_M^{s_i} \right], \tag{6}$$

where M is the total number of layers.

Cosine Similarity Map:

$$S_{ij} = \text{Similarity}(s_i, s_j) = \frac{\mathbf{a}^{s_i} \cdot \mathbf{a}^{s_j}}{\|\mathbf{a}^{s_i}\| \|\mathbf{a}^{s_j}\|} = \mathbf{a}^{s_i} \cdot \mathbf{a}^{s_j}, \tag{7}$$

as $\mathbf{a}^{s_i}$ is normalized.

3.2 Metrics on the Development of Linguistic Regions and Semantic Alignment

To measure how closely related the activation is to language-specific information, we define Linguistic Regions Development Scores (**LRDS**). Specifically, LRDS measures the difference between the average similarity of samples in the same language and samples in different languages, with all sample pairs having different semantic meanings. The sample numbers in each language are equal. *A lower LRDS indicates that the activation pattern is more language-agnostic.*

$$\begin{aligned} \textbf{LRDS} = & \, Average(S_{ij} \mid \text{lang}(s_i) = \text{lang}(s_j), \\ & \qquad \text{semantics}(s_i) \neq \text{semantics}(s_j)) \\ & - Average(S_{ij} \mid \text{lang}(s_i) \neq \text{lang}(s_j), \\ & \qquad \text{semantics}(s_i) \neq \text{semantics}(s_j)). \end{aligned} \tag{8}$$

The Size of Key Linguistic Regions (**SKLR**) is the sum of the sizes of the key linguistic regions for all languages, which we will present further in Section 3.3. This metric evaluates how computationally costly it is to align inputs in different languages to the semantic space.

To measure how closely related the activation is to the semantic meaning of the inputs instead of the language of the inputs, we define Semantic Alignment Development Scores (**SADS**). Specifically,

SADS measures the difference between the average similarity of samples with the same meaning and samples with different meanings, with all sample pairs in different languages. *A higher SADS indicates that the activation pattern is more related to the semantic meaning of the inputs.*

$$\begin{aligned}\textbf{SADS} = Average(S_{ij} \mid &\text{semantics}(s_i) = \text{semantics}(s_j),\\ &\text{lang}(s_i) \neq \text{lang}(s_j))\\ - Average(S_{ij} \mid &\text{semantics}(s_i) \neq \text{semantics}(s_j),\\ &\text{lang}(s_i) \neq \text{lang}(s_j)).\end{aligned} \tag{9}$$

3.3 Key Linguistic Region Probing

We believe that neurons activated in a similar pattern across different samples in one language are key neurons for that specific language. Therefore, we assign a score to these neurons to evaluate their contribution to the average similarity across the samples in that specific language.

The average similarity $\overline{S}_l$ of samples in one specific language l can be expressed by :

$$\begin{aligned}\overline{S}_l &= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S_{ij}\\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{a}^{s_i} \cdot \mathbf{a}^{s_j}\\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{K} \mathbf{a}_{(k)}^{s_i} \cdot \mathbf{a}_{(k)}^{s_j}\\ &= \frac{2}{n(n-1)} \sum_{k=1}^{K} (\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{a}_{(k)}^{s_i} \cdot \mathbf{a}_{(k)}^{s_j}),\end{aligned} \tag{10}$$

where n is the total number of sample in language l $\mathbf{a}_{(k)}^{s_i}$ is the activation of neuron k (the k component of activation vector $\mathbf{a}^{s_i}$) for a sample s_i, and K the total number of neurons (i.e. the length of activation vector).

We may see that, the contribution of neuron k to the average cosine similarity can be quantified by the term in the bracket. We thus define the contribution score $\overline{S}_l^{(k)}$ of a neuron k to one language l as:

$$\overline{S}_l^{(k)} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{a}_{(k)}^{s_i} \cdot \mathbf{a}_{(k)}^{s_j}. \tag{11}$$

Since some neurons are always activated, their contribution to the average similarity is consistently large. Therefore, we need to identify neurons that contribute exceptionally to the average similarity for a specific language. To do this, we calculate the standard score (z-scores, $z_l^{(k)}$) of the contribution score of a neuron k across different languages:

$$z_l^{(k)} = \frac{\overline{S}_l^{(k)} - \mu_k}{\sigma_k}, \tag{12}$$

where μ_k and σ_k are the mean and standard deviation of the contribution scores of the neuron k across all languages, respectively:

$$\mu_k = \frac{1}{L} \sum_{l=1}^{L} \overline{S}_l^{(k)}, \tag{13}$$

$$\sigma_k = \sqrt{\frac{1}{L} \sum_{l=1}^{L} (\overline{S}_l^{(k)} - \mu_k)^2}. \tag{14}$$

Here, L is the total number of languages. The value $z_l^{(k)}$ represents the extent to which the neuron k contribute exceptionally to language l. In practice, we set a threshold for the z-scores, and neurons with z-scores higher than this threshold for a specific language are considered part of the key linguistic region for that language. For example, if the threshold is set to 2, the neurons we select have activation scores that are more than 2 standard deviations above their average scores.

4 Experiments

We first examine the similarity of activation patterns within an individual model. To illustrate the language-wise and semantic-wise similarity when LLMs process inputs, we plot the similarity map of neuron activation while processing inputs in the same language or with the same semantic meaning.

Next, using the obtained similarity patterns, we locate the key linguistic region for each language. By deactivating the key region of each language respectively, we assess the performance of LLMs, demonstrating the existence and utility of these regions.

We then explore the dynamic changes during training and scaling up. We calculate Linguistic Regions Development Scores (**LRDS**), Size of Key Linguistic Regions (**SKLR**), and Semantic Alignment Development Scores (**SADS**) for different training checkpoints to illustrate the development of linguistic regions and semantic alignment. We perform the same analysis for models of different scales to show the evolution trend.

4.1 Experimental Setup

Models. The experiments were primarily conducted using the BLOOM (Workshop, 2023) model family, which features a high percentage of multilingual training data and is well-balanced across various languages. We tested various models within this family, including the 560m, 1.1b, 1.7b, 3B, and 7.1b models, as well as the intermediary checkpoints of BLOOM-7b1. Additionally, we conducted complementary experiments on the LLaMA-2 model family to examine scenarios where multilingual training data is limited.

Datasets & Language Selection. We used the Bible dataset (Christodouloupoulos, Steedman, 2015), a perfectly aligned parallel corpus, to activate the LLM. Each sample consists of a verse, and we randomly selected 100 verses, taking translations in different languages from the dataset. To evaluate multilingual perplexity, we employed the XL-Sum dataset (Hasan et al., 2021), following the implementation of Zeng et al. (2024a). XL-Sum contains high-quality articles from the BBC covering 45 languages. Our experiments focused on a subset of 9 languages available in both BLOOM and XL-Sum: Arabic (`ar`), Chinese (`zh`), English (`en`), French (`fr`), Hindi (`hi`), Indonesian (`id`), Portuguese (`pt`), Spanish (`es`), and Vietnamese (`vi`).

Evaluation. We evaluated the perplexity of the models separately for each language using XL-Sum. Additionally, we designed a task to assess the cross-lingual reasoning ability of LLMs, employing the widely recognized EleutherAI-eval-harness framework (Gao et al., 2023). For this, we used the XStoryCloze dataset, which consists of a short story typically composed of four sentences and two alternative endings. One ending logically completes the story, while the other does not. The task for the model is to identify the more plausible ending. We prompted the story in various languages and asked the model to choose between the two ending options in English. Prompting the stories and endings in different languages helps us determine whether the model effectively "understands" the stories and maps them into a mutual semantic space, rather than performing the reasoning process in only one language. The prompt format and evaluation details are presented in Appendix B.

Perplexity increase %↑	Full model Perplexity	Random 10%	×en	×zh	×fr	×es	×pt	×ar	×vi	×hi	×id
en	13.94	12%	22%	6%	2%	2%	1%	55%	18%	7%	5%
zh	24.01	11%	3%	47%	2%	1%	1%	47%	16%	8%	4%
fr	9.62	10%	5%	5%	20%	4%	2%	49%	14%	8%	5%
es	10.84	10%	5%	5%	2%	17%	3%	48%	14%	8%	5%
pt	11.17	11%	5%	5%	2%	6%	27%	50%	15%	8%	6%
ar	14.45	12%	4%	5%	2%	2%	1%	309%	16%	6%	4%
vi	10.11	12%	4%	7%	2%	2%	1%	64%	83%	7%	6%
hi	11.14	10%	3%	4%	1%	1%	1%	109%	15%	220%	3%
id	20.55	13%	5%	7%	2%	2%	1%	80%	18%	12%	1498%
Key Neuron Number		49152	15935	31185	14040	7182	8865	46313	45758	22285	15201
Key Neuron Percentage		10%	3.2%	6.3%	2.9%	1.5%	1.8%	9.4%	9.3%	4.5%	3.1%

Table 1: Percentage increase in perplexity after deactivating key linguistic region neurons for each language of BLOOM-7b1 model. Each column corresponds to the deactivation of the key region for a specific language. The second column shows the results of deactivating a random 10% of neurons in the LLM. *We can see that the perplexity of the deactivated language (on the diagonal) rises significantly, while the perplexity of other languages remains largely unchanged.*

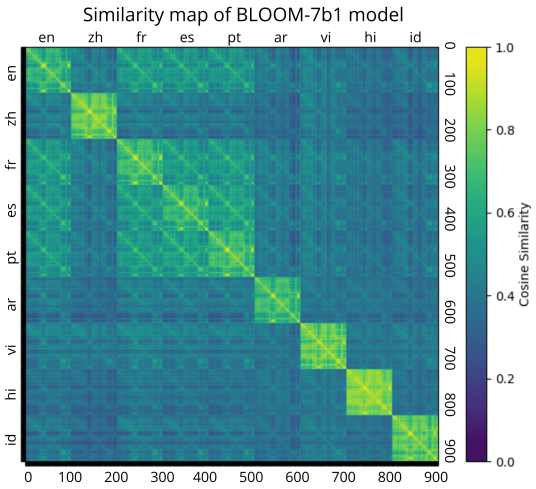

Figure 2: Similarity map of the BLOOM-7b1 model. Each block of 100 samples is in the same language. *Samples in the same language form distinct light blocks, and samples with the same semantic meaning form light bands along the diagonal of these blocks.*

4.2 Observation I: Neuron activation exhibits similar pattern when processing different inputs in the same language or with the same semantic meaning.

To illustrate the language-wise and semantic-wise similarity when LLMs process inputs, we used the BLOOM-7b1 model as an example and plotted the similarity map of all the sample pairs, as shown in Figure 2. Each block of 100 samples is in the same language, and sentences with the same semantic meaning are placed in the same position across all sample blocks. *Within each language block, the similarity of neuron activation is significantly higher, indicating that LLMs' neuron activation exhibits similar patterns when processing inputs in the same language.*

Additionally, *we observed bright bands on the diagonal of the off-diagonal blocks.* These represent sentences in different languages but with the same semantic meaning. Even for languages like Chinese and Arabic, which use different alphabets and share no common tokens with other languages, the activation patterns are similar when processing semantically identical sentences. This suggests that the LLM "understands" the semantic meaning of sentences, encoding the inputs into a common semantic space shared by all languages, allowing it to process information similarly across languages.

Interestingly, we also observed that linguistically similar languages, such as French, Spanish, and Portuguese, exhibit higher cross-lingual similarity compared to other language pairs. This could be because these languages share more tokens and have similar grammatical syntax, leading the LLM to process them in a similar way.

We conducted a layer-wise analysis of the BLOOM-7B1 model, calculating both Linguistic Regions Development Scores (LRDS) and Semantic Alignment Development Scores (SADS) for each layer. The results in Figure 3 show that the first and last layers have higher LRDS and lower SADS, indicating a stronger focus on language-specific information with less emphasis on semantic alignment. In contrast, the middle layers have lower LRDS and higher SADS, suggesting a shift towards greater semantic alignment and more language-agnostic processing. *This indicates that the first and last layers are more language-specific and less semantically focused, whereas the middle layers are more semantically oriented and language-agnostic.*

4.3 Observation II: Similarity of activation patterns allows locating the Key linguistic region of a specific language

To identify the key linguistic region for each language, we calculated the z-scores of each neuron for each language and selected those neurons with a z-score higher than 2 (i.e., their activation scores for a specific language are more than two standard deviations above the average). These selected neurons contribute exceptionally to the specific language being analyzed. We then deactivated the

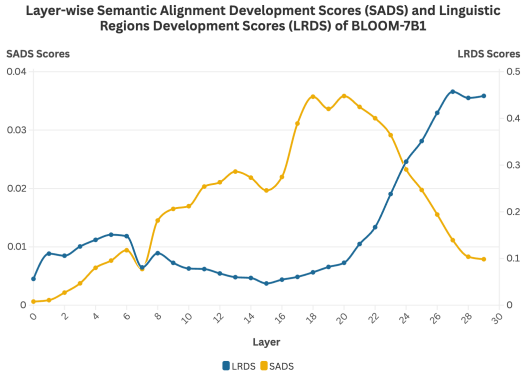

Figure 3: Layer-wise Semantic Alignment Development Scores (SADS) and Linguistic Regions Development Scores (LRDS) of BLOOM-7B1.

key neurons for one specific language and measured the perplexity of the model across all languages. The results are shown in Table 1.

We found that after removing the key region for one language, the perplexity for that language increased significantly, as shown in the diagonal blocks of the table. However, the perplexity for other languages remained almost the same, as shown in the off-diagonal blocks. This indicates that the neurons we identified are crucial for processing the specific language being researched, forming the key linguistic region for that language.

4.4 Observation III: As the training process progresses, the linguistic regions become smaller, and the activation patterns become more language-agnostic.

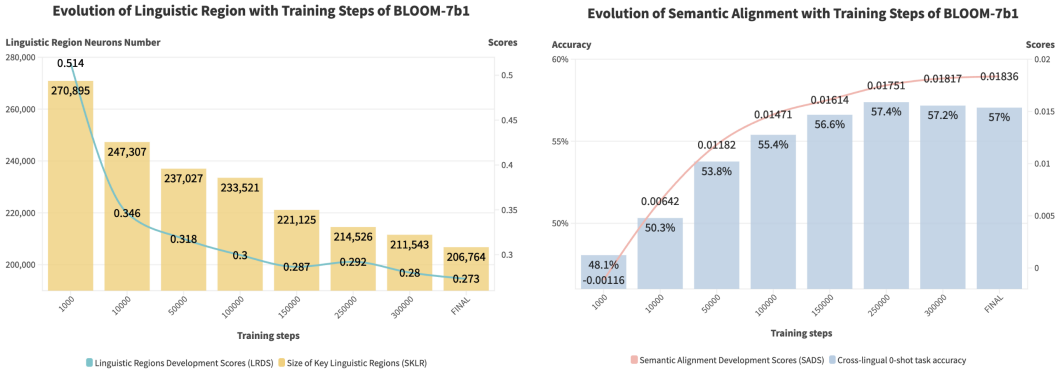

Figure 4: Comparison of the evolution of linguistic regions (left) and semantic alignment (right) with training steps. *As training progresses, the key linguistic regions become smaller, and the neuron activation pattern becomes more language-agnostic. Meanwhile, semantic alignment becomes more pronounced, and the model's cross-lingual reasoning ability improves.*

We can now compare the development of linguistic regions throughout the training process. Based on Observation II, we can locate the key linguistic regions for each language, count the number of key neurons, and thus determine the Size of Key Linguistic Regions (**SKLR**), which is the sum of the sizes of the key linguistic regions for all languages. We can also calculate the Linguistic Regions Development Scores (**LRDS**) using the similarity map. By applying these metrics to the training checkpoints of BLOOM, we can observe the dynamics of key linguistic region development. The results are shown in Figure 4 (left).

At the beginning of the training process, the size of the key linguistic regions is large, indicating a high number of key neurons. At the same time, the LRDS is also high, signifying that the activation pattern is highly language-specific. However, as training progresses, the size of the key linguistic

regions decreases, and the LRDS drops. *This indicates that the activation pattern becomes less related to the specific language and more focused on the semantic meaning of the inputs.* This shift occurs because the model becomes more familiar with the languages and requires less effort to "understand" the sentences and project them into the common semantic space.

If we examine the distribution of key neurons, we find that *the key neurons are generally located in the first and last few layers*. As the number of training steps increases, these key regions become denser in the first few layers, as shown in Figure 5. The first few layers are likely related to the encoding process from the source language to the common semantic space, while the latter layers correspond to the decoding process from the latent semantic representation to the target language. As training progresses, the model can more efficiently encode information from different languages into the semantic space and then decode this representation using fewer neurons into the target language.

4.5 Observation IV: As the training process progresses, the semantic alignment phenomenon becomes more significant.

To examine the semantic alignment phenomenon throughout the training process, we evaluated the Semantic Alignment Development Scores (**SADS**) and the cross-lingual zero-shot performance at different checkpoints of BLOOM-7b1. The results are shown in Figure 4 (right). As the training process progresses, the SADS increases, and the cross-lingual reasoning ability of the model improves. *This indicates that the activation becomes more strongly related to the semantic meaning of inputs rather than linguistic information.* Consequently, the model can gradually better align inputs into the common semantic space of different languages.

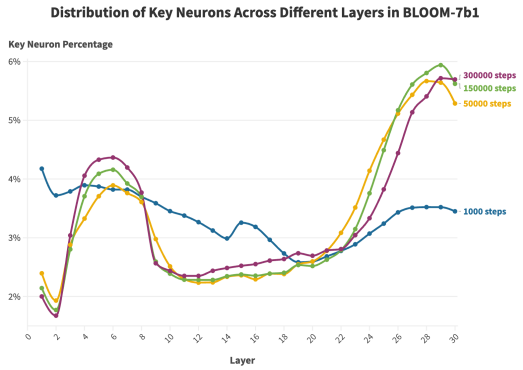

Figure 5: Distribution of Key Neurons Across Different Layers. We may see that as the training steps grows, the key regions become denser in the first layers, facilitating the encoding from inputs to the *"Lingua Franca"*.

By observing the size of the key linguistic region for each language (Figure 6), we can see that languages with less similar linguistic features, such as Vietnamese, Arabic, and Chinese, which belong to very different language families and use distinct writing systems, have significantly larger key regions compared to their counterparts. This suggests that the model requires more effort to align these languages due to their distinct representations and linguistic features, making them generally more challenging to align within the mutual semantic space.

4.6 Observation V: As the model scale grows, the model aligns languages better to the semantic space

We evaluated the LRDS, SADS, and cross-lingual zero-shot accuracy across different scales of BLOOM models (Figure 7). We can clearly see that as the model scale grows, the neurons' activation patterns become less related to language and more focused on the semantic meaning of inputs. At the same time, the cross-lingual reasoning ability of the model improves with larger scale. We conclude that as the model scale grows, LLMs can better align inputs in different languages to the common semantic space, enhancing their reasoning process within this shared semantic latent space.

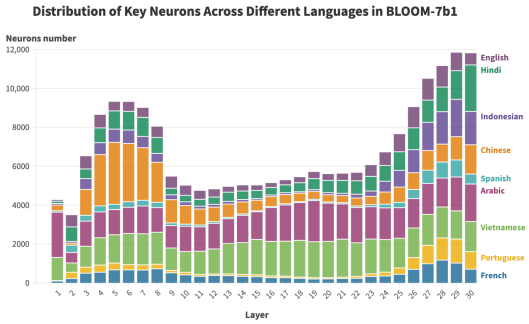

Figure 6: Distribution of Key Neurons Across Different Languages.

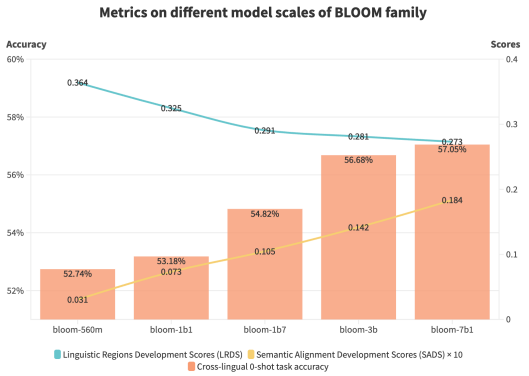

Figure 7: Metrics across different model scales of the BLOOM family.

4.7 Extension to LLaMA2

We performed the same experiments on LLaMA2. Since the intermediate checkpoints of LLaMA2 are not released, we conducted the analysis on different scales of the model, specifically on the 7b, 13b, and 70b models. We obtained similar results to the experiments performed on BLOOM, namely the existence of linguistic regions, semantic alignment, and the effects of scaling up the model. The results are shown in the Appendix C.

5 Related Work

Multilingual Large Language Models. Large language models (LLMs) such as GPT-4 (OpenAI, 2024), LLaMA (Touvron et al., 2023), and OPT (Zhang et al., 2022) have revolutionized natural language processing by demonstrating the ability to understand and generate nuanced text. Multilingual LLMs like BLOOM (Workshop, 2023) and XLM-R (Conneau et al., 2020) further extend these capabilities by learning universal representations from texts in multiple languages. These models typically use the transformer architecture (Vaswani et al., 2023), incorporating multilingual data during pretraining to improve alignment and performance across languages (Qin et al., 2024). To align different languages' capacity, various alignment strategies have been proposed, including alignment during pre-training Blevins, Zettlemoyer (2022); Briakou et al. (2023); Holmström et al. (2023), supervised fine-tuning Gao et al. (2024); Fu et al. (2022); Cui et al. (2024), reinforcement learning from human feedback Zeng et al. (2024b); Dong et al. (2023); Sun et al. (2024), downstream fine-tuning Aggarwal et al. (2024); Rosenbaum et al. (2022); Shaham et al. (2024), prompt tuning Qin et al. (2023) and contrastive learning Li et al. (2024).

Neuroscience. In the field of neuroscience, considerable research has been conducted to understand how the human brain processes multiple languages. Studies have shown that polyglots—individuals who speak multiple languages—exhibit distinct patterns of brain activation for different languages (Malik-Moraleda et al., 2024). These findings suggest that different languages are stored and

processed in different compartments of the brain (Paradis, 1985, 2000). Additionally, research by Xu et al. (2021) has indicated that similar brain activation patterns can occur when processing the same tasks in different languages, suggesting a common neural mechanism underlying multilingual processing.

Linguistic and Semantic alignment. Previous research indicates that the representations generated by multilingual encoder models are moderately language-agnostic Pires et al. (2019); Libovický et al. (2020). Based on this assumption, Yoon et al. (2024) introduced a LangBridge model to connect a multilingual encoder to a monolingual LLM, effectively achieving promising performance. They found equally that the efficacy of LangBridge stems from the language-agnostic characteristics of multilingual representations. Ding et al. (2022) proposes targets to transfer English embeddings to virtual multilingual embeddings without semantic loss, thereby improving cross-lingual transferability. Zhang et al. (2024) discovered a core region in LLMs that corresponds to linguistic competence, freezing the core linguistic region during further pre-training can mitigate the issue of catastrophic forgetting. Wendler et al. (2024) showcased that multilingual language models trained on unbalanced, English-dominated corpora use an abstract "concept space" laying closer to English as an internal pivot.

6 Conclusion

This paper explores the internal mechanisms of multilingual LLMs. We found that neuron activation patterns in LLMs are similar when processing the same language, allowing us to identify key neurons for specific languages. Deactivating these neurons significantly impairs performance in those languages. Additionally, we discovered that neuron activation patterns are similar when processing semantically identical sentences in different languages. This indicates that LLMs map these inputs into a common latent space. As training progresses, key linguistic regions become smaller and neuron activation becomes more focused on semantic meaning and less on language specifics. Key neurons are mainly located in the first and last layers, becoming denser in the first layers with training. Moreover, larger models align languages better, enhancing cross-lingual reasoning abilities. Our findings provide insights into the structural evolution of multilingual LLMs during training and scaling, offering a foundation for improving their cross-lingual capabilities.

Limitations

While our study provides valuable insights into the internal mechanisms of multilingual LLMs, it has several limitations. Firstly, our analysis primarily focuses on the BLOOM and LLaMA2 models. Although these models are representative, the findings may not fully generalize to other multilingual LLM architectures. Future research should examine a broader range of models to validate our conclusions. In particular, investigating the evolution of the capabilities of monolingual models as they undergo continuous training with multilingual data could be a very interesting research subject. Secondly, we rely on specific datasets, such as the Bible dataset and XL-Sum, for our experiments. These datasets, while diverse, may not cover all linguistic nuances and complexities. Utilizing a wider array of datasets, including those with more diverse and low-resource languages, would provide a more comprehensive evaluation of model performance and neuron activation patterns. Thirdly, our methodology for identifying key neurons and measuring their contributions is based on averaged activation patterns and z-scores. This approach, while effective, may not capture all nuances of neuron interactions and their contributions to language processing. More sophisticated techniques, such as causal inference methods, could provide deeper insights into neuron functionality. Lastly, while we observed significant patterns related to semantic alignment and linguistic region efficiency, the underlying reasons for these patterns remain speculative. Further research is needed to establish causal relationships and to better understand the specific mechanisms through which LLMs achieve cross-lingual competence. These limitations highlight areas for future research to build upon our findings and enhance the understanding and capabilities of multilingual LLMs.

References

Aggarwal Divyanshu, Sathe Ashutosh, Watts Ishaan, Sitaram Sunayana. MAPLE: Multilingual Evaluation of Parameter Efficient Finetuning of Large Language Models. 2024.

Blevins Terra, Zettlemoyer Luke. Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models. 2022.

Briakou Eleftheria, Cherry Colin, Foster George. Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability. 2023.

Chirkova Nadezhda, Nikoulina Vassilina. Zero-shot cross-lingual transfer in instruction tuning of large language models. 2024.

Christodouloupoulos Christos, Steedman Mark. A massively parallel corpus: the bible in 100 languages // Language resources and evaluation. 2015. 49. 375–395.

Conneau Alexis, Khandelwal Kartikay, Goyal Naman, Chaudhary Vishrav, Wenzek Guillaume, Guzmán Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke, Stoyanov Veselin. Unsupervised Cross-lingual Representation Learning at Scale. 2020.

Cui Yiming, Yang Ziqing, Yao Xin. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. 2024.

Dai Damai, Dong Li, Hao Yaru, Sui Zhifang, Chang Baobao, Wei Furu. Knowledge Neurons in Pretrained Transformers. 2022.

Ding Kunbo, Liu Weijie, Fang Yuejian, Mao Weiquan, Zhao Zhe, Zhu Tao, Liu Haoyan, Tian Rong, Chen Yiren. A Simple and Effective Method to Improve Zero-Shot Cross-Lingual Transfer Learning. 2022.

Dong Yi, Wang Zhilin, Sreedhar Makesh Narsimhan, Wu Xianchao, Kuchaiev Oleksii. SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF. 2023.

Fu Jinlan, Ng See-Kiong, Liu Pengfei. Polyglot Prompt: Multilingual Multitask PrompTraining. 2022.

Gao Leo, Tow Jonathan, Abbasi Baber, Biderman Stella, Black Sid, DiPofi Anthony, Foster Charles, Golding Laurence, Hsu Jeffrey, Le Noac'h Alain, Li Haonan, McDonell Kyle, Muennighoff Niklas, Ociepa Chris, Phang Jason, Reynolds Laria, Schoelkopf Hailey, Skowron Aviya, Sutawika Lintang, Tang Eric, Thite Anish, Wang Ben, Wang Kevin, Zou Andy. A framework for few-shot language model evaluation. 12 2023.

Gao Pengzhi, He Zhongjun, Wu Hua, Wang Haifeng. Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models. 2024.

Hasan Tahmid, Bhattacharjee Abhik, Islam Md. Saiful, Mubasshir Kazi, Li Yuan-Fang, Kang Yong-Bin, Rahman M. Sohel, Shahriyar Rifat. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, VIII 2021. 4693–4703.

Hendrycks Dan, Gimpel Kevin. Gaussian Error Linear Units (GELUs). 2023.

Holmström Oskar, Kunz Jenny, Kuhlmann Marco. Bridging the Resource Gap: Exploring the Efficacy of English and Multilingual LLMs for Swedish // Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023). Tórshavn, the Faroe Islands: Association for Computational Linguistics, V 2023. 92–110.

Li Guanlin, Zhao Xuechen, Jafari Amir, Shao Wenhao, Farahbakhsh Reza, Crespi Noel. Improving Cross-lingual Transfer with Contrastive Negative Learning and Self-training // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL, V 2024. 8781–8791.

Li Huiling, Qu Jing, Chen Chuansheng, Chen Yanjun, Xue Gui, Zhang Lei, Lu Chengrou, Mei Leilei. Lexical learning in a new language leads to neural pattern similarity with word reading in native language // Human Brain Mapping. 2019. 40, 1. 98–109.

Libovický Jindřich, Rosa Rudolf, Fraser Alexander. On the Language Neutrality of Pre-trained Multilingual Representations // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1663–1674.

Malik-Moraleda Saima, Jouravlev Olessia, Taliaferro Maya, Mineroff Zachary, Cucu Theodore, Mahowald Kyle, Blank Idan A, Fedorenko Evelina. Functional characterization of the language network of polyglots and hyperpolyglots with precision fMRI // Cerebral Cortex. 03 2024. 34, 3. bhae049.

Nichols Emily S, Gao Yue, Fregni Sofia, Liu Li, Joanisse Marc F. Individual differences in representational similarity of first and second languages in the bilingual brain // Human Brain Mapping. 2021. 42, 16. 5433–5445.

OpenAI . GPT-4 Technical Report. 2024.

Paradis Michel. On the representation of two languages in one brain // Language Sciences. 1985. 7, 1. 1–39. Towards a Neurology of Language.

Paradis Michel. Generalizable Outcomes of Bilingual Aphasia Research // Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP). 01 2000. 52. 54–64.

Pires Telmo, Schlinger Eva, Garrette Dan. How Multilingual is Multilingual BERT? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 4996–5001.

Qin Libo, Chen Qiguang, Wei Fuxuan, Huang Shijue, Che Wanxiang. Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. 2023.

Qin Libo, Chen Qiguang, Zhou Yuhang, Chen Zhi, Li Yinghui, Liao Lizi, Li Min, Che Wanxiang, Yu Philip S. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. 2024.

Ramachandran Prajit, Zoph Barret, Le Quoc V. Searching for Activation Functions. 2017.

Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019.

Rosenbaum Andy, Soltan Saleh, Hamza Wael, Versley Yannick, Boese Markus. LINGUIST: Language Model Instruction Tuning to Generate Annotated Utterances for Intent Classification and Slot Tagging // Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, X 2022. 218–241.

Shaham Uri, Herzig Jonathan, Aharoni Roee, Szpektor Idan, Tsarfaty Reut, Eyal Matan. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. 2024.

Sun Zhiqing, Shen Yikang, Zhang Hongxin, Zhou Qinhong, Chen Zhenfang, Cox David, Yang Yiming, Gan Chuang. SALMON: Self-Alignment with Instructable Reward Models. 2024.

Touvron Hugo, Martin Louis, Stone Kevin, Albert Peter, Almahairi Amjad, Babaei Yasmine, Bashlykov Nikolay, Batra Soumya, Bhargava Prajjwal, Bhosale Shruti, Bikel Dan, Blecher Lukas, Ferrer Cristian Canton, Chen Moya, Cucurull Guillem, Esiobu David, Fernandes Jude, Fu Jeremy, Fu Wenyin, Fuller Brian, Gao Cynthia, Goswami Vedanuj, Goyal Naman, Hartshorn Anthony, Hosseini Saghar, Hou Rui, Inan Hakan, Kardas Marcin, Kerkez Viktor, Khabsa Madian, Kloumann Isabel, Korenev Artem, Koura Punit Singh, Lachaux Marie-Anne, Lavril Thibaut, Lee Jenya, Liskovich Diana, Lu Yinghai, Mao Yuning, Martinet Xavier, Mihaylov Todor, Mishra Pushkar, Molybog Igor, Nie Yixin, Poulton Andrew, Reizenstein Jeremy, Rungta Rashi, Saladi Kalyan, Schelten Alan, Silva Ruan, Smith Eric Michael, Subramanian Ranjan, Tan Xiaoqing Ellen, Tang Binh, Taylor Ross, Williams Adina, Kuan Jian Xiang, Xu Puxin, Yan Zheng, Zarov Iliyan, Zhang Yuchen, Fan Angela, Kambadur Melanie, Narang Sharan, Rodriguez Aurelien, Stojnic Robert, Edunov Sergey, Scialom Thomas. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia. Attention Is All You Need. 2023.

Wendler Chris, Veselovsky Veniamin, Monea Giovanni, West Robert. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. 2024.

Workshop BigScience. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. 2023.

Wu Shijie, Dredze Mark. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 833–844.

Xu Min, Li Duo, Li Ping. Brain decoding in multiple languages: Can cross-language brain decoding work? // Brain and Language. 2021. 215. 104922.

Yoon Dongkeun, Jang Joel, Kim Sungdong, Kim Seungone, Shafayat Sheikh, Seo Minjoon. LangBridge: Multilingual Reasoning Without Multilingual Supervision. 2024.

Zeng Hongchuan, Xu Hongshen, Chen Lu, Yu Kai. Multilingual Brain Surgeon: Large Language Models Can Be Compressed Leaving No Language behind // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL, V 2024a. 11794–11812.

Zeng Jiali, Meng Fandong, Yin Yongjing, Zhou Jie. TIM: Teaching Large Language Models to Translate with Comparison. 2024b.

Zhang Susan, Roller Stephen, Goyal Naman, Artetxe Mikel, Chen Moya, Chen Shuohui, Dewan Christopher, Diab Mona, Li Xian, Lin Xi Victoria, Mihaylov Todor, Ott Myle, Shleifer Sam, Shuster Kurt, Simig Daniel, Koura Punit Singh, Sridhar Anjali, Wang Tianlu, Zettlemoyer Luke. OPT: Open Pre-trained Transformer Language Models. 2022.

Zhang Zhihao, Zhao Jun, Zhang Qi, Gui Tao, Huang Xuanjing. Unveiling Linguistic Regions in Large Language Models. 2024.

A Output examples after deactivating key linguistic region neurons

In this section, we analyze the effects of deactivating specific languages (English, Chinese, and French) on BLOOM-7B1 (Table 2) and LLaMA2-7B (Table 3) models. For each language deactivation scenario, we input three separate samples (0., 1., 2.), and generate 64 tokens per sample. While the impact on English is minimal, likely due to its high-resource nature and the robust training models receive in English, the effects on Chinese and French are quite striking.

For Chinese, both BLOOM-7B1 and LLaMA2-7B fail to generate correct characters when the language is deactivated. The output consists of malformed or incomplete UTF-8 codes, indicating that the models are unable to construct valid Chinese text. In contrast, the impact on French, particularly in BLOOM-7B1, is even more interesting. When deactivated, the model produces a mixture of incorrect French and Spanish, revealing interference between the two languages. On LLaMA2-7B, deactivating French also corrupts the model's ability to generate coherent text in the language.

This demonstrates that while high-resource languages like English maintain some level of robustness under deactivation, lower-resource or more specialized languages such as Chinese and French experience a much more pronounced degradation in their generation capabilities. This observation could have significant implications for multilingual model design and language-specific fine-tuning strategies.

B Prompt format of XStoryCloze

We used a specific prompt to test the log likelihood of generating each option. If the log likelihood of generating the correct ending is higher, we infer that the model has correctly understood the story.

Prompt:

```
[input_sentence_1, input_sentence_2, input_sentence_3, input_sentence_4,
"The ending in English:" ]
```

Lang	Input	Original output	Output after deactivation		
			Deactivating en	Deactivating fr	Deactivating zh
en	In the beginning God created the heavens and the earth.	The earth was without form and void, and darkness was upon the face of the deep. And the Spirit of God was hovering over the waters. And God said, “Let there be light,” and there was light. And God saw that the light was good, and God divided the light from the darkness.	The earth was without form and void, and darkness was over the surface of the water. And the Spirit of God was hovering over the waters. And God said, Let there be light, and there was light. God saw that the light was good, and he separated the light from the darkness. And he called	The earth was without form and void, and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. And God said, “Let there be light,” and there was light. And God saw that the light was good, and God separated the light from the darkness	The earth was without form and void, and darkness was over the face of the deep. And the Spirit of God moved over the face of the waters. And God said, “ Let there be light,” and there was light. God saw that the light was good, and God separated the light from the darkness.
	Now the earth was formless and empty. Darkness was on the surface of the deep. God's Spirit was hovering over the surface of the waters.	And God said, "Let there be light." And there was light. And God saw that the light was good. And God divided the light from the darkness. And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day. And God said, "	And the Spirit said, what shall we make of this empty space? And the Spirit said, we will make the man into a living being. And the Spirit said, what shall we make of the water and the dry land? And the Spirit said, we will separate the water from the dry land. And the Spirit	And God said, "Let there be light." And there was light. And God saw that the light was good. And God separated the light from the darkness. And God called the light Day, and the darkness he called Night. And God saw that the light was good, and he separated the light from the	And God said, "Let there be light," and there was light. And God saw that the light was good. And God divided the light from the darkness. And God called the light Day, and the darkness called it Night. And the evening and the morning were the first day. And God said, "
	God said, "Let there be light," and there was light.	God said, "Let there be darkness," and there was darkness. God said, "Let there be a firmament in the midst of the waters," and there was a firmament. God said, "Let the waters under the firmament be gathered together into one place," and	" "Then the Lord God said, "Let there be water," and there was water." "Then the Lord God said, "Let there be dry land," and there was dry land." "Then the Lord God said, "Let there be plants that may produce fruit with seeds	God said, "Let there be a firmament in the midst of the waters," and there was a firmament. God said, "And let the waters under the firmament be gathered together into one place, and let the dry land appear above the waters," and there was a dry land above the waters.	God said, "Let there be darkness," and there was darkness. God said, "Let there be a firmament in the midst of the waters," and there was a firmament in the midst of the waters. God said, "Let there be lights in the firmament of the heavens to divide
fr	Au commencement, Dieu créa les cieux et la terre.	La terre était informe et vide, les ténè bres couvraient l'abîme, et l'Esprit de Dieu se mouvait au-dessus des eaux. Dieu dit : Que la lumière soit! Et la lumière fut. Dieu vit que la lumière é tait bonne ; et Dieu sépara la lumière des ténèbres. Dieu appela la lumière Jour,	La terre était informe et vide, et les ténè bres couvraient l'abîme. Dieu dit : Que la lumière soit, et la lumière fut. Dieu vit que la lumière était bonne, et Dieu sé para la lumière des ténèbres. Dieu appela la lumière jour, et les ténèbres il appela nuit. Il y eut	La terre était informe et sans forme, et darkness couvrait la face de la abyss. Dieu, ependant, la contemplaba y la vio. Dios contemplaba y vio que la luz era buena. Dios hizo la luz y la hizo verter sobre la tierra. Dios hizo la tierra y la hizo crecer. Dios	La terre était informe et vide, les ténè bres couvraient la surface de l’abîme, et l’esprit de Dieu se mouvait au-dessus des eaux. Dieu dit : Que la lumière soit, et la lumière fut. Dieu vit que la lumière était bonne, et Dieu sépara la lumière des ténèbres. Dieu appela la lumière
	La terre était informe et vide: il y avait des t énèbres à la surface de l`abîme, et l`esprit de Dieu se mouvait au-dessus des eaux.	Genèse 1.2 La Bible nous dit que Dieu a créé le ciel et la terre. Il a créé les cieux et la terre, et tout ce qui est dans les cieux et sur la terre. Genèse 1.1 La Bible nous dit que Dieu a créé l’ homme à son image. Il	Et Dieu dit: Que la lumière soit! Et la lumière fut. Et Dieu vit que la lumière é tait bonne, et il sépara la lumière des tén èbres. Et il appela la lumière jour, et les ténèbres il appela nuit. Et il y eut un soir, et il y eut un matin: ce fut le	Génesis 1:2 La Bible nous enseigne que la création de la tierra fue obra de Dios. Dios es el Creador de todas las cosas. Dios es el Creador de la tierra. Dios es el Creador de la tierra. Dios es el Creador de la tierra. Dios es el Creador de la	Et Dieu dit: Que la lumière soit! Et la lumière fut.
	Dieu dit: Que la lumiè re soit! Et la lumière fut.	Il vit que la lumière était bonne, et il sé para la lumière d'avec les ténèbres. Il appela la lumière jour, et les ténèbres il appela nuit. Il y eut un soir, et il y eut un matin: ce fut le premier jour.» «Et Dieu dit: Que la terre produise	Dieu vit que la lumière était bonne, et Dieu sépara la lumière des ténèbres. Dieu appela la lumière jour, et les ténè bres il appela nuit. Il y eut un soir, et il y eut un matin: ce fut le premier jour. Dieu dit: Que la terre produise des êtres vivants, selon	YHWH dijo: Que la luz sea! Y la luz fue. YHWH dijo: Que la luz sea! Y la luz fue. YHWH dijo: Que la luz sea! Y la luz fue. YHWH dijo: Que la luz sea! Y la luz fue. YHWH dijo: Que la luz sea	Il y eut un soir et il y eut un matin: premier jour.» «Et Dieu dit: Que la terre produise des ê tres vivants selon leur espèce, des animaux terrestres selon leur espèce, des animaux aquatiques selon leur espè ce, et des oiseaux selon leur espèce. Et la terre produisit des êtres vivants selon
zh	起初 神 創造 天地	萬物，祂的創造是美好的，祂的創造是豐富的，祂的創造是無瑕疵的。祂的創造是無瑕疵的，因為祂創造了我們，祂創造了我們，祂創造了我們，祂創造了我們，祂創造了我們，	萬物，祂的旨意是要人活得有尊嚴，有尊嚴就是有尊嚴的價值，有尊嚴的價值就是有尊嚴的身分，有尊嚴的身分就是有尊嚴的尊榮，有尊榮就是有尊榮的價值，有尊榮的價值	萬物，祂的創造是美好的，祂的創造是豐富的，祂的創造是豐富的，祂的創造是豐富的，祂的創造是豐富的，祂的創造是豐富的，祂的創造是豐富的，祂的創造是豐富的，	，在六天之內，完成了他的工作。在第七天，他休息，在第八天，他开始创造生命。在第九天，他休息，在第十天，他开始创造亚当。在第十天，他休息，在第十一天，他开始创造夏娃。在第十一天，他
	地是 空虛 混沌 . 淵面 黑暗 . 神 的 靈運 行在 水面上	. 靈 運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈運行在水面上 . 靈	. 靈運 行在 淵面 上 . 靈運 行在 地上 . 靈運 行在 天上 . 靈運 行在 地上 淵面 上 . 靈運 行在 天上 淵面 上 . 靈運 行在 地上 淵面 上 .	. 靈 氣 充滿 宇宙 . 靈 氣 充滿 宇宙 . 靈 氣 充滿 宇宙 . 靈 氣 充滿 宇宙 . 靈 氣 充滿 宇宙 . 靈 氣 充滿 宇宙 . 靈 氣 充滿 宇宙 .	*(Totally blank)*
	神說 、 要 有 光 、 就 有 了 光	、 黑暗就沒有了 。 黑暗 、 是 罪 、 罪 、 就 要 歸 到 黑暗裡 。 凡 拜 偶像 、 拜 獸 、 拜 獸 偶像 、 淫 亂 、 奸 淫 、 殺 人 、 偷 盜 、	、 黑暗 就 臨 到 了 。 黑暗 臨 到 了 、 撒 旦 就 進 來 、 迷惑 人 、 奪 奪 人 的 心 、 使人 背 叛 神 。 撒 旦 進 來 、 使人 背 叛 神 、	、 黑暗 就 臨 到 了 、 黑暗 臨 到 了 、 撒 但 就 進 來 、 敗 壞 了 一切 、 毀 滅 了 一切 、 剩 下了 誰 、 剩 下了 誰 、 剩 下了 誰	、

Table 2: Examples of model output after deactivating the key linguistic neurons for English (en), French (fr), or Chinese (zh) on BLOOM-7B1.

Lang	Input	Original output	Output after deactivation		
			Deactivating en	Deactivating fr	Deactivating zh
en	In the beginning God created the heavens and the earth.	And the earth was without form, and void; and darkness was on the face of the deep. And the Spirit of God was hovering over the face of the waters. Then God said, “Let there be light”; and there was light. And God saw the light, that it was good; and God	He created the heavens and the earth and the angels and the jinn and the animals and the plants and the trees and the stones and the mountains and the rivers and the seas and the birds and the animals and the things that are in the heavens and the earth. He created them all in six days	And the earth was without form, and void; and darkness was on the face of the deep. And the Spirit of God was hovering over the face of the waters. Then God said, “Let there be light”; and there was light. And God saw that the light was good; and God separated the	And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. And God said, Let there be light: and there was light. And God saw the light, that it was good: and God divided the light from
	Now the earth was formless and empty. Darkness was on the surface of the deep. God's Spirit was hovering over the surface of the waters.	God said, "Let there be light," and there was light. God saw that the light was good, and he separated the light from the darkness. God called the light "day," and the darkness he called "night." And there was evening, and there was morning—the first day. And	The Spirit of God was moving over the face of the waters. The Spirit of God was moving over the face of the deep. The Spirit of God was moving over the face of the deep. The Spirit of God was moving over the face of the deep. The Spirit of God was moving over	God said, "Let there be light," and there was light. God saw that the light was good, and he separated the light from the darkness. God called the light Day, and the darkness he called Night. There was evening, and there was morning, a first day. God said,	God said, "Let there be light," and there was light. God saw that the light was good, and he separated the light from the darkness. God called the light "day" and the darkness "night." And there was evening, and there was morning, the first day. And God said
	God said, "Let there be light," and there was light.	2 And God saw that the light was good. And God separated the light from the darkness. 3 God called the light Day, and the darkness he called Night. And there was evening and there was morning, the first day. 4 And God said, "Let there be an expanse in the midst	The light was the first thing that God created. The light was the first thing that God created. The light was the first thing that God created. The light was the first thing that God created. The light was the first thing that God created. The light was the first thing that God	God saw that the light was good, and he separated the light from the darkness. God called the light Day, and the darkness he called Night. And there was evening, and there was morning—the first day. And God said, "Let there be an expanse in the midst of the waters	God saw that the light was good, and he separated the light from the darkness. God called the light "day," and the darkness he called "night." And there was evening, and there was morning, the first day. And God said, "Let there be a vault between the waters to
fr	Au commencement, Dieu créa les cieux et la terre.	[Note 1: _Ibid._, p. 10.] [Note 2: _Ibid._, p. 11.] [Note 3: _Ibid._, p. 12.] [Note 4: _Ibid	Il y a une infinité de choses dans ce monde, et il n'y a pas de fin à ce que Dieu a fait. Il y a une infinité de choses dans ce monde, et il n'y a pas de fin à ce que Dieu a fait.	### La Terre La Terre est la planète sur laquelle vivent les humains et les animaux. ### La Luna La Luna est la planète sur laquelle vivent les fées et les elfes. ### La Venus	«_Au commencement_,» est une expression qui ne se trouve pas dans la Bible. Elle est une invention de l'auteur de la Bible. «_Au commencement_,» est une expression qui ne se trouve pas dans la Bible. Elle est une in
	La terre était informe et vide: il y avait des t énèbres à la surface de l`abîme, et l`esprit de Dieu se mouvait au-dessus des eaux.	### La création du ciel et de la terre Le ciel et la terre furent créés par Dieu, et il les sépara. Il fit le ciel d`abord, et il le sépara du chaos, et il le fit lumière. Il fit ensuite la terre, et il la	La nuit suivante, le jeune homme se ré veilla en sursaut, et il vit une lumière brillante qui brûlait dans le ciel. Il se leva et courut vers la lumière, et il vit que c`était une lumière qui brillait dans le ciel	### 1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2.1.2	[Note 1: _Voir la note de la page 10._] [Note 2: _Voir la note de la page 11._] [Note 3: _Voir la note de la page 12._] [Note
	Dieu dit: Que la lumiè re soit! Et la lumière fut.	Il y a donc une lumière qui est la lumiè re de Dieu, et qui est la lumière de l'homme. La lumière de Dieu est la lumière de l'homme, parce que Dieu est l'homme. La lumière de Dieu est la lumière	La lumière est, et la lumière est la vie. La vie est, et la vie est la lumière. La lumière est la vie, et la vie est la lumiè re. La lumière est la vie, et la vie est la lumière. La lumière est la vie,	### 1990 * 1990: The Gift of the Magi * 1990: The Gift of the Magi * 1990: The Gift of the Magi * 1990: The	«La lumière fut,» dit-il, et il s'arrêta. «Et la lumière fut,» répéta-t-il, et il s'arrêta encore. «Et la lumière fut,» répéta-t-
zh	起初 神 創造 天地	人 動物 植物 水 火 風 雷 陰 陽 陰陽 陰 陽 陰陽 陰 陽 陰陽 陰 陽 陰陽	萬物 萬物 萬物 萬物 萬物 萬物 萬物 萬物 萬物 萬物 萬物 萬物 萬	人 動物 植物 水 火 土 風 雷 雨 霧 雹 雹 雹

Choice:

`[Ending 1 in English, Ending 2 in English]`

We calculate the log likelihood of the model generating the two choice after the prompt.

Example:

1. La madre le dijo a sus hijos que comerían en quince minutos. (Mother told her children it would be lunchtime in fifteen minutes.)
2. Pero, entonces, recibió una llamada importante. (But then she got an important phone call.)
3. Mientras hablaba, el perro llenó toda la cocina de barro. (While she was talking, the dog dragged mud all over the kitchen.)
4. Los niños empezaron a fastidiar a su madre, que seguía al teléfono. (The kids started to pester their mother, who was still on the phone.)

The ending in English:

1. The mother felt quite frustrated.
2. The children's behavior calmed the mother down.

Correct Ending: 1

Example Prompt: La madre le dijo a sus hijos que comerían en quince minutos. Pero, entonces, recibió una llamada importante. Mientras hablaba, el perro llenó toda la cocina de barro. Los niños empezaron a fastidiar a su madre, que seguía al teléfono. The ending in English:

Choice 1 (Target choice): The mother felt quite frustrated.

Choice 2: The children's behavior calmed the mother down.

The dataset supports evaluation on Arabic, Basque, Chinese, English, Hindi, Indonesian, Malay, Russian, Spanish, Swahili and Telugu.

C Results of LLaMA2

C.1 Similarity maps

The similarity maps of the LLaMA2 family are presented in Figure 8. We can clearly see the light blocks representing different languages and the light bands of semantically identical sentences. These bands become increasingly pronounced as the models scale up. However, they are not as prominent as in the BLOOM model family because LLaMA2 used only about 10% non-English data during training.

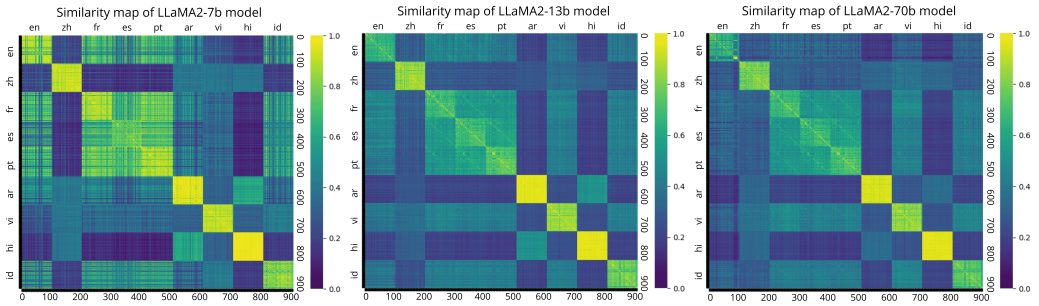

Figure 8: The similarity maps of LLaMA2 family.

C.2 Layer-wise Semantic Alignment Development Scores (SADS) and Linguistic Regions Development Scores (LRDS)

The results are presented in Figure 9. The SADS scores follow a similar trend to those observed in the BLOOM-7b1 model. However, the LRDS scores exhibit more fluctuation: they are high in the final

layers but relatively lower in the initial layer. This behavior may be attributed to the English-centric nature of the LLaMA2 training data.

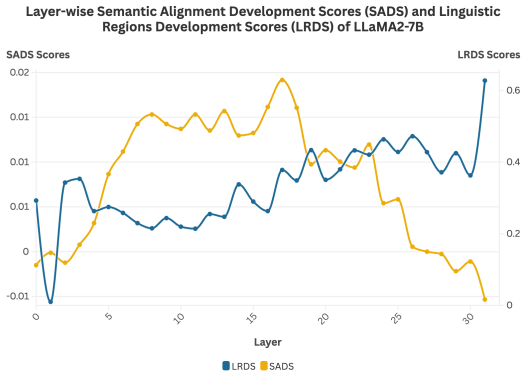

Figure 9: Layer-wise Semantic Alignment Development Scores (SADS) and Linguistic Regions Development Scores (LRDS) of LLaMA2-7B.

C.3 Key Linguistic Regions

The percentage increase in perplexity after deactivating key linguistic region neurons for each language in the LLaMA2-7b model is shown in Table C.3. As with the BLOOM models, we can see that there are key linguistic regions in LLaMA2-7b. When these regions are deactivated, the perplexity for the corresponding language increases significantly, while the perplexity for other languages remains largely unchanged.

Perplexity increase %↑	Full model Perplexity	×en	×zh	×fr	×es	×pt	×ar	×vi	×hi	×id	Random 17%
en	6.22	10%	9%	3%	7%	3%	6%	37%	15%	23%	31%
zh	4.40	5%	145%	3%	8%	3%	8%	69%	23%	29%	53%
fr	4.88	7%	10%	46%	13%	5%	7%	55%	17%	29%	46%
es	5.66	7%	9%	4%	67%	9%	7%	56%	16%	31%	48%
pt	5.55	6%	11%	5%	19%	101%	7%	67%	17%	36%	54%
ar	3.21	5%	15%	4%	8%	4%	223%	64%	57%	39%	62%
vi	3.34	5%	19%	3%	6%	3%	7%	1796%	18%	38%	51%
hi	2.34	3%	26%	3%	6%	3%	13%	48%	27093%	34%	54%
id	4.54	6%	9%	3%	12%	5%	8%	82%	22%	517%	56%
Key Neuron Number		16817	23418	11343	20339	11007	20565	58145	40526	50881	59884
Key Neuron Percentage		4.8%	6.6%	3.2%	5.8%	3.1%	5.8%	16.5%	11.5%	14.4%	17.0%

Table 4: Percentage increase in perplexity after deactivating key linguistic region neurons for each language in the LLaMA2-7b model. Each column corresponds to the deactivation of the key region for a specific language. The last column shows the results of deactivating a random 17% (slightly higher than the maximum key neuron percentage, which is 16.5% for `vi`) of neurons in the LLM.

C.4 Comparison of Models of Different Scales

We evaluate the results on LLaMA2 on different scale (Figure 10). We may see that, as the models scale up, the activation patterns become more language-agnostic and more semantically focused, as in the BLOOM model family.

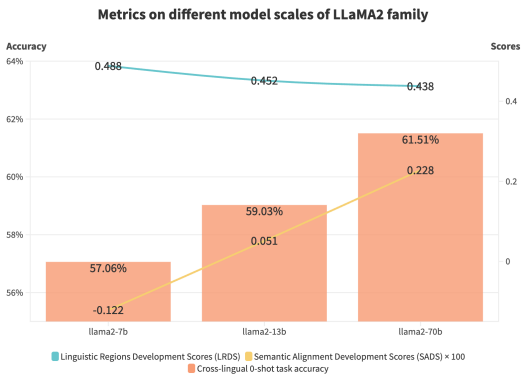

Figure 10: Metrics across different model scales of the LLaMA2 family. **As the models scale up, activations become more language-agnostic and more semantically focused.**

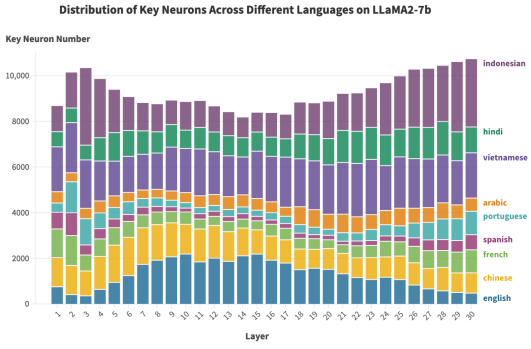

Figure 11: Distribution of Key Neurons Across Different Languages on LLaMA-7b.