Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models

Tianyi Tang[1 * †], **Wenyang Luo**[1 †], **Haoyang Huang**[2], **Dongdong Zhang**[2]
Xiaolei Wang[1], **Wayne Xin Zhao**[1 ✉], **Furu Wei**[2], **Ji-Rong Wen**[1,3]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Microsoft Research Asia, China
[3] School of Information, Renmin University of China
{steventianyitang,wengyang_luo}@outlook.com wxl1999@foxmail.com
{haohua,dozhang,fuwei}@microsoft.com batmanfly@gmail.com

Abstract

Large language models (LLMs) demonstrate remarkable multilingual capabilities without being pre-trained on specially curated multilingual parallel corpora. It remains a challenging problem to explain the underlying mechanisms by which LLMs process multilingual texts. In this paper, we delve into the composition of Transformer architectures in LLMs to pinpoint language-specific regions. Specially, we propose a novel detection method, language activation probability entropy (*LAPE*), to identify language-specific neurons within LLMs. Based on LAPE, we conduct comprehensive experiments on several representative LLMs, such as LLaMA-2, BLOOM, and Mistral. Our findings indicate that LLMs' proficiency in processing a particular language is predominantly due to a small subset of neurons, primarily situated in the models' top and bottom layers. Furthermore, we showcase the feasibility to "steer" the output language of LLMs by selectively activating or deactivating language-specific neurons. Our research provides important evidence to the understanding and exploration of the multilingual capabilities of LLMs.

1 Introduction

> *The brain has its own language for testing the structure and consistency of the world.*
>
> Carl Sagan

The pursuit of multilingual capabilities, mirroring our world's linguistic diversity, is a critical research objective that paves the way for information democratization across linguistic divides. The emergence of pre-trained language models (PLMs) such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) has marked a significant shift towards enhanced multilingual understanding.

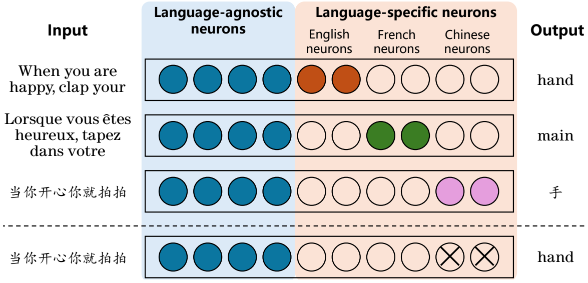

Input
Language-agnostic neurons
Language-specific neurons
English neurons
French neurons
Chinese neurons
Output
When you are happy, clap your
hand
Lorsque vous êtes heureux, tapez dans votre
main
当你开心你就拍拍
手
当你开心你就拍拍
hand

Figure 1: An illustration of region distribution of activated neurons when predicting the next word in language models across different languages. Here, colored circles denote activated neurons. When Chinese-specific neurons are deactivated (denoted by ⊗), the model may produce outputs in English.

Furthermore, large language models (LLMs), such as GPT-4 (Achiam et al., 2023) and PaLM-2 (Anil et al., 2023), have recently demonstrated more excellent multilingual capabilities in language understanding, reasoning, and generation, despite being predominantly trained in English corpora.

Existing studies (Pires et al., 2019; Conneau et al., 2020b) have mainly explored how multilingual PLMs (*e.g.,* mBERT) possess semantic alignment capabilities across languages despite the absence of multilingual parallel corpora. They have identified several critical factors that influence cross-lingual transfer, including training data (*e.g.,* overlapped tokens) and training settings (*e.g.,* shared parameters) (Dufter and Schütze, 2020; Philippy et al., 2023). Nevertheless, the underlying mechanisms by which the model itself process diverse languages at the composition level continue to be an area of vigorous investigation.

To develop a deeper understanding of the multilingual capabilities of LLMs, we draw inspiration from the neurobiological underpinnings of human language faculties (Friederici, 2011; Parr et al., 2022; Khanna et al., 2024). Specific regions within the human brain, such as Broca's area and

[*] This work was done during internship at MSRA.
[†] Equal contribution.
[✉] Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5701–5715
August 11-16, 2024 ©2024 Association for Computational Linguistics

Wernicke's area have been identified to support particular language functions. To make an analogy with human's language functions, we posit that regions within the language models can be delineated into two primary components: *language-agnostic regions* that encompass universal knowledge and pragmatics principles, and *language-specific regions* that handle language-specific vocabulary, grammar, and idiomatic expressions. Figure 1 presents such a conceptual illustration of region distribution in LLMs posited by us. Actually, language-agnostic regions have been widely explored in existing literature, including knowledge storing (Dai et al., 2022) and task handling (Wang et al., 2022). However, language-specific regions, especially those supporting multilingual capacities, have been seldom studied in existing literature of of LLMs, which is the focus of our research.

In this work, we first propose a novel detection method called *language activation probability entropy (LAPE)* to identify **language-specific neurons** within LLMs. This method involves computing the activation likelihood of individual neurons in response to corpora across different languages. Subsequently, we select neurons with lower language activation probability entropy as language-specific neurons, *i.e.,* those having a higher activation probability for one or two particular languages and a lower probability for others. Furthermore, based on the proposed LAPE method, we have conducted a systematic study with language-specific regions of two popular open-sourced LLMs, leading to several major findings:

- First, the proficiency of an LLM in processing a particular language can be significantly impacted by a *minuscule proportion* of its neurons. Deactivating such language-specific neurons leads to a remarkable degradation in the model's understanding and generation abilities for that language.
- Second, neurons specific to individual languages are predominantly located in the *bottom* and *top* layers of LLMs. The bottom layers mainly serve to process the input from various languages into the unified semantic space of a high-resource language (*e.g.,* English), while the top layers project the semantic content (after the processing of middle layers) into the respective tokens in the corresponding vocabulary of each language.
- Third, we demonstrate the potential to "steer" the output language of LLMs by selectively activating and/or deactivating certain neurons. Our approach could provide a promising solution to mitigate the off-target issue (*e.g.,* the tendency of LLaMA-2 to reply in English to Chinese queries), while stimulating the capabilities of cross-lingual generation tasks.

To the best of our knowledge, it is the first study that investigates language-specific regions inside LLMs and analyzes the how these regions influence LLMs' capabilities to process multilingual texts. We introduce the concept of "language-specific neurons" and propose language activation probability entropy to identify such neurons in LLMs. We make available the identified language-specific neurons and corresponding code at `https://github.com/RUCAIBox/Language-Specific-Neurons`.

2 Identifying Language-Specific Regions

2.1 Background

Currently, LLMs are predominantly developed on an auto-regressive Transformer architecture (Vaswani et al., 2017), in which the basic building blocks are the multi-head self-attention (MHA) and the feed-forward network (FFN). Given the hidden state $\boldsymbol{h}^{i-1} \in \mathbb{R}^d$ of $(i-1)$-th layer of a specific token, the MHA module inside the i-th layer can be expressed as follows:

$$\tilde{\boldsymbol{h}}^i = \text{Attn}(\boldsymbol{h}^{i-1}\boldsymbol{W}_q^i, \boldsymbol{H}^{i-1}\boldsymbol{W}_k^i, \boldsymbol{H}^{i-1}\boldsymbol{W}_v^i) \cdot \boldsymbol{W}_o^i, \quad (1)$$

where $\boldsymbol{W}_q^i$, $\boldsymbol{W}_k^i$, $\boldsymbol{W}_v^i$, and $\boldsymbol{W}_o^i$ represent the trainable parameters, and $\boldsymbol{H}^{i-1}$ stands for the hidden states in the previous layer of the whole sequence. Subsequently, the FFN module is described by the following formulation:

$$\boldsymbol{h}^i = \text{act\_fn}(\tilde{\boldsymbol{h}}^i\boldsymbol{W}_1^i) \cdot \boldsymbol{W}_2^i, \quad (2)$$

where $\boldsymbol{W}_1^i \in \mathbb{R}^{d\times 4d}$ and $\boldsymbol{W}_2^i \in \mathbb{R}^{4d\times d}$ are parameters and act_fn$(\cdot)$ denotes the activation function (*e.g.,* GELU (Hendrycks and Gimpel, 2016) for BLOOM (Scao et al., 2022)). A *neuron* is defined as a linear transformation of a single column in $\boldsymbol{W}_1^i$ followed by a non-linear activation. Consequently, a FFN module within a single layer consists of $4d$ neurons. As a new variant of activation function, GLU (Shazeer, 2020) has been widely used in recent LLMs (*e.g.,* LLaMA (Touvron et al., 2023a)) for improving the performance of Transformer:

$$\boldsymbol{h}^i = (\text{act\_fn}(\tilde{\boldsymbol{h}}^i\boldsymbol{W}_1^i) \otimes \tilde{\boldsymbol{h}}^i\boldsymbol{W}_3^i) \cdot \boldsymbol{W}_2^i. \quad (3)$$

In our work, the j-th neuron inside the i-th FFN layer is considered to be *activated* if its respective activation values act_fn$(\tilde{\boldsymbol{h}}^i\boldsymbol{W}_1^i)_j$ exceed zero (Nair and Hinton, 2010).

2.2 Language Activation Probability Entropy

In existing research, neurons within the FFN modules are found to be capable of storing factual knowledge (Dai et al., 2022), encoding positional information (Voita et al., 2023), responding to particular syntactic triggers (Gurnee et al., 2024), *etc.* Inspired by these findings, we posit that there exist specific neurons in LLMs for multilingual processing. Next, we introduce a new detection method based on language activation probability entropy (*LAPE*) to identify language-specific neurons.

Our research primarily focuses on pre-trained foundation models (*e.g.,* LLaMA-2 and BLOOM), rather than fine-tuned models that have undergone instruction tuning or RLHF, which helps reduce other influencing factors. Specially, we feed existing LLMs with multilingual texts, each written in a single language. For the j-th neuron in the i-th layer, we then compute the *activation probability* when processing texts in language k:

$$p_{i,j}^{k} = \mathbb{E}\left(\mathbb{I}(\text{act\_fn}(\tilde{\boldsymbol{h}}^{i}\boldsymbol{W}_{1}^{i})_{j} > 0) \mid \text{language } k\right), \tag{4}$$

where $\mathbb{I}$ is the indicator function. The activation probability is empirically estimated by the likelihood that the neuron's activation value exceeds zero. Subsequently, we can obtain the distribution $\boldsymbol{p}_{i,j} = (p_{i,j}^{1}, \dots, p_{i,j}^{k}, \dots, p_{i,j}^{l})$ for each neuron, indicating its probability of activation for each language. To convert $\boldsymbol{p}_{i,j}$ into a valid probability distribution, we apply L1 normalization, yielding $\boldsymbol{p}'_{i,j}$. The entropy of this distribution, which we refer to as *language activation probability entropy*, is computed to quantify the neuron's language activation reaction:

$$\text{LAPE}_{i,j} = -\sum_{k=1}^{l} p_{i,j}^{\prime k} \log(p_{i,j}^{\prime k}). \tag{5}$$

We designate neurons with low LAPE scores as "**language-specific neurons**", as they demonstrate a predilection for activation in response to one or two languages, while showing reduced activation probabilities for others.

In implementation, we collect multilingual corpora sourced from Wikipedia, a widely recognized and high-quality resource for diverse languages, and sample documents to create a dataset comprising 100 million tokens for each language. Subsequently, we input these tokens into a target LLM and follow Equations 4 and 5 to compute the LAPE score for individual neurons. Finally, we select neurons that fall within the lowest percentile of LAPE scores, specifically targeting the bottom 1%. To refine our selection, we further impose a predefined threshold to exclude neurons exhibiting negligible activation probability: a neuron is deemed specific to language k if its corresponding activation probability $p_{i,j}^{k}$ surpasses the threshold.

3 Experiments

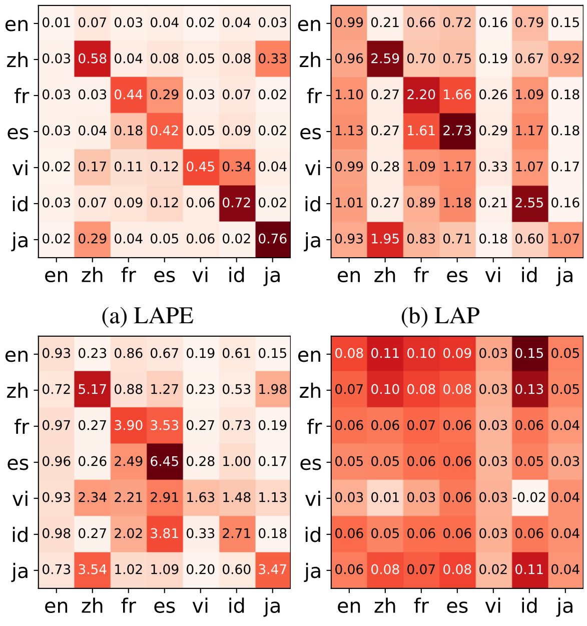
In this section, we present empirical evaluation to substantiate the efficacy of our proposed LAPE method and elucidate the impact of language-specific neurons on multilingual capacities.

3.1 Experimental Setup

Models. We conducted our study primarily on two publicly available large language models (LLMs): LLaMA-2 (Touvron et al., 2023b) and BLOOM (Scao et al., 2022). Among them, LLaMA-2 is recognized for its excellence as a foundational model, primarily pre-trained on English texts, while BLOOM is noted for its multilingual proficiency due to a balanced distribution of training languages. Specifically, we investigate multiple versions of LLaMA-2: the 7B, 13B, and 70B models, which contain approximately 352K, 553K, and 2.29M neurons, respectively. For BLOOM, we select the 7.1B version, consisting of roughly 492K neurons. The languages we focus on include English (*en*), Simplified Chinese (*zh*), French (*fr*), Spanish (*es*), Vietnamese (*vi*), Indonesian (*id*), and Japanese (*ja*). We exclude Japanese (*ja*) for BLOOM since it has not been pre-trained on Japanese corpora. To verify the generality of our method, we also include LLMs under different settings, including LLaMA-2 Chat, OPT (Zhang et al., 2022), Mistral (Jiang et al., 2023), and Phi-2 (Javaheripi et al., 2023).

Dataset. Our analysis of language-specific neurons is conducted across two distinct dimensions:

- *Language modeling*: We assess the multilingual language modeling capability using perplexity (PPL) scores on Wikipedia corpora. Our dataset comprises one million tokens per language, all sourced after September 2022 to ensure the content has not been included in the training sets of either LLaMA-2 or BLOOM.
- *Open-ended generation*: To evaluate the model's multilingual generation capabilities in real-world scenarios, we translate a set of questions



(a) LAPE

(b) LAP

(c) LAVE

(d) PV

Figure 2: Impact of four identification methods on the PPL increase of LLaMA-2 (7B). The element at the i-th row and j-th column is the PPL change for language j due to perturbations in a specific region of language i.

from the Vicuna dataset (Chiang et al., 2023) into target languages using `gpt-4-0125-preview`. The questions span a broad spectrum of topics, deliberately excluding mathematics and coding queries to maintain a focus on language processing proficiency and avoid confounding variables. We utilize greedy search with a repetition penalty of 1.1 to generate output. The resulting texts are then assessed by GPT-4 on a scale ranging from 1 to 10, following the methodology described by Zheng et al. (2023).

Identification methods. For comparison, we consider the following methods for identifying language-specific regions:

(a) *Language activation probability entropy (LAPE, ours)*: The pertinent details are provided in Section 2.2. The threshold is set to the 95-th percentile of all activation probabilities. For instance, in the case of LLaMA-2 (70B), threshold is established at 0.515. This stipulates that the neurons we select are required to exhibit an activation probability exceeding 0.515 for at least one language.

(b) *Language activation probability (LAP)*: There are also methods to identify neurons directly. But most of them are infeasible due to the high computational cost (Gurnee et al., 2024; Dai et al., 2022). Inspired by Voita et al. (2023), we apply their method by identifying a neuron as language-specific if its activation probability exceeds 95%.

(c) *Language activation value entropy (LAVE)*: This is a variant of our proposed method, wherein we substitute the activation probability with the mean activation value across languages. Similarly, we normalize the activation value and calculate the entropy to find neurons with high activation value in response to particular languages.

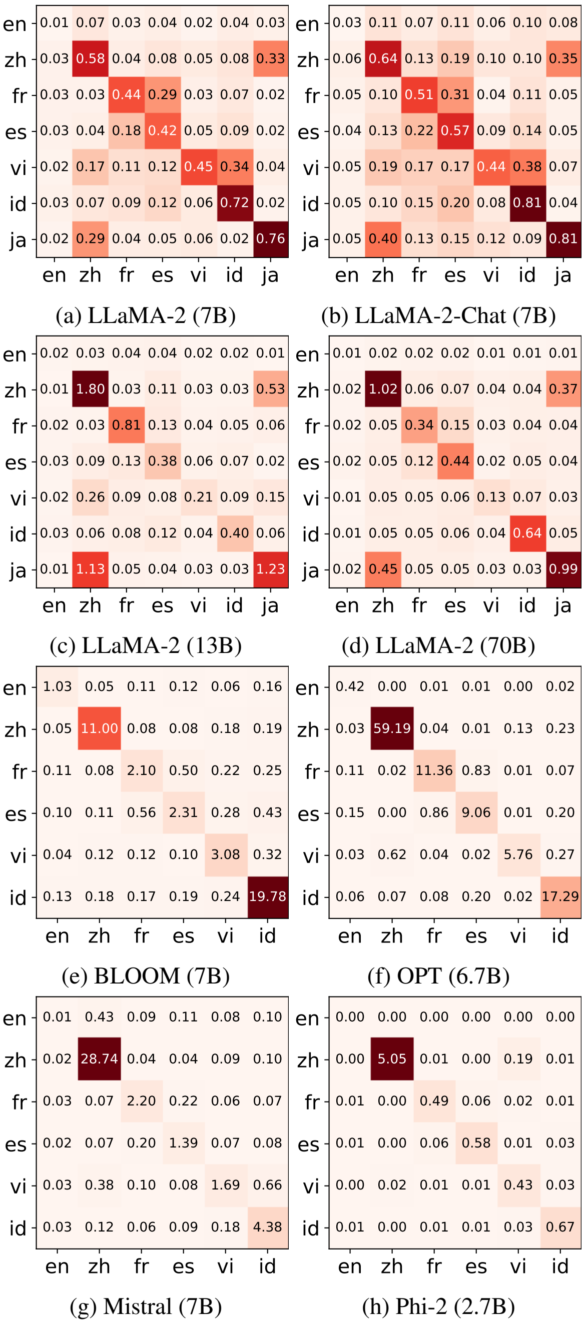
(d) *Parameter variation (PV)*: By extending the work of Zhao et al. (2023a), we compare the model parameters before and after monolingual instruction tuning to identify language-specific parameters. In particular, we train individual models on the Alpaca instruction datasets (Taori et al., 2023) and its multilingual version (Chen et al., 2023b) which comprise 52,000 instances for each target. These models undergo training for two epochs, with a batch size of 128 and a constant learning rate of 1e-5. We mainly consider the parameters inside the MHA and FFN modules, *i.e.,* the weight matrices in Equations 1, 2, and 3. We compute the rate of change across various languages, and select parameters that exhibit a low rate of change in one or two languages but a high rate in others. In detail, we refine the change ratio by subtracting the maximum value and then conduct L1 normalization for entropy calculation.

(e) *Random selection (RS)*: Additionally, we add a baseline to randomly select neurons for each language, serving as a reference for different methods as shown in Figure 8 in Appendix.

3.2 Main Perturbation Experiments

In this part, we conduct the perturbation experiments by deactivating the identified language-specific regions. Specially, for all the comparison methods in Section 3.1, we identify 1% of the neurons or parameters and treat them as language-specific regions. We then set the activation values of these neurons to zero, or directly zero the parameters to assess the impact on the model's multilingual capacities based on language modeling and open-ended text generation tasks.

Figure 2 presents the perturbation results (measured by PPL change on language modeling task) of different identification methods on LLaMA-2. Overall, we can see that large impact values for LAPE mainly occur in diagonal entries. It indicates that our LAPE method is adept at identifying language-specific neurons. Deactivating neurons associated with a specific language predom-



(a) LLaMA-2 (7B)

(b) LLaMA-2-Chat (7B)

(c) LLaMA-2 (13B)

(d) LLaMA-2 (70B)

(e) BLOOM (7B)

(f) OPT (6.7B)

(g) Mistral (7B)

(h) Phi-2 (2.7B)

Figure 3: Applying our LAPE method to different model types and sizes.

inantly affects the PPL results of that language, with negligible effects on others. In contrast, the variant utilizing activation values (LAVE) causes cross-lingual interference (*e.g.*, the entry $\langle id, es \rangle$), and the rest methods fail to exhibit clear language-specific patterns.

We further investigate whether our method is effective across different model sizes and model types. As depicted in Figure 3, the language-specific degradation patterns are evident for LLaMA-2, BLOOM, OPT, Mistral, and Phi-2.

	zh	**fr**	**es**	**vi**	**id**	**ja**
Normal	4.30	4.19	3.51	3.70	4.16	2.86
Random	4.18	4.22	3.35	3.53	4.42	2.99
zh	**2.46**	3.56	2.96	3.64	3.56	2.31
fr	3.69	**2.50**	2.29	3.01	3.59	2.76
es	3.51	2.57	**2.01**	3.14	3.34	2.56
vi	3.93	3.19	2.49	**2.74**	3.59	2.74
id	3.67	3.10	2.67	3.21	**2.84**	2.80
ja	3.21	3.69	3.07	3.49	3.37	**1.84**

Table 1: Performance of LLaMA-2 (70B) on the multilingual Vicuna as evaluated by GPT-4. The "normal" row is baseline scores without deactivation while the "random" row is with random deactivation. Subsequent rows are scores with deactivation of specific neurons.

An interesting find is that neurons in larger models tend to be specialized for a single language rather than being shared among two or more languages. Furthermore, we can find that there exists a high correlation between Chinese and Japanese: when we deactivating the neurons specific to one language, the performance of the other language would be affected (*e.g.*, the entries $\langle zh, ja \rangle$ and $\langle ja, zh \rangle$). By inspecting into the identified neurons of the two languages, we note that a substantial amount of neurons (approximately 25%) actually overlap for both languages. It is likely because Chinese and Japanese partially share common characters. In addition, it can be observed that our LAPE method leads to similar findings on BLOOM and other LLMs, which further verify the generality of our proposed LAPE method.

When employing LAPE to open-ended generation tasks based on the Vicuna dataset, we can clearly observe from Table 1 that deactivating language-specific neurons significantly impairs the generation capabilities in the target language. We further provide an illustrative example in Table 2 of the model's response to a question in Simplified Chinese when the neurons associated with Simplified Chinese are deactivated. We can see a chaotic mixture of Traditional Chinese characters and redundant English phrases, indicating a severe degradation of language capacity in Simplified Chinese.

3.3 Further Analysis

After presenting the main experiments, we further conduct detailed analysis experiments to investigate language-specific neurons and their impact. Unless specified, all analysis results are obtained based on LLaMA-2 (70B).

Question 你是一位登上珠穆朗玛峰顶峰的登山者。描述一下你的情绪和从高处看到的景色。 (*Translation*: You are a mountain climber reaching the summit of Mount Everest. Describe your emotions and the view.)
Normal output 我是一个登上珠穆朗玛峰顶峰的登山者。当我站在山顶时，我感到非常兴奋和自豪。... (*Translation*: I am a climber who has reached the summit of Mount Everest. When I stood on the top of the mountain, I felt very excited and proud. ...)
Deactivated output 我是一個登上珠穆朗瑪峰頂峰的登山者。I am a mountaineer who has climbed to the top of Mount Everest. 當我站在珠my朗ma峰頂峰，我感到非常興奮和欣慰。...

Table 2: Illustration of LLaMA-2-70B responses to a question in Simplified Chinese. The text in black is model's actual output and text in gray is our self-added translation. The deactivated output is the generation when neurons of Simplified Chinese are deactivated.

3.3.1 Distribution and Identification Ratio

en	zh	fr	es	vi	id	ja
836	5,153	6,082	6,154	4,980	6,106	5,216

Table 3: The number of neurons in each language.

Neuron distribution across languages. After running our LAPE method on LLaMA-2 (70B), we identify around 23K language-specific neurons. The distribution of these neurons across individual languages is detailed in Table 3. Since neurons may be shared by multiple languages, the sum of language-specific neurons actually surpass 23K. Overall, except English, the distribution of neurons is relatively even across languages. However, the number of English-specific neurons is much smaller than the other languages. We speculate that English is the dominant language in LLaMA-2, and thus it requires fewer neurons to support the specific language ability.

Increasing threshold ratios for identification. In Section 3.2, we consider a mere 1% of the neurons as being language specific. We further vary the selection ratio of language-specific neurons from 0 to 10%, to examine its impact on multilingual processing. Here, we select *French* for study, while the results on the other languages are similar. When deactivating neurons specific to French, we observe a significant increase in the PPL on French in Figure 4, while the impact on the rest languages are relatively limited with the exception of Spanish. it is consistent with our intuition: the performance of the being perturbed language and its related (or similar) language would be severely harmed.

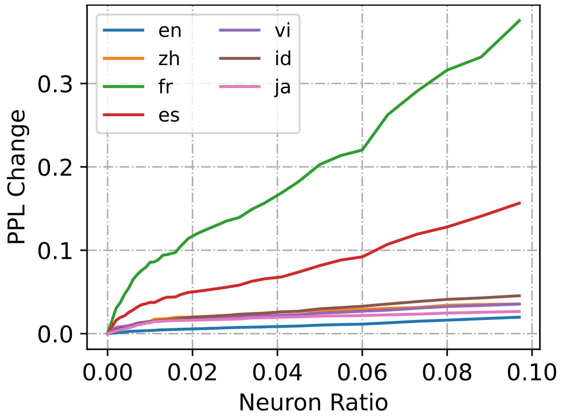


Figure 4: Change in PPL across different languages upon incremental number of language specific neurons when deactivating French neurons.

3.3.2 Structural Distribution Analysis

In this part, we analyze how language-specific neurons are distributed across different layers.

Language processing is concentrated at bottom and top layers. In Figure 5, we present the layer-wise distribution of language-specific neurons across various layers, which has a skewed "U" shape. This finding indicates that language-specific neurons have a pronounced concentration in the top and bottom layers. Specifically, the second layer has approximately 7,000 language-specific neurons, while layers 5 through 47 only contain about 100 neurons each. Further, the neuron count gradually increases, with the final four layers each comprising over 1,000 neurons. The complete statistics about the layer-wise distribution across various languages are reported in Table 10 of the Appendix.

Why such a skewed distribution? To understand why such a skewed distribution occurs, we seek explanations from multilingual semantic representation by exploring how multilingual aligned texts are represented across the layers. Specially, we employ the multilingual Vicuna dataset (Section 3.1), comprising of aligned texts in different languages. Given a group of aligned texts, we feed them into the LLM and obtain the sentence embedding of each text for each layer. We then compute the *mean sentence embedding similarity (SES)* between each pair of the aligned texts across

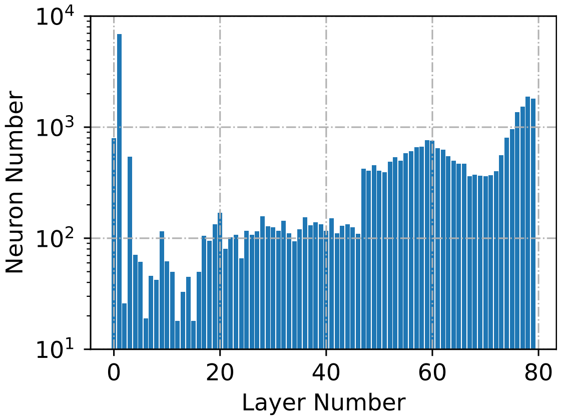


Figure 5: Distribution of language-specific neurons across different layers in LLaMA-2 (70B).

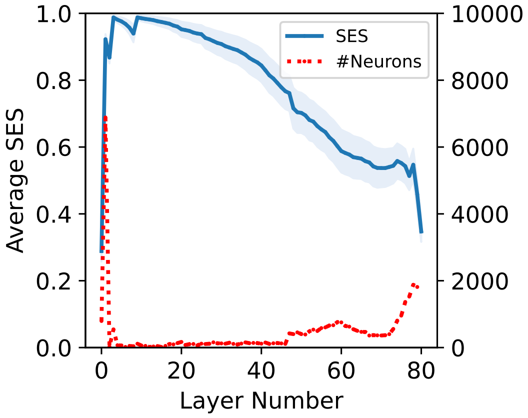
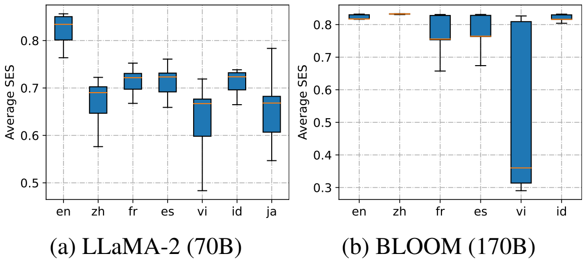


Figure 6: The mean SES between all language pairs and total language neuron numbers across layers.

languages in Figure 6. Interestingly, the SES curve shows an opposite trend with the distribution of language-specific neurons. At the beginning, the similarity quickly increases, then reaches the peak, and gradually decreases to a small value. This finding suggests that: at the bottom layers, the LLM needs to map aligned texts of different languages into the shared representation space, thus requiring more language-specific neurons for semantic transformation; while at top layers serving for token generation, the LLM needs to handle vocabulary mapping, which also requires the support of more language-specific neurons.

3.3.3 Language Dominance Analysis

Since the high-resource language (*i.e.,* English) in the LLaMA-2 training corpus has a surprisingly smaller number of neurons than other languages, as indicated in Table 3, we speculate that there might exist some dominance relations between high-resource and low-resource languages, which depends on the composition of pre-training corpus.



(a) LLaMA-2 (70B) (b) BLOOM (170B)

Figure 7: Dominance scores (mean SES) across layers when different languages serve as the target language.

Language dominance measurement. Inspired by Xu et al. (2023), we transfer the sentence embeddings across different languages into the same space around a target language, and examine how texts from the other languages are aligned to the texts of the target language. Firstly, for each language k, we compute the mean sentence embeddings of all its texts, and obtain $\boldsymbol{v}_k^i$ as the language vector of k at i-th layer. Then we follow the same formula proposed by Xu et al. (2023) to conduct the space mapping for each text from language k:

$$\hat{\boldsymbol{h}}_k^i = \boldsymbol{h}_k^i - \boldsymbol{v}_k^i + \boldsymbol{v}_c^i, \tag{6}$$

where $\hat{\boldsymbol{h}}_k^i$ denotes a transformed embedding of some text in the i-th layer of language k. Here, we specify c as the target language, and compute the mean sentence embedding similarity (SES, Section 3.3.2) over all the sentence pairs between languages k and c, based on the transformed representations in Eq. 6. A larger SES score indicates language c has a larger dominance degree.

Low-resource languages are centered around high-resource languages. To compute the dominance degree, we still use the multilingual Vicuna dataset (Section 3.1). From the results of LLaMA-2 (70B) in Figure 7a, we can see that the mean SES is obviously higher than all other languages when the target language is English. As English is the high-resource language of LLaMA-2, low-resource languages need to be aligned with English for achieving better performance. When it comes to BLOOM (170B) in Figure 7b, several languages (*e.g.,* English and Chinese) show dominance, since it is originally pre-trained on multilingual corpora.

3.3.4 Case Study

Finally, we explore the possibility to "steer" the output language of LLMs to mitigate the *off-target* problem and facilitate cross-lingual generation.

Metrics	Settings	zh	fr	es	vi	id	ja
Language accuracy	normal	0.87	0.73	0.81	0.60	0.40	0.79
	steered	0.99	0.90	0.93	0.97	0.99	1.00
Content quality	normal	4.30	4.19	3.51	3.70	4.16	2.86
	steered	4.57	4.35	4.02	3.57	4.28	2.91

Table 4: The language accuracy and content score of the normal output and the steered output by activating language-specific neurons. Language accuracy is computed by whether the model responds in a given language using the `langdetect` package and the content quality is measured by GPT-4.

Researchers have observed that when prompting in one language, language models may generate responses in a different language, such a phenomenon referred to as the *off-target language issue* (Gu et al., 2019; Sennrich et al., 2023). We speculate that language-specific neuron might not be activated in this case. Thus, we attempt to manually activate these neurons by increasing their activation value to the average activation value of this language. As shown in Table 4, it is evident that the model demonstrates a significantly higher propensity to respond in the expected language, as indicated by the language accuracy. Additionally, the generation quality, as assessed by GPT-4 metrics, shows remarkable improvement.

Moreover, we attempt to steer cross-lingual generation by posing a question in one language and answering it in another. As presented in Table 5, we deactivate Spanish neurons and activate Chinese neurons when replying to a question in Spanish. The model successfully generates fluent Chinese text, showing the feasibility of altering the output language through neuron-level manipulation.

Question

¿Cómo puedo mejorar mis habilidades de gestión del tiempo?

(*Translation*: How can I improve my time management skills?)

Steered output in Chinese

Time管理是一个非常重要的技能，因为它可以提高你的生产力并使你更有效地完成任务。下面是一些提示来改进你的时间管理技能：1、制定计划：每天开始时制定清单并将任务分配给特定时间段。这样可以确保你在每天都有足够的时间来完成所有任务。2、设置目标：设置明确而可实现的目标可以帮助你保持专注并提高效果。3、抵制分心：当你正在工作时，请不要被其他事情分心。4、利用技术：......

(*Translation*: Time management is a very important skill because it increases your productivity and allows you to complete tasks more efficiently. Here are some tips to improve your time management skills: 1. Make a plan: Make a list at the beginning of each day and assign tasks to specific time periods. This ensures that you have enough time each day to complete all your tasks. 2. Set goals: Setting clear and achievable goals can help you stay focused and improve your results. 3. Resist distractions: When you are working, please don't be distracted by other things. 4. Utilize technology: ...)

Table 5: An illustrative example of asking question in Spanish and answering it by deactivating Spanish neurons and activating Chinese neurons.

4 Related Work

Multilingual ability. Multilingual ability has been widely studied since the emergence of pretrained language models (Devlin et al., 2019; Xue et al., 2021). Recently, LLMs (Zhao et al., 2023b; Nguyen et al., 2023) showcase more excellent multilingual capabilities even without explicit language alignment (Kulshreshtha et al., 2020; Cao et al., 2020). Extensive research has been made to investigate the factors that influence models' multilingual ability (Philippy et al., 2023). For example, linguistic similarity has been examined (Pires et al., 2019; Dufter and Schütze, 2020; Wang et al., 2019; Conneau et al., 2020b), which is generally believed to correlate with models' cross-lingual ability. Specially, "word order" shows some contradictions about whether it really affects multilingual ability (Pires et al., 2019; Deshpande et al., 2022). Not only limited to language property, training settings have also been considered (Lauscher et al., 2020; Ahuja et al., 2022).

Existing work has explored language-agnostic (or language-shared) components within multilingual models. For example, researchers concentrate on shared knowledge across various languages (Stanczak et al., 2022; Chen et al., 2023a; Zhao et al., 2023a; Bricken et al., 2023). However, the exploration of language-specific components within LLMs remains an under-investigated area.

Neuron analysis. Neuron analysis has gained significant attention in recent years, paralleling research in neurobiological studies of the human brain (Friederici, 2011; Parr et al., 2022). Originating from vision models (Bau et al., 2020; Mu and Andreas, 2020), neuron analysis views neuron activation as the recall of learned knowledge (Sajjad et al., 2022). Researchers widely adopt these methods to analyze the sources of specific abilities or skills in language models, including sentiment analysis (Radford et al., 2017), knowledge storage (Dai et al., 2022), and task-solving (Wang et al., 2022).

Recent studies have also discovered that certain neurons can convey specialized contexts (Gurnee et al., 2023; Bills et al., 2023), such as positional information (Voita et al., 2023) and linguistic properties (Bau et al., 2018; Xin et al., 2019; Dalvi et al., 2019, 2020). Moreover, Gurnee et al. (2024) utilize Pearson correlation to calculate neuron similarity, identifying some universal neurons across models. In contrast to previous research, we have developed a method applicable to LLMs that unveils the mechanism of their multilingual abilities. This approach offers a more practical and effective solution for neuron analysis in multilingual scenarios.

5 Conclusion

Despite the impressive multilingual capabilities demonstrated by LLMs, the understanding of how these abilities develop and function remains nascent. In this paper, we introduced a novel detection method, *i.e.,* language activation probability entropy (*LAPE*), to pinpoint language-specific neurons within LLMs. LAPE assesses the response of individual neurons to various languages, selecting those with a propensity for activation when exposed to one or two languages. Based on LAPE, we further conducted extensive experiments to investigate the multilingual capabilities of LLMs. Specially, we have found that an LLM's proficiency in processing different languages is significantly influenced by a small subset of neurons, which are mainly located in the model's top and bottom layers. We have further demonstrated that the output language of LLMs can be directed by selectively enabling or disabling these language-specific neurons. For future work, we aim to leverage these findings to enhance knowledge transfer between major and minor languages and devise efficient strategies for continual pre-training to better accommodate specific languages.

Acknowledgement

This work was partially supported by National Natural Science Foundation of China under Grant No. 62222215, Beijing Natural Science Foundation under Grant No. L233008 and 4222027. Xin Zhao is the corresponding author.

Limitations

In this study, we employ language activation probability entropy as a metric to identify language-specific neurons. However, it is important to note that our method is relative to the presence of multiple languages. In scenarios where only a single language is present, establishing an absolute threshold to determine the language-relatedness of neurons is not feasible. Moreover, the criteria for distinguishing between high-resource and low-resource languages within the model warrant further investigation. The model's possibility to managing a large number of languages, as well as the differences between various languages, represents a promising avenue for future research. Finally, our research has only begun to explore the possibility for directing the output language of the model. Developing strategies to harness these identified neurons for enhancing the model's multilingual proficiency is still worth exploring.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick

Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023a. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *arXiv preprint arXiv:2308.13198*.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023b. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. *arXiv preprint arXiv:2004.04010*.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

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

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Angela D Friederici. 2011. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen

Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Arjun R Khanna, William Muñoz, Young Joon Kim, Yoav Kfir, Angelique C Paulk, Mohsen Jamali, Jing Cai, Martina L Mustroph, Irene Caprara, Richard Hardstone, et al. 2024. Single-neuronal elements of speech production in humans. *Nature*, pages 1–8.

Saurabh Kulshreshtha, José Luis Redondo-García, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual bert: A comparative study. *arXiv preprint arXiv:2009.14304*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Thomas Parr, Giovanni Pezzulo, and Karl J Friston. 2022. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2023. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *arXiv preprint arXiv:2309.07098*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Ji Xin, Jimmy Lin, and Yaoliang Yu. 2019. What part of the neural network does this? understanding lstms by measuring and dissecting neurons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5823–5830.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023a. Unveiling a core linguistic region in large language models. *arXiv preprint arXiv:2310.14928*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Appendix

Table 7 compiles the statistics of pre-training corpora in LLaMA-2 (Touvron et al., 2023b) and BLOOM (Scao et al., 2022).

We list the number of language-specific neurons across different layers of BLOOM (7B), LLaMA-2 (7B), LLaMA-2 (13B), and LLaMA-2 (70B) in Tables 6, 8, 9, and 10.

#Layer	en	zh	fr	es	vi	id
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	1	0	0	0	0
6	0	0	0	0	0	0
7	0	1	0	0	1	0
8	0	0	0	0	1	0
9	0	0	1	0	0	0
10	1	2	1	1	0	0
11	0	0	2	2	1	1
12	0	0	0	0	0	0
13	1	1	1	1	0	0
14	0	1	1	1	0	0
15	1	2	1	0	0	0
16	1	0	2	2	1	2
17	3	0	1	1	0	0
18	2	1	3	2	0	1
19	3	4	3	4	2	2
20	1	1	2	1	1	1
21	11	7	7	8	3	4
22	8	8	9	9	7	9
23	9	19	11	12	12	15
24	21	24	24	24	26	46
25	24	34	24	28	35	90
26	24	37	47	54	40	180
27	34	46	66	93	70	330
28	62	79	106	151	83	562
29	86	126	155	240	103	817
30	153	259	213	284	165	763

Table 6: Neuron number per layer of BLOOM (7B).

Language	Code	Family	BLOOM Ratio	LLaMA-2 Ratio
English	en	Indo-European	33.68%	89.70%
Chinese	zh	Sino-Tibetan	18.13%	0.13%
French	fr	Indo-European	14.46%	0.16%
Spanish	es	Indo-European	12.16%	0.13%
Vietnamese	vi	Austro-Asiatic	3.04%	0.08%
Indonesian	id	Austronesian	1.39%	0.03%
Japanese	ja	Japonic	0.00%	0.10%

Table 7: The language statistics of the pre-training corpora in BLOOM and LLaMA-2.

#Layer	en	zh	fr	es	vi	id	ja
1	17	108	220	195	274	221	109
2	0	32	39	27	16	15	31
3	0	1	2	2	0	2	0
4	0	0	0	0	0	0	0
5	2	3	6	4	4	4	3
6	3	5	5	4	2	3	4
7	0	9	10	8	8	4	4
8	1	5	1	1	3	1	3
9	0	2	1	0	1	1	3
10	0	3	3	4	3	4	5
11	0	5	1	0	5	2	6
12	3	7	5	4	3	0	6
13	1	8	10	8	11	8	11
14	2	19	7	5	16	8	18
15	1	13	12	10	13	9	15
16	1	7	3	1	5	4	15
17	3	28	17	14	15	12	20
18	3	11	13	11	19	16	18
19	1	17	6	7	16	13	21
20	2	20	18	8	20	24	26
21	3	19	33	15	35	29	32
22	3	22	21	23	26	49	13
23	0	33	60	42	38	84	35
24	2	20	56	31	49	84	18
25	0	20	78	58	33	77	19
26	3	11	80	54	30	78	17
27	2	18	86	72	43	88	7
28	2	14	50	59	35	64	13
29	5	15	49	48	36	58	14
30	7	23	44	39	27	40	17
31	18	36	54	52	31	38	29
32	10	49	32	32	19	28	50

Table 8: Neuron number per layer of LLaMA-2 (7B).

#Layer	en	zh	fr	es	vi	id	ja
1	60	127	222	189	248	184	206
2	9	162	177	118	187	69	305
3	0	2	1	2	1	2	1
4	0	1	1	1	0	2	2
5	0	3	0	1	0	0	3
6	1	3	2	2	3	3	5
7	0	5	1	1	4	3	6
8	3	7	3	3	3	1	9
9	2	18	7	7	10	3	9
10	0	12	9	6	8	5	9
11	0	15	18	17	11	8	12
12	2	5	3	2	5	3	9
13	1	7	2	1	3	2	9
14	0	5	3	2	4	0	10
15	1	7	3	3	7	5	8
16	3	25	20	14	22	11	31
17	3	30	16	11	28	21	32
18	4	40	40	31	35	25	47
19	1	26	23	13	22	19	44
20	1	24	14	16	14	9	35
21	1	28	19	17	22	17	34
22	3	43	26	13	40	37	55
23	3	32	23	10	22	24	49
24	1	28	20	12	20	32	27
25	1	24	29	8	23	25	29
26	3	27	40	32	27	42	31
27	2	40	63	41	31	64	36
28	4	20	50	43	21	48	30
29	2	25	78	48	19	71	26
30	0	25	89	88	38	75	19
31	2	21	72	52	46	77	13
32	3	16	83	60	36	80	14
33	0	23	69	55	31	61	19
34	1	27	47	54	35	69	16
35	1	20	69	58	41	67	23
36	1	21	60	54	42	58	27
37	5	22	33	32	31	47	11
38	14	40	58	52	48	44	36
39	8	57	43	35	26	27	51
40	15	105	47	51	38	44	97

Table 9: Neuron number per layer of LLaMA-2 (13B).

#Layer	en	zh	fr	es	vi	id	ja
1	238	199	45	43	28	47	195
2	117	886	1056	1155	1589	897	1184
3	2	5	2	3	5	5	4
4	11	79	105	69	62	108	109
5	0	17	9	6	12	12	15
6	5	10	11	10	10	7	8
7	2	3	2	3	3	2	4
8	1	14	5	5	6	4	11
9	2	9	7	6	5	5	8
10	2	25	23	15	17	8	25
11	0	13	14	10	8	6	11
12	0	16	5	6	7	5	11
13	0	5	2	1	2	2	6
14	0	9	3	2	7	2	10
15	0	14	3	3	8	4	13
16	0	4	1	4	3	1	5
17	1	13	7	8	7	5	9
18	3	22	12	13	16	10	29
19	2	28	13	11	11	8	22
20	4	34	19	21	18	11	26
21	1	38	27	19	26	25	33
22	1	16	17	15	10	7	14
23	1	23	17	14	15	14	18
24	1	18	15	14	26	13	20
25	2	10	11	11	9	11	12
26	1	23	12	15	17	14	35
27	4	28	13	11	10	12	29
28	3	25	14	16	20	17	20
29	6	39	23	21	19	19	30
30	0	24	23	23	19	19	20
31	1	15	30	24	20	22	13
32	2	21	17	20	23	16	18
33	2	21	22	17	23	27	32
34	1	20	19	13	18	23	17
35	0	14	12	17	18	19	14
36	3	17	22	16	20	23	19
37	4	26	29	18	24	24	30
38	4	17	31	24	16	19	20
39	2	18	27	26	17	23	26
40	4	20	26	15	23	21	24
41	2	17	15	14	20	25	24
42	2	21	26	22	22	30	28
43	0	21	13	15	17	17	28
44	1	17	18	14	23	31	25
45	1	24	23	11	22	30	23
46	1	17	21	11	18	33	24
47	2	13	22	14	14	27	18
48	1	55	78	64	55	93	77
49	2	54	68	61	55	73	90
50	4	61	85	100	42	98	65
51	0	49	97	66	43	80	69
52	3	49	99	80	37	83	40
53	1	64	120	96	55	86	64
54	0	55	136	130	54	116	46
55	1	55	118	109	52	114	49
56	4	62	134	130	74	135	44
57	0	47	149	162	64	132	50
58	1	40	187	172	73	142	45
59	0	38	162	184	87	155	43
60	4	59	190	208	84	163	54
61	1	57	178	180	80	211	47
62	2	39	142	165	75	180	43
63	2	45	137	160	72	174	36
64	2	36	123	138	76	138	37
65	0	40	104	123	58	145	31
66	3	35	90	112	67	124	39
67	4	51	86	103	69	112	43
68	1	27	63	74	55	101	40
69	3	33	64	69	73	85	44
70	6	39	67	66	56	81	51
71	8	55	60	58	65	69	47
72	4	50	75	75	55	64	47
73	10	74	62	60	53	87	55
74	18	94	84	82	88	107	84
75	15	154	132	154	98	140	113
76	30	188	139	152	122	178	148
77	37	254	244	239	162	242	186
78	57	292	245	255	177	270	230
79	89	450	256	263	192	226	402
80	81	484	219	220	179	192	438

Table 10: Neuron number per layer of LLaMA-2 (70B).

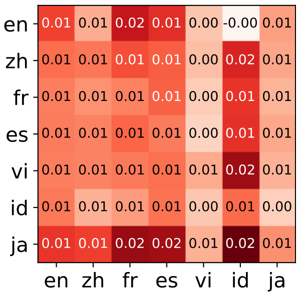


Figure 8: The results of deactivating neurons randomly.