Sharing Matters: Analysing Neurons Across Languages and Tasks in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) have revolutionized the field of natural language processing (NLP), and recent studies have aimed to understand their underlying mechanisms. However, most of this research is conducted within a monolingual setting, primarily focusing on English. Few studies have attempted to explore the internal workings of LLMs in multilingual settings. In this study, we aim to fill this research gap by examining how neuron activation is shared across tasks and languages. We classify neurons into four distinct categories based on their responses to a specific input across different languages: *all-shared*, *partial-shared*, *specific*, and *non-activated*. Building upon this categorisation, we conduct extensive experiments on three tasks across nine languages using several LLMs and present an in-depth analysis in this work. Our findings reveal that: (i) deactivating the *all-shared neurons* significantly decreases performance; (ii) the shared neurons play a vital role in generating responses, especially for the *all-shared neurons*; (iii) neuron activation patterns are highly sensitive and vary across tasks, LLMs, and languages. These findings shed light on the internal workings of multilingual LLMs and pave the way for future research. We will release the code to foster research in this area.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in recent studies, excelling in both understanding and generating text across various languages (OpenAI, 2023; Zhang et al., 2023; Zhao et al., 2024a). Despite their proven effectiveness, the intricate mechanisms underlying their processing remain largely opaque. This opacity has given rise to a growing field of research aimed at interpreting the internal workings of the Transformer architecture (Elhage et al., 2021; Yu et al., 2023). To enhance interpretability and in-

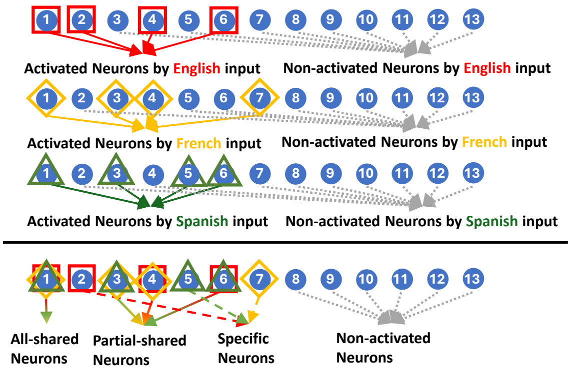


Figure 1: A comparison of neuron analysis with different type designs in multilingual settings with the same semantic input, in which we define four types of neurons in one layer of LLM.

vestigate specific aspects of model behavior, researchers have increasingly focused on the components of these models. Recent studies have explored the role of Feed-Forward Networks (FFNs) within LLMs, proposing that these components function as key-value memories for storing factual and linguistic knowledge (Geva et al., 2020, 2022; Ferrando et al., 2023). While these studies have analyzed neuron behaviors based on activation states in monolingual settings, there remains a significant gap in our understanding of how neurons behave in multilingual contexts.

To address this research gap, recent research attempts to unveil the mechanistic interpretability of multilingual LLMs. Bhattacharya and Bojar (2023) categorized neurons into two coarse-grained groups: language-agnostic (shared across languages) and language-specific (unique to a language). However this categorization oversimplifies the complexity observed in cross-lingual studies, where neuron overlap varies significantly between languages (Stanczak et al., 2022; Zhao et al., 2023; Liu et al., 2024). Additionally, most research has been confined to single-task analyses, overlooking how neuron types might shift across diverse tasks (Bhattacharya and Bojar, 2023; Tang et al., 2024; Tan et al., 2024). This underscores the need for a more nuanced, fine-grained classification method to enhance our understanding of the multifaceted roles of neurons in multilingual LLMs.

In this work, our research introduces a fine-grained classification of neurons, enabling a detailed exploration of their functions across languages. For a specific English example and its translations in eight other languages, we categorize neurons into four distinctive types (see Figure 1): *all-shared neurons*, which remain active for all the inputs regardless of language; *partial-shared neurons*, which are activated only for inputs in certain languages; *specific neurons*, which are activated exclusively for inputs in one language; and *non-activated neurons*, which are not activated for any inputs. We begin by analysing the importance of each neuron type by deactivating them individually. Then we probe their contributions to generating answers using the Generation Impact Score (Geva et al., 2022) and the Correctness Impact Score (Voita et al., 2023). Furthermore, by examining the percentage of neurons in each type, we analyse activation patterns to gain insights into the internal workings of LLMs. We systematically study neuron behaviours across three distinct tasks, including reasoning, fact probing, and question answering, in nine languages. This analysis utilizes diverse model backbones such as BLOOMZ-7B, LLAMA2-7B-CHAT, BLOOM-7B, and XGLM.

We provide substantial empirical evidence detailing neuron contributions and activation patterns in this study, leading to several significant findings. Here are the main takeaways:

- ***All-shared neurons* have a significant impact on model performance.** We individually deactivate each type of neurons in LLMs and observe substantial performance declines (up to 87.39%) across tasks (see Section 5).

- ***All-shared neurons* are crucial in generating responses.** Both the Generation Impact Score and Correctness Impact Score highlight the significance of the shared neurons in the generation process, and the *all-shared neurons* make substantially more contributions compared to other neuron types (see Section 6).

- **Neuron activation patterns vary across tasks, LLMs, and languages.** We observe that the patterns of four types of neurons vary across tasks (see Section 7.2) and LLMs (see Section 7.3). Moreover, our empirical results show that languages from the same language family do not always exhibit a higher degree of neuron sharing compared with languages from distinct language families (see Section 7.4).

2 Related Work

The black-box nature of LLMs has given rise to an area of research which aims to interpret the internal mechanism of the Transformer architecture (Elhage et al., 2021; Yu et al., 2023). More recently, several studies on LLMs have advanced our understanding of how neurons acquire task-specific knowledge. For instance, Ferrando et al. (2023); Dai et al. (2022); Geva et al. (2020, 2022) investigated how FFN blocks function as key-value memories and proved that factual knowledge is stored in the neurons. Research work on the sparsity of neurons in FFN blocks showed that many neurons are inactive in various tasks (Zhang et al., 2022; Li et al., 2023). Voita et al. (2023) located these "dead" neurons in the lower part of the model (close to inputs) in the English scenario. Despite the insights obtained, these studies have focused exclusively on a monolingual setting.

For multilingual neuron analysis, Bhattacharya and Bojar (2023) explored the neuron sharing between two languages. Tang et al. (2024); Tan et al. (2024); Liu et al. (2024); Kojima et al. (2024) classified neurons in an FFN block to language-specific and language-agnostic based on predefined threshold. However, the broad classification into two groups is inadequate for detailed multilingual analysis. Additionally, these studies classified neurons based on the single task (Tan et al., 2024; Liu et al., 2024), without considering the potential adaptation of neurons under various languages and semantics brought forth by inputs from various multilingual tasks. We investigate neurons' behaviors across multiple languages and tasks to this end.

3 Fine-Grained Neuron Classification

In this section, we provide a detailed description of the 4-way neuron classification that we propose. We begin with some background concerning neurons in the FFN block (Section 3.1). Following this, we define the four types of neurons (Section 3.2).

	XNLI			KE (EN → ALL)			KE (ALL → EN)			Fact Probing		
	pct.	μ_{acc}	Δ_{acc}	pct.	μ_{acc}	Δ_{acc}	pct.	μ_{acc}	Δ_{acc}	pct.	μ_{acc}	Δ_{acc}
baseline	0.00%	41.99	0.00%	0.00%	38.39	0.00%	0.00%	41.74	0.00%	0.00%	41.98	0.00%
w/o. all	9.92%	9.38	**-77.66%**	8.71%	4.84	**-87.39%**	10.17%	13.19	**-68.40%**	0.28%	21.86	-50.31%
w/o. partial	10.33%	42.65	1.57%	13.36%	40.67	5.94%	10.55%	39.59	-5.15%	36.73%	26.86	-36.02%
w/o. specific	3.14%	42.07	0.19%	4.91%	40.78	6.23%	3.82%	40.77	-2.32%	16.56%	12.68	**-67.41%**
w/o. non-act.	76.61%	35.90	-14.50%	73.22%	21.96	-42.80%	75.46%	19.58	-53.09%	46.43%	26.68	-36.45%
w/o. random	5.00%	42.30	0.74%	5.00%	30.98	-19.30%	5.00%	41.29	-1.08%	1.00%	37.86	-9.81%
	15.00%	43.13	2.71%	15.00%	31.74	-17.32%	15.00%	42.14	0.96%	15.00%	35.38	-15.72%
	25.00%	43.98	4.74%	25.00%	32.40	-15.60%	25.00%	42.28	1.29%	35.00%	41.78	-0.48%
	75.00%	36.58	-12.88%	75.00%	13.29	-65.38%	75.00%	16.50	-60.47%	45.00%	17.06	-59.36%

Table 1: The performance on `XNLI`, `Cross-lingual KE`, and `Fact Probing` tasks, using BLOOMZ-7B, when deactivating *all-shared neurons*, *specific neurons*, *partial-shared neurons*, *non-activated neurons*, and random selected neurons, respectively. The largest reductions are highlighted in **bold**. "pct." indicates the percentage of the deactivated neurons. μ_{acc} indicates the macro-average accuracy across languages. Δ_{acc} indicates the macro-average of relative change (%) in accuracy across languages.

3.1 Neurons in FFN Blocks

A neuron inside the FFNs is defined as a linear transformation of an input representation followed by a non-linear activation (Tang et al., 2024). Every FFN block at layer l involves two linear transformations separated by a point-wise activation function. Biases are omitted for brevity:

$$FFN^l(x^l) = Act(W_K^l x^l) W_V^l \tag{1}$$

where $W_K^l \in \mathbb{R}^{d \times d_m}$,$W_V^l \in \mathbb{R}^{d_m \times d}$ are linear parameter matrices, and $Act(\cdot)$ is a non-linear activation function, where rows in W_K^l and columns in W_V^l are viewed as d-dimensional keys k^l and values v^l, respectively. d_m is the count of neurons. And the output of neurons $A^l := Act(W_K^l x^l) \in \mathbb{R}^{d_m}$ determines the weighting of the corresponding values in W_V^l.

For the i-th neuron and corresponding key k_i^l, value v_i^l and activation value A_i^l, we can express this relationship using the following formulation:

$$FFN^l(x^l) = \sum_{i=1}^{d_m} Act(x^l \cdot k_i^l) v_i^l = \sum_{i=1}^{d_m} A_i^l v_i^l \tag{2}$$

Following Voita et al. (2023); Bhattacharya and Bojar (2023); Tang et al. (2024), we define a neuron as activated when its activation value satisfies $A_i^l > 0$. Conversely, if the activation value is $A_i^l \leq 0$, the neuron is considered deactivated.

3.2 Definitions of Four Types of Neurons

In this work, we categorize the neurons into four types based on their activation values and detail the neuron classification in this section. To ablate the impact of semantic discrepancies across languages, the datasets used in this work are initially in English and then translated into foreign languages (see Section 4.1), so we can formulate the s-th example as $\boldsymbol{X}^s = \{X_p^s\}_{p=1}^P$, where p indicates the p-th language and P is the total number of languages. Given the s-th example $\boldsymbol{X}^s$, the set of ***all-shared neurons*** at the l-th layer can be defined as:

$$N_{\text{all}}^{s,l} := \bigcap_{p}^{P} \{n^i \in N^l : A_{i,p}^{s,l} > 0\}. \tag{3}$$

where N^l is the set of all the neurons at the l-th layer and n^i is the i-th neuron in N^l. Furthermore, the ***non-activated neurons*** is the set of neurons whose activation value is less than or equal to zero in all languages, as follows:

$$N_{\text{non}}^{s,l} := \bigcap_{p}^{P} \{n^i \in N^l : A_{i,p}^{s,l} \leq 0\}. \tag{4}$$

Moreover, the ***specific neurons*** are the neurons only activated in one specific language and not activated in any other languages, defined as follows:

$$N_{\text{spec}}^{s,l} := \bigcup_{p'}^{P} \{\{n^i \in N^l : A_{i,p'}^{s,l} > 0\} \bigcap_{\substack{p \\ p \neq p'}}^{P} \{n^i \in N^l : A_{i,p}^{s,l} \leq 0\}\} \tag{5}$$

Lastly, the remaining neurons are ***partial-shared neurons*** as they are activated by inputs from a subset of languages:

$$N_{\text{part}}^{s,l} := N^l \setminus \{N_{\text{all}}^{s,l} \bigcup N_{\text{non}}^{s,l} \bigcup N_{\text{spec}}^{s,l}\} \tag{6}$$

Note that, we only examine the activation state of the last token of the input, as that is when the LLM performs the prediction task.

4 Experimental Setting

4.1 Multilingual Tasks

We perform analysis on neurons in FFN blocks of various LLMs, harnessing their multilingual capabilities in three diverse tasks which consist of multilingual parallel sentences, including `XNLI` (Conneau et al., 2018), `Fact Probing` (Fierro and Søgaard, 2022), and `Cross-lingual Knowledge Editing (KE)` (Wang et al., 2023). For the `Cross-lingual KE`, we analyse the LLMs in two setups, including `EN (Edit)` $\rightarrow$ `ALL (Test)` and `ALL (Edit)` $\rightarrow$ `EN (Test)`. These test sets across languages are translated from the original English test set. More details are described in Appendix B.

These tasks cover nine diverse languages, including English (en), German (de), Spanish (es), French (fr), Russian (ru), Thai (th), Turkish (tr), Vietnamese (vi), and Chinese (zh). Prompts are detailed in Appendix C.

4.2 Model Backbones

We mainly analyse the contributions and activation patterns of neurons in an instruction-finetuned multilingual model BLOOMZ-7B (Muennighoff et al., 2023). We also include the analysis of other multilingual LLMs: BLOOM-7B (Scao et al., 2022), LLAMA2-7B-CHAT (Touvron et al., 2023), and XGLM (Lin et al., 2022). We use one NVIDIA A100 (40G) for all experiments.

settings	pct.	en	de	es	fr	ru	th	tr	vi	zh
					`XNLI` task					
baseline	0%	53.8	41.8	50.3	49.0	47.6	40.9	34.9	50.5	51.1
w/o. all	9.9%	**16.7**	**3.5**	**10.1**	**10.0**	**6.6**	**9.0**	**1.4**	**12.1**	**14.5**
w/o. partial	10.3%	52.9	40.4	49.7	47.6	49.2	40.3	36.1	50.0	50.0
w/o. specific	3.1%	53.7	41.7	50.3	48.9	47.4	40.6	35.3	50.4	49.3
w/o. non-act.	76.7%	36.6	31.6	33.6	33.4	29.5	31.3	28.3	34.5	23.5
w/o. random	5%	53.2	42.2	50.7	48.8	47.4	40.2	34.5	50.1	50.9
	15%	53.1	41.8	50.1	48.9	47.3	40.8	33.8	50.1	50.4
	25%	52.6	41.7	50.3	48.8	46.0	38.8	36.2	50.7	49.7
	75%	36.0	28.7	40.7	36.7	28.9	25.4	23.0	38.5	32.9
					`Cross-lingual KE (EN (Edit)` $\rightarrow$ `ALL (Test))` task					
baseline	0%	96.2	48.8	36.9	49.5	24.6	6.3	38.8	49.4	33.4
w/o. all	8.7%	**11.0**	**4.9**	**6.1**	**4.9**	1.9	**0.4**	**1.5**	**6.1**	**2.9**
w/o. partial	13.3%	90.2	51.7	46.9	48.9	25.4	5.5	35.5	50.9	38.4
w/o. specific	4.9%	96.1	54.4	48.7	48.9	30.4	6.3	37.9	51.7	28.5
w/o. non-act.	73.2%	36.1	15.8	18.2	17.8	1.9	9.8	19.9	10.6	16.3
w/o. random	5%	96.1	46.9	36.6	40.4	0.8	4.4	28.7	40.8	10.1
	15%	94.8	47.1	36.1	39.4	0.8	4.3	28.4	40.4	11.1
	25%	91.5	46.8	36.2	38.6	1.1	4.4	27.9	40.4	12.1
	75%	11.1	5.5	9.3	11.3	**0.1**	2.7	1.9	8.9	7.1
					`Cross-lingual KE (ALL (Edit)` $\rightarrow$ `EN (Test))` task					
baseline	0%	96.2	55.1	49.2	49.5	30.6	9.2	39.3	51.7	36.6
w/o. all	10.2%	24.4	**19.5**	**13.8**	**13.1**	**8.0**	**1.4**	**14.5**	19.9	**7.1**
w/o. partial	10.5%	85.1	51.3	47.8	48.0	25.4	5.2	35.1	51.4	36.1
w/o. specific	3.8%	96.1	54.4	48.7	48.9	29.7	6.3	38.0	51.8	30.0
w/o. non-act.	75.5%	**19.2**	19.8	15.5	14.7	11.4	2.0	18.2	**12.0**	7.5
w/o. random	5%	95.6	54.2	48.7	50.2	29.9	6.5	38.5	51.3	33.0
	15%	93.9	56.1	49.1	49.9	29.2	6.4	38.2	51.0	32.6
	25%	91.3	55.9	48.5	49.3	28.3	6.8	37.7	50.3	29.7
	75%	10.2	17.0	11.7	15.7	4.7	1.1	7.8	14.8	7.0
					`Fact Probing` task					
baseline	0%	72.4	41.6	56.6	58.1	37.3	5.7	39.3	57.4	51.4
w/o. all	0.2%	43.4	12.9	34.4	22.4	11.4	**5.2**	15.2	34.5	29.0
w/o. partial	36.7%	43.3	20.8	31.2	30.9	14.4	2.8	24.5	34.6	29.4
w/o. specific	16.6%	**18.1**	**9.1**	**30.1**	**7.7**	**5.7**	5.7	**9.4**	**12.3**	**22.1**
w/o. non-act.	46.4%	42.5	27.6	39.9	28.1	1.7	0.0	22.9	40.2	17.5
w/o. random	1%	76.4	50.6	48.6	56.0	3.2	0.0	36.2	59.5	47.1
	15%	71.5	48.5	45.6	63.5	4.3	0.0	22.7	44.5	38.2
	35%	77.3	51.4	50.3	56.3	4.6	0.0	37.1	57.5	48.3
	45%	29.0	21.3	16.4	17.2	0.3	0.0	9.5	25.9	6.0

Table 2: The performance on three tasks, using BLOOMZ-7B, when deactivating *all-shared neurons*, *specific neurons*, *partial-shared neurons*, *non-activated neurons*, and randomly selected neurons, respectively. The largest reductions are highlighted in **bold**. "pct." indicates the percentage of the deactivated neurons.

5 Shared Neurons Are Crucial to Performance

In this section, we explore how different neuron types affect the performance of the BLOOMZ-7B model by selectively deactivating specific groups of neurons. By setting the activation values of these neurons to zero, we assess their impact on the model's output across various tasks. Specifically, we compare the effects of deactivating four distinct types of neurons and include a control group of randomly selected neurons to evaluate their respective contributions to the model performance. Our experiments involve tasks such as `XNLI`, `cross-lingual KE`, and `fact probing`.

***All-shared neurons* play a crucial role in model performance across different tasks.** As shown in Table 1, we observe that *all-shared neurons* significantly contribute to the model's performance across various tasks. For instance, in the `Cross-lingual KE (EN (Edit)` $\rightarrow$ `ALL (Test))` task, deactivating the *all-shared neurons*, which account for only 8.71% of the total neurons, results in an 87.39% decrease in accuracy. Moreover, for the `Fact Probing` task, deactivating the *all-shared neurons*, which constitute only 0.28% of the total neurons, causes a substantial 50.31% performance drop. Furthermore, deactivating the *specific neurons*, which account for 16.56% of the total neurons, leads to the largest performance decline of 67.41%. In comparison, deactivating a comparable number of random selected neurons typically results in smaller performance drops, suggesting that *all-shared neurons* are crucial to the performance.

Deactivating neurons does not always result in performance declines. Interestingly, we some-times observe small performance gains when a small number of neurons are deactivated, as shown in Table 1, regardless of the neuron type. To explore this phenomenon further, we provide a breakdown of the results by language in Table 2. Our analysis reveals that only deactivating the *all-shared neurons* consistently leads to a decline in model performance across various tasks and languages. In contrast, deactivating either *partial-shared neurons* or *specific neurons* can occasionally improve performance for certain languages. For example, in the `Cross-lingual KE (EN (Edit)` $\rightarrow$ `ALL (Test))` task, we observe substantial performance improvements in German (de), Spanish (es), and Chinese (zh) when the *partial-shared neurons* are deactivated. We hypothesize that this phenomenon stems from knowledge conflicts encoded in the LLM (Xu et al., 2024). By deactivating certain neurons, these knowledge conflicts may be mitigated, resulting in enhanced performance.

It is important to note that we conduct similar experiments using the LLAMA2-7B-CHAT and present the results in Appendix D. These additional experiments yield observations and conclusions consistent with those using BLOOMZ-7B.

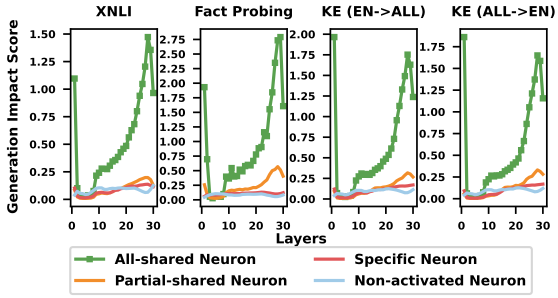
6 Probing Neuron Contributions

We demonstrate the significant role of neurons shared across languages, particularly the *all-shared neurons*, in generating answers, as discussed in Section 5. To gain a deeper understanding of the model's behavior, we conduct further analysis using two metrics: **Generation Impact Score** and **Correctness Impact Score**. First, we introduce the definitions of these two metrics in Section 6.1. Then, we analyse and quantify the contributions of each type of neuron in Section 6.2 and Section 6.3, respectively.

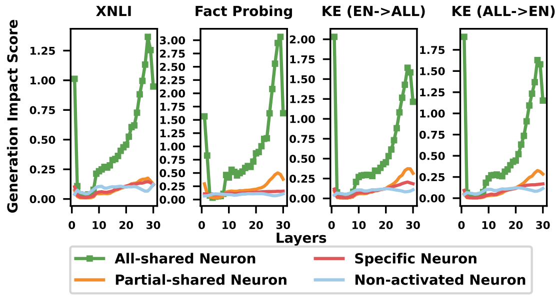
6.1 Generation Impact Score and Correctness Impact Score

In this section, we introduce two measures to quantify the contributions of neurons: Generation Impact Score and Correctness Impact Score.

Generation Impact Score Inspired by Geva et al. (2022), the Generation Impact Score (GIS) evaluates the importance of neurons in generating answers. For the i-th neuron at l-th layer, the GIS is



(a) English test set.



(b) German test set.

Figure 2: Average Generation Impact Score of the four types of neurons on the English and German test sets across tasks given by BLOOMZ-7B.

defined as:

$$GIS_i^l := \frac{|A_i^l| \, ||v_i^l||}{\sum_{j=1}^{d_m} |A_j^l| \, ||v_j^l||} \tag{7}$$

which is the proportion of its weight to the sum of weights of all neurons in the FFN block. $|A_i^l|$ is the absolute value of activation value and $||v_i^l||$ is the L2-norm of value v_i^l.

Correctness Impact Score Following Geva et al. (2022) and Voita et al. (2023), Correctness Impact Score (CIS) assesses a neuron's influence on generating the correct answer.

$$CIS_i^l = E_r \cdot A_i^l v_i^l \tag{8}$$

where E_r is the embedding of the correct answer r. A larger CIS_i^l has a higher probability to produce the correct answer r, while a negative CIS_i^l reduces the probability in generating r. Detailed descriptions of the neuron projection are provided in Appendix A.

Comparison While both Generation Impact Score (GIS) and Correctness Impact Score (CIS) measure neuronal influence, they serve different purposes. The GIS quantifies a neuron's overall contribution to the generation process, regardless

	all-shared				partial-shared				specific				non-activated			
	max	min	mean	var	max	min	mean	var	max	min	mean	var	max	min	mean	var
en	1.85	-0.94	0.07	0.36	0.22	-0.16	1.2e-4	1.9e-4	0.02	-0.02	2.5e-4	3.5e-5	0.04	-0.03	2.1e-4	5.8e-6
de	1.03	-0.60	0.02	0.07	0.13	-0.13	6.7e-5	7.1e-5	0.07	-0.03	2.3e-5	2.3e-5	0.02	-0.01	2.9e-6	2.9e-6
es	1.15	-0.84	0.02	0.07	0.12	-0.11	1.3e-4	6.3e-5	0.01	-0.01	9.4e-5	7.6e-6	0.02	-0.02	7.7e-5	3.1e-6
fr	1.06	-0.78	0.01	0.05	0.15	-0.11	2.1e-4	7.8e-5	0.03	-0.04	3.6e-5	9.9e-6	0.02	-0.02	5.6e-5	2.8e-6
ru	0.70	-0.45	3.3e-3	5.9e-3	0.24	-0.13	2.6e-4	7.6e-5	0.08	-0.03	1.1e-4	1.8e-5	0.01	-0.01	3.9e-5	1.7e-6
th	0.50	-0.90	2.6e-3	0.02	0.17	-0.10	2.2e-4	6.8e-5	0.03	-0.05	3.4e-5	1.8e-5	0.01	-0.01	7.1e-5	1.9e-6
tr	0.82	-0.51	0.03	0.07	0.12	-0.12	1.6e-4	7.4e-5	0.04	-0.03	6.6e-5	1.1e-5	0.02	-0.02	9.1e-5	3.4e-6
vi	0.86	-0.68	6.8e-3	0.03	0.15	-0.11	9.3e-5	6.8e-5	0.04	-0.04	2.8e-5	1.3e-5	0.02	-0.02	3.2e-5	2.6e-6
zh	0.52	-0.42	1.9e-3	0.02	0.17	-0.20	1.5e-4	7.7e-5	0.08	-0.07	8.5e-5	2.6e-5	0.02	-0.01	1.7e-5	3.1e-6

Table 3: Maximum, minimum, average, and variance of Correctness Impact Score of the four types of neurons on the `Cross-lingual KE (EN (edit)` $\rightarrow$ `ALL (Test))` task given by BLOOMZ-7B.

of output correctness. In contrast CIS specifically measures a neuron's impact on producing accurate responses by incorporating the correct answer's embedding. Thus, the key distinction lies in their consideration of answer correctness: GIS focuses on general generation ability, whereas CIS emphasizes correctness.

6.2 The Generation Impact of Neuron Types

In this section, we explore the contribution of each neuron type using the Generation Impact Score (GIS) described in Section 6.1.

***All-shared neurons* have the greatest impact on generation outputs.** As shown in Figure 2, we analyse the GIS across layers on the English and German test sets of three tasks (with overall results provided in Appendix E). For both English and German, it can be observed that the *all-shared neurons* almost always achieve the highest GIS across all layers, indicating their significant influence on the model's output generation. The *partial-shared neurons* are the second most influential, particularly in the upper layers. Notably, there is a decrease in the influence of *all-shared neurons* between layers 5 and 10. This can be attributed to the fact that GIS assesses the impact on generating answers, while the lower layers are primarily responsible for input understanding (Zhao et al., 2024b). Consequently, all types of neurons exhibit lower GIS in these layers. Moreover, previous studies have demonstrated that higher layers capture more abstract, high-level information essential for generation (Gao et al., 2024). These findings suggest that shared neurons play a more significant role in the model's generation capabilities.

6.3 The Correctness Impact of Neuron Types

In this section, we assess the effectiveness of each neuron type using the Correctness Impact Score (CIS) described in Section 6.1.

***All-shared neurons* have the greatest impact on generating *correct* answers.** In the `Cross-lingual KE (EN (Edit)` $\rightarrow$ `ALL (Test))` task, we present the maximum, minimum, average, and variance of CIS for each neuron type across all layers of the BLOOMZ-7B model, as shown in Table 3. The results reveal that *all-shared neurons* have both the highest maximum and the lowest minimum CIS values, indicating that they have strong impact on generating correct outputs. While *all-shared* and *partial-shared neurons* display a wide variance in CIS (e.g., 1.85 vs. -0.94 and 0.22 vs. -0.16 in English, respectively), *specific neurons* and *non-activated neurons* exhibit much narrower score ranges (approximately $\pm$ 0.07). Furthermore, the *all-shared neurons* also exhibit the largest mean and variance of CIS among all kinds of neurons.

In conclusion, these findings presented in Section 6.2 and Section 6.3 demonstrate that *all-shared neurons* also have the greatest impact on generating both answers and correct answers, highlighting their importance in the model's performance across different languages and tasks.

7 Understanding Neuron Activations

We demonstrate in Section 5 that shared neurons have a significant impact on model performance and investigate their influence on the generation process in Section 6. However, the inner patterns of neurons across layers remain unexplored. In this section, we firstly introduce the measure of quantifying neuron activation in Section 7.1, and then we further illustrate how neuron activation patterns vary across tasks (Section 7.2), LLMs (Section 7.3) and languages (Section 7.4).

7.1 Measuring Neuron Activation

In this section, we explain how to quantify neuron activation patterns based on the definitions in Section 3.2. Specifically, we measure the percentage

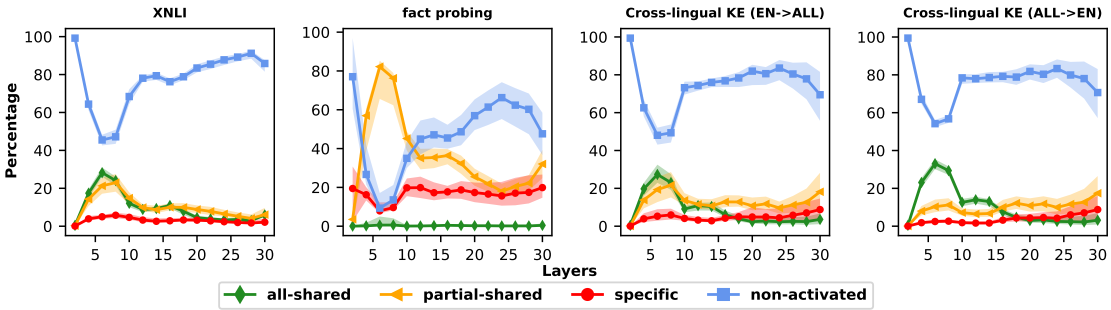


Figure 3: Neuron activation pattern ($R^l_{\{\cdot\}}$) in the `XNLI`, `Fact Probing`, `Cross-lingual KE (EN (Edit)` $\rightarrow$ `ALL (Test)`, and `Cross-lingual KE (ALL (Edit)` $\rightarrow$ `EN (Test)` tasks with BLOOMZ-7B backbone. It shows the percentage of each type of neuron relative to the total number of neurons across layers.

of each type of neuron relative to the total number of neurons. Given the s-th test instance, the percentage of each neuron type $R^{s,l}_{\{\cdot\}}$ at the l-th layer can be defined as follows:

$$R^{s,l}_{\{\cdot\}} = 100 \times \frac{|N^{s,l}_{\{\cdot\}}|}{|N^l|}, \tag{9}$$

where $|\cdot|$ denotes the number of elements in the set. Consequently, the aggregated neuron activation pattern at the l-th layer for one dataset containing S instances can be defined as:

$$R^l_{\{\cdot\}} = \frac{1}{S}\sum_{s=1}^{S} R^{s,l}_{\{\cdot\}}. \tag{10}$$

7.2 Neuron Activations Across Tasks

Neuron activations are task-related. As shown in Figure 3, *non-activated neurons* are typically more prevalent than other types of neurons, except in the `Fact Probing` task. In this task, there are more *partial-shared neurons* and *specific neurons*, with a negligible amount of *all-shared neurons*, whereas other tasks involve far more *all-shared neurons*. Referring to Table 1, deactivating the *specific neurons* and *all-shared neurons* results in the largest and second largest performance declines. These findings demonstrate that the some factual knowledges in LLMs are language-specific and minimally shared across languages, while others are universally shared. We leave more in-depth investigation to the future work.

Neuron sharing peaks at early layers for universal features, declining later for specific ones. We present the percentage of each neuron type at each layer in Figure 3. The number of *all-shared neurons* and *partial-shared neurons* typically peaks between the 5th and 10th layers and then gradually

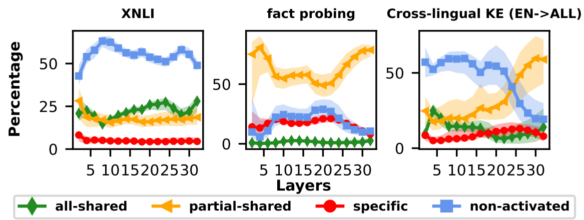


Figure 4: Neuron activation patterns in the `XNLI`, `Fact Probing`, `Cross-lingual KE (EN (Edit)` $\rightarrow$ `ALL (Test)` tasks with LLAMA2-7B-CHAT backbone.

decreases in subsequent layers. This trend can be explained by the functional roles of different layers in the model. The initial layers, which are closer to the input data, primarily focus on capturing low-level features such as basic lexical and syntactic patterns. As the network progresses to the later layers (between the 5th and 10th layers), it begins to learn abstract concepts that are relatively universal across different tasks and languages. This universality leads to a higher number of shared neurons in these layers. In contrast, the higher layers specialize in task-specific features and nuances unique to each task, resulting in a decline in neuron sharing. These findings highlight the importance of neuron sharing in LLMs, as shared neurons in the early layers facilitate the transfer of universal knowledge across tasks and languages. They also align with previous research (Yosinski et al., 2014; de Vries et al., 2020; Zhao et al., 2024b; Gao et al., 2024).

7.3 Neuron Activations Across LLMs

Different LLMs exhibit different neuron activation patterns. To investigate whether neuron activation patterns vary across different multilingual LLMs, we present additional results from LLAMA2-7B-CHAT in Figure 4. Our analysis re-

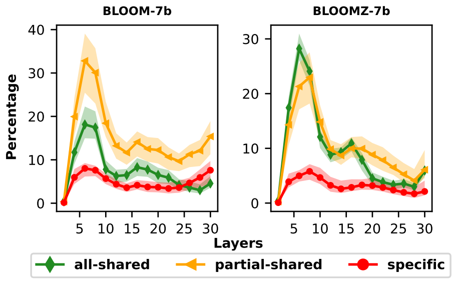


Figure 5: Comparison of neuron activations with foundation LLM BLOOM-7B (left) and instruction finetuned LLM BLOOMZ-7B (right).

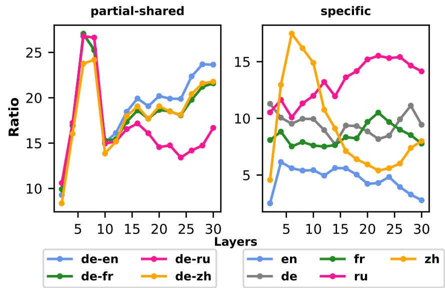


Figure 6: Neuron activation pattern across languages in the `Fact Probing` task with BLOOMZ-7B backbone. Left: The ratio of *partial-shared neurons* representing {en, fr, ru, zh} shared with German (de). Right: The percentage of {en, de, fr, ru, zh} in *specific neurons*.

veals that the activation patterns in LLAMA2-7B-CHAT differ significantly from those observed in BLOOMZ-7B, highlighting the variability across models. Notably, LLAMA2-7B-CHAT demonstrates a higher degree of neuron sharing, particularly for *partial-shared neurons*. This phenomenon can be attributed to the English-centric nature of LLAMA2-7B-CHAT. When processing multilingual inputs, the model heavily relies on knowledge transfer from English to other languages, resulting in a substantial number of *partial-shared neurons*. We also present additional results using XGLM (Lin et al., 2022) in Figure 9 of Appendix F, aligning with our observations.

Instruction finetuned LLMs exhibit larger proportion of the *all-shared neurons*. We conduct additional experiments using the foundation model BLOOM-7B to explore the impact of instruction finetuning on neuron activation patterns. As shown in Figure 5, the instruction-finetuned BLOOMZ-7B demonstrates a higher percentage of *all-shared neurons* compared to BLOOM-7B. This observation suggests that instruction finetuning may encourage neuron sharing within LLMs, potentially aligning their internal representations across languages. Therefore, instruction-finetuned LLMs, such as BLOOMZ-7B, generally outperform their foundational counterparts.

7.4 Neuron Activations Across Languages

Neuron sharing does not completely align with language similarity. We investigate the relationship between language similarity and neuron sharing by analysing the proportion of *partial-shared neurons* for language pairs involving German and several other languages on the `Fact Probing` task. As shown in Figure 6, our findings reveal that similar languages (e.g., German and French) do not always exhibit higher levels of neuron sharing. For instance, the proportion of *partial-shared neurons* between German and Chinese is nearly identical to that between German and French, despite German and French both belonging to the Indo-European language family, while Chinese belongs to the Sino-Tibetan language family. Furthermore, we observe no consistent pattern in the percentage of *specific neurons* across the languages studied, suggesting that neuron specialization may not directly correlate with language similarity. We leave further exploration of this phenomenon to future work. Additional results on the `XNLI` task are in Appendix H.

Furthermore, we conduct ablation studies to investigate the impact of two key factors on the neuron activation patterns: the size of the backbone model with 0.56b, 1b, 3b, 7b parameters (Appendix I), and the number of demonstrations in the few-shot setting (Appendix J).

8 Conclusion

In this study, we explored the complex mechanisms of neuron activation within multilingual LLMs, addressing the significant research gap in understanding these models beyond a monolingual context. We developed a fine-grained classification for analysing how neurons respond to different tasks and languages. We categorized neurons into four distinct groups: *all-shared*, *partial-shared*, *specific*, and *non-activated*. Our research revealed that neurons shared across all languages proved essential for generating accurate responses, highlighting their pivotal role in multilingual processing. Furthermore, we demonstrate that neuron sharing is task-related, and, it does not always align with language similarity. Our study improves the understanding of the internal workings of multilingual LLMs and fosters future research in this direction.

9 Limitations

In this paper, we develop a method to analyse neuron behaviors in detail by categorizing them into four distinct neuron types w.r.t the degree of their responses to input languages. Although this enables a fine granularity neuron analysis on LLM backbones across various linguistic characteristics and task complexity, the scope of the experiments can be extended to accommodate larger LLMs with large amounts of parameters (i.e., BLOOMZ-176B) on a more comprehensive range of tasks. While this study demonstrates that the number of languages slightly impacts the percentage of *all-shared neurons*, it is limited to nine languages. Exploring the effects of incorporating a larger number of languages into the proposed method warrants further investigation. Additionally, other network components, for example, attention heads, are not in the scope of this analysis.

References

Sunit Bhattacharya and Ondrej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. *CoRR*, abs/2310.15552.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5486–5513. Association for Computational Linguistics.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3046–3052. Association for Computational Linguistics.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Baochen Sun, Ruibo Liu, Daiyi Peng, Yawen Zhang, Xiaoyuan Guo, Jie Yang, and V. S. Subrahmanian. 2024. Higher layers need more lora experts. *CoRR*, abs/2402.08562.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 30–45. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6919–6971. Association for Computational Linguistics.

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, Felix Chern, Felix X. Yu, Ruiqi Guo, and Sanjiv Kumar. 2023. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin

Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024. Unraveling babel: Exploring multilingual activation patterns within large language models. *arXiv preprint arXiv:2402.16367*.

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

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Karolina Stanczak, Edoardo M. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1589–1598. Association for Computational Linguistics.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. *CoRR*, abs/2404.11201.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023. Retrieval-augmented multilingual knowledge editing. *arXiv preprint arXiv:2312.13040*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3320–3328.

Zeping Yu, Kailai Yang, Zhiwei Liu, and Sophia Ananiadou. 2023. Exploring the residual stream of transformers. *arXiv preprint arXiv:2312.12141*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Moefication: Transformer feed-forward layers are mixtures of experts. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 877–890. Association for Computational Linguistics.

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

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling A core linguistic region in large language models. *CoRR*, abs/2310.14928.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? *CoRR*, abs/2402.18815.

A Detailed Interpretation of Projection in Vocabulary Space

There is a residual connection in the each layer of transformer, where the hidden state is:

$$h^l = x^l + FFN^l(x^l) \tag{11}$$

In order to analyze the attribution of neurons, we explore how the output distribution in the vocabulary space changes when the representation x^l (before the FFN update) is added with the output of neurons $A_i^l v_i^l$. With the embedding matrix E, we map each vector into the vocabulary space ν. For each token w, the probability is calculate with the softmax function:

$$\begin{aligned} & p(w|x^l + A_i^l v_i^l, E) \\ &= \frac{exp(E_w \cdot x^l + E_w \cdot A_i^l v_i^l)}{Z(E(x^l + A_i^l v_i^l))} \\ &\propto exp(E_w \cdot x^l) \cdot exp(E_w \cdot A_i^l v_i^l) \end{aligned} \tag{12}$$

where E_w is the embedding of w, and $Z(\cdot)$ is the constant softmax normalization factor. The $E_w \cdot x^l$ can be viewed as a static score of w that is independent of the input to the model. Thus, the projection $E_w \cdot A_i^l v_i^l$ induces a ranking over the vocabulary. So we use the projection as effective score to detect the responsibility of neurons.

B Tasks

- `XNLI`. Natural Language Inference (Conneau et al., 2018) is a multilingual natural languages inference dataset, containing 5000 items. Each test sample consists of a premise and a hypothesis, requiring an LLM to determine whether a hypothesis is entailed, contradicted, or neutral conditioned on the premise.

- `Fact Probing`. LLMs are used to predict factual answers in response to corresponding probing prompts. A multilingual factual knowledge dataset (mParaRel (Fierro and Søgaard, 2022)) capturing 38 binary relations (e.g., *X born-in Y*) is used in the analysis. We seletc the relation of "capital" subset (*X capital Y*) as testset, including 348 items.

- `Cross-lingual Knowledge Editing (KE)`. MzsRE (Wang et al., 2023) is a multilingual question-answering dataset, containing 743 items for each language. It provides counterfactual edited knowledge in the context and requires an LLM to produce the corresponding answer according to the context. We evaluate LLMs in two Cross-lingual KE scenarios: 1) EN (Edit) → ALL (Test): edit in English and test in other languages and 2) ALL (Edit) → EN (Test): edit in other languages and test in English.

settings	pct.	en	de	es	fr	ru	th	tr	vi	zh
baseline	0%	59.1	47.6	50.1	47.0	49.1	41.4	40.2	51.6	46.1
w/o. all-shared	22.42%	**3.0**	**3.6**	**4.4**	**1.9**	**4.7**	**6.9**	**3.6**	**13.5**	**4.8**
w/o. partial-shared	17.48%	59.1	48.4	51.5	47.9	49.7	42.9	41.5	50.8	48.0
w/o. specific	4.75%	59.2	47.3	49.9	47.0	49.1	41.9	40.1	51.4	46.2
w/o. non-activated	55.35%	30.5	13.8	12.0	11.9	12.4	5.0	14.2	13.4	5.2
w/o. random	5%	58.7	47.7	50.2	48.2	49.0	41.7	40.0	49.9	45.7
	15%	52.7	44.6	47.2	46.4	44.5	38.4	40.1	48.6	45.2
	25%	46.1	42.4	41.3	43.3	40.1	34.5	39.7	38.7	40.7
	55%	28.7	30.2	28.6	30.3	25.8	19.0	27.1	28.2	25.0

Table 4: The accuracy in `XNLI` task with LLAMA2-7B-CHAT backbone when deactivating four types of neurons.

C Prompts

For the `Fact Probing` task, we use the P36 subtestset, which describe facts of entities in a relation of "capital". The prompt is framed as " The capital of {X} is " where "{X}" is the subject (sovereign state) and LLMs are required to predict the object (capital city). We keep at least three paraphrase prompts from mParaRel for each language to ensure a level of diversity.

For the Natural Language Inference (`XNLI`) task, we frame the prompt as " Take the following as truth: {premise} Then the following statement: '{hypothesis}' is 'true', 'false', or 'inconclusive'? "

For the `Cross-lingual KE` task, we format the prompt as " {context} Question: {question} Answer: ". The same language is used for the questions and the answers, but the context is in a different language.

D Supplemental Results on Deactivating Neurons

In order to further prove the importance of *all-shared neurons* across LLMs, we conduct the experiments with deactivating neurons on the `XNLI` task with LLAMA2-7B-CHAT backbone. The results in Table 4 show that there is more significant decline when *all-shared neurons* are deactivated. It demonstrates that *all-shared neurons* play a key role in predicting correct answers across LLMs.

E Generation Impact Score of Different Tasks

The Generation Impact Score of the four types of neurons evaluated on the `Cross-lingual KE (EN (edit) → ALL (Test))` and `XNLI` tasks across languages are shown in Figure 7 and Figure 8.

F Supplemental Results on Neurons Activation Patterns across LLMs

We further study the neuron activation patterns in another multilingual LLM (XGLM). The results of XGLM backbone are captured in Figure 9.

G Supplemental Results on Neurons Activation Patterns of Foundation LLM BLOOM-7B

We further explore the neuron activation patterns across various tasks in the foundation LLM (BLOOM-7B). The results of BLOOM-7B backbone are captured in Figure 10.

H Neuron Activation Across Languages on XNLI Task

We analyze the shared proportion of German with other languages in *partial-shared neurons* and the *specific neuron* ratios for each language derived from the `XNLI` task in Figure 11. The shared ratio of German with Russian (in different language family) is higher than the ratio of German with French (in the same language family), confirming the conclusion in Section 7.4.

I Influence of Model Scale

We investigate neuron activation patterns across the BLOOMZ series with 0.56b, 1b, 3b, 7b parameters in a `XNLI` task. As shown in the results captured in Figure 12, no identifiable pattern difference can be observed to indicate a scale law effect. However, the scale of the model is limited, potentially leading to unreliable results in this experiment. More *non-activated neurons* in the upper layers of BLOOMZ-7B may reflect on a higher level of sparsity for a larger LLM (consistent with Voita et al. (2023); Li et al. (2023)).

J Neuron Activation Patterns in Few-shot In-context Learning

According to Wang et al. (2023), in-context learning (ICL) can improve the performance of an LLM under the guidance of few-shot examples in a `Cross-lingual KE` task. We further explore the impact of few-shot examples on neuron activation patterns. We compare the results of an LLM with 0-shot, 2-shot, 4-shot, 6-shot examples in a `Cross-lingual KE (EN (edit) → ALL (Test))` task. Four types of neurons in scope have almost identical activation patterns across various few-shot examples (Figure 13). Although in-context examples lead to no observable neuron activation pattern changes, more examples lead to better performances. Could ICL lead to a better neuron activation composition instead of invoking more neurons? We leave this to a future study.

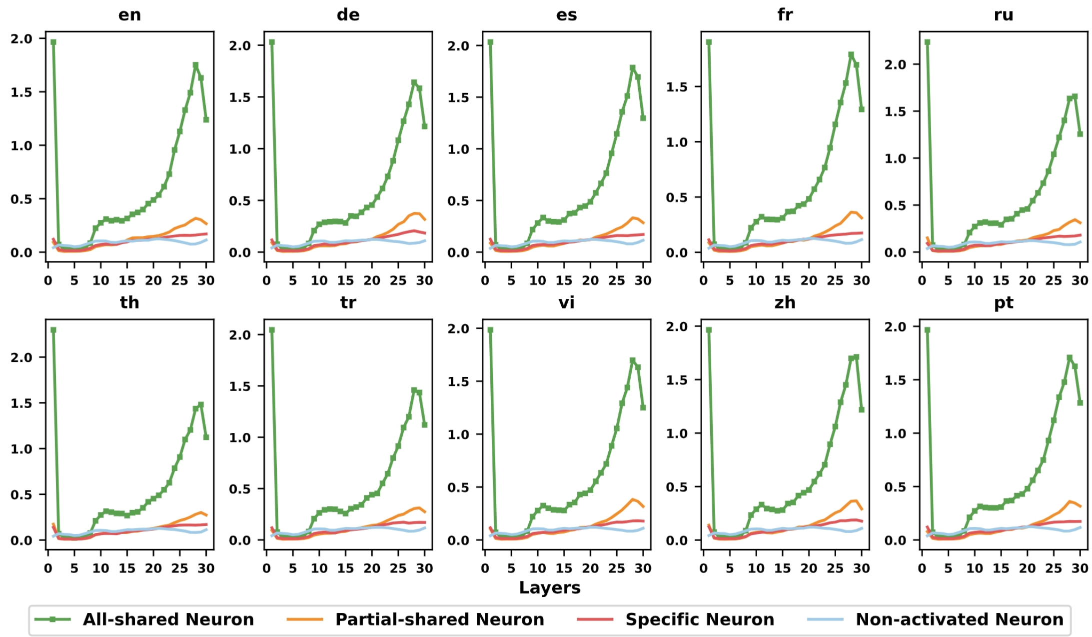


Figure 7: Generation Impact Score on the `Cross-lingual KE (EN (edit)` $\rightarrow$ `ALL (Test))` task with BLOOMZ-7B backbone.

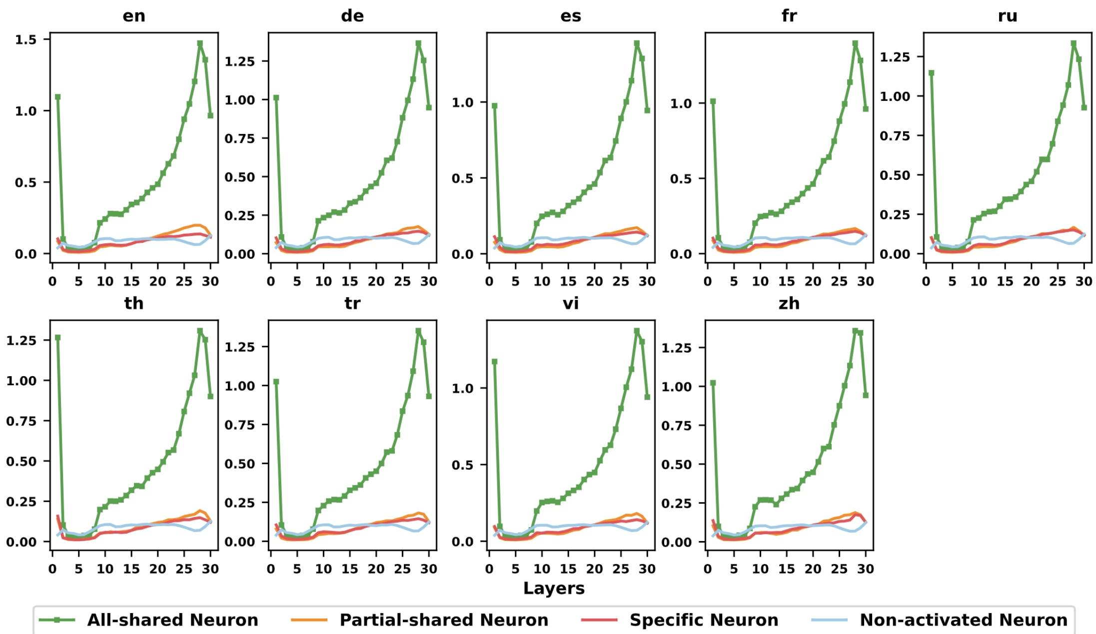


Figure 8: Generation Impact Score on the `XNLI` task with BLOOMZ-7B backbone.

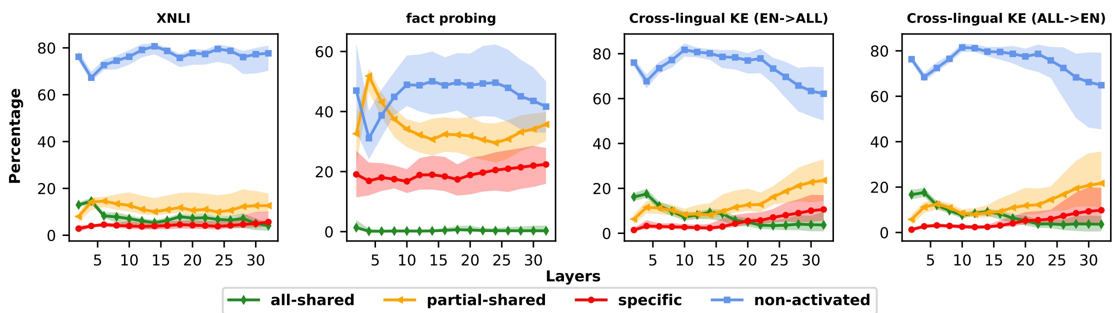


Figure 9: Neuron activation pattern in `XNLI`, `Fact Probing`, and `Cross-lingual KE` tasks with XGLM backbone.

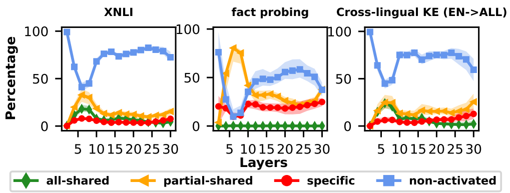


Figure 10: Neuron activation pattern in `XNLI`, `Fact Probing`, and `Cross-lingual KE` tasks with BLOOM-7B backbone.

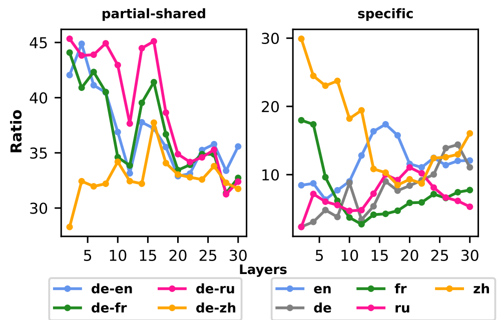


Figure 11: Aggregated neuron activation pattern across languages in the `XNLI` task. Left: The ratio of partially-shared neurons representing {en, fr, ru, vi} shared with German (de). Right: The percentage of {en, de, fr, ru, vi} in specific neurons.

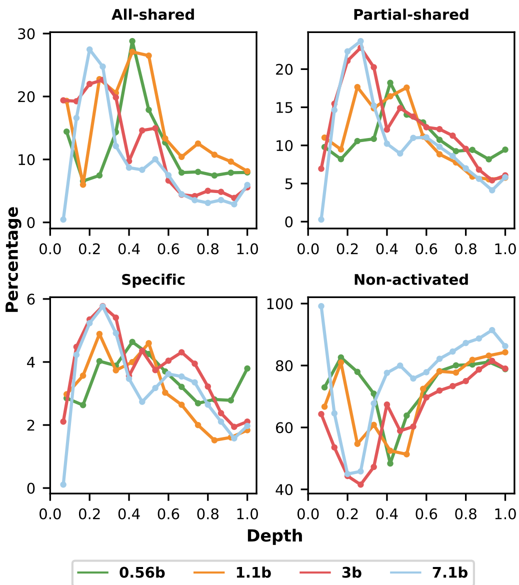


Figure 12: Neuron activation patterns in a `XNLI` task with the BLOOMZ size as 0.56b, 1b, 3b, 7b.

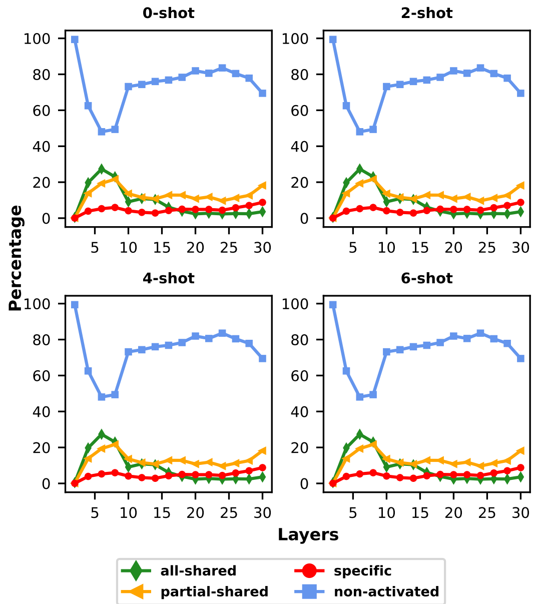


Figure 13: Neuron activation patterns in `Cross-lingual KE (EN (edit)` $\rightarrow$ `ALL (Test))` task with BLOOMZ-7B backbone under the in-context learning.