Sharing Matters: Analysing Neurons Across Languages and Tasks in LLMs

Anonymous ACL submission

Abstract

Multilingual large language models (LLMs) have greatly increased the ceiling of performance on non-English tasks. However, the mechanisms behind multilingualism in these LLMs are poorly understood. Of particular interest is the degree to which internal representations are shared between languages. Recent work on neuron analysis of LLMs has focused on the monolingual case, and the limited work on the multilingual case has not considered the interaction between tasks and linguistic representations. In our work, we investigate how neuron activation is shared across languages by categorizing neurons into four distinct groups according to their responses across different languages for a particular input: ***all-shared***, ***partial-shared***, ***specific***, and ***non-activated***. This categorization is combined with a study of neuron attribution, i.e. the importance of a neuron w.r.t an output. Our analysis reveals the following insights: (i) the linguistic sharing patterns are strongly affected by the type of task, but neuron behavior changes across different inputs even for the same task; (ii) *all-shared neurons* play a key role in generating correct responses; (iii) boosting multilingual alignment by increasing *all-shared neurons* can enhance accuracy on multilingual tasks. We will release the code to foster research in this area.

1 Introduction

The black-box nature of large language models (LLMs) has given rise to an area of research which aims to interpret the internal mechanism of the Transformer architecture (Elhage et al., 2021a; Yu et al., 2023). In order to investigate specific aspects of model behavior, previous studies choose to focus on specific model components to encourage interpretability. Differently from e.g. attention heads in the Transformer layers which are responsible for moving information from one token to another token (Elhage et al., 2021b), feed-forward networks

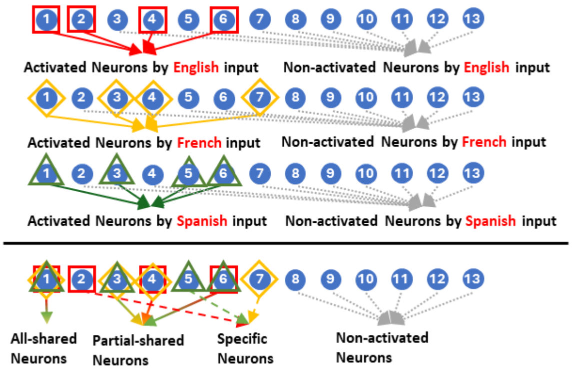


Figure 1: A comparison of neuron analysis with different role designs in multilingual settings with the same semantic input, in which we define four types of neurons in one layer of LLM. English input "The capital of France is" corresponds to its Spanish and French counterparts "La capital de Francia es" (Spanish), "La capitale de la France est" (French).

(FFNs) are more likely to represent semantic features (Black et al., 2022). The FFN consists of two linear layers and an activation function is applied between them. 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). Geva et al. (2020, 2022); Ferrando et al. (2023) have earlier demonstrated the importance of neurons in the FFNs for encoding factual and linguistic knowledge. Furthermore, various neuron analytic methods have been developed to prune inactive neurons, i.e. those which have activation value less than or equal to zero (Zhang et al., 2022; Li et al., 2023; Voita et al., 2023).

While the above-mentioned methods analyze neuron behaviors based on the activation state (active or inactive means activation value > 0 vs. ≤ 0) in a monolingual setting, there is still a lack of understanding regarding how neurons behave across various tasks in a multilingual setting. As shown in Figure 1, different sets of neurons are activated when an LLM is presented with corresponding inputs in different languages. An inactive neuron in response to an English input is often activated and influential in response to an input in another language. It remains unclear how LLMs actually manage to learn multilingual representations. One typical example is that we may need to explore sets of neurons responsible for generating the output "Paris"[1] when presented with the English input "The capital of France is" and its Spanish counterpart "La capital de Francia es".

In a similar line of reasoning, Bhattacharya and Bojar (2023) explored the behavior of language-specific and language-agnostic (shared between two languages) neurons and their distribution over layers of a model. They concluded that the layers close to the model's input and output exhibit more language-specific behavior. However, their results are restricted to the English-Czech. Due to the potential variation in neuron behavior of different language types and the reasoning and memory abilities of different multilingual tasks, we argue that there is a need for further exploration.

Our research aims to establish a more fine-grained classification of neurons, enabling a more detailed analysis of LLM behavior in reasoning style tasks (XNLI), fact-retrieval based tasks (fact probing), and explicitly multilingual question-answer tasks (knowledge editing). As shown in Figure 1, we reformulate the activation state of neurons for a particular input and its corresponding translations in 10 languages to four distinctive types to represent multilingual behaviors. ***All-shared neurons*** are neurons that remain active for all inputs regardless of language. ***Partial-shared neurons*** are activated only for inputs in certain languages. ***Specific neurons*** are activated exclusively for inputs in one language. ***Non-activated neurons*** are not activated at all for any input.

We perform neuron analysis addressing two research questions: (a) what are the behaviors of the four types of neurons in various multilingual tasks? (b) What attribution (Dhamdhere et al., 2019) does each type have in multilingual generation tasks, meaning which neurons are responsible for a prediction?

We examine the percentage of each type of neuron and neuron attributions in the multilingual tasks. We discover that the pattern of four types of neurons is determined by the tasks they encountered, and that the behavior of a neuron changes with different inputs for the same task, indicating the potential implications of pruning neurons.

Furthermore, we demonstrate the importance of *all-shared neurons* in generating the correct output from the neuron attribution study. Converting other types of neurons to *all-shared neurons* improves the accuracy of an LLM in multilingual tasks.

Our main contributions are listed below:

- **Fine-grained neuron analysis**: We define four categories of neurons and use these to analyze neuron behaviors in various types of tasks across 10 languages. We reveal that reasoning style tasks (e.g. XNLI) involve more *all-shared neurons* than fact-retrieval based tasks (e.g. fact probing) which utilize more *specific neurons*.

- **Neuron attribution**: By studying the contribution of neurons to the output of multilingual tasks, we are the first to reveal the importance of *all-shared neurons* within FFNs in multilingual tasks. For instance, for the XNLI task, the *all-shared neurons* comprise less than 30% of the neurons, but they contribute 91.6% to the generation of the correct responses in the German test set.

- **Multilingual alignment**: We demonstrate that increasing the percentage of *all-shared neurons* (by converting other types of neurons) can significantly enhance the accuracy of an LLM in multilingual tasks.

2 Related Work

Prior interpretability studies focused on understanding attention heads, while others have analyzed neuron behaviors. 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 disclosed 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 are focused exclusively on a monolingual setting.

For multilingual model analysis, Bhattacharya and Bojar (2023); Tang et al. (2024); Tan et al.

[1]"Paris" is identical in both English, Spanish and French

(2024) classified neurons in an FFN block to language-specific and language-agnostic based on the threshold, which presumed that a neuron's behavior remained consistent across examples. However, they do not consider the potential adaptation of neurons under various language types and semantics brought forth by inputs from various multilingual tasks. We investigate neurons' behaviors across multiple languages and tasks to this end.

3 Definitions

3.1 Neurons in FFN Blocks

Every feed-forward block at layer l involves two linear transformations separated by a point-wise activation function. Biases are omitted for the sake of clarity:

$$FFN^l(x^l) = Act(W_K^l x^l)W_V^l \quad (1)$$

where $W_K^l, W_V^l \in \mathbb{R}^{d_m \times d}$ 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. We define the behavior of a neuron to be the output of function immediately after the elementwise nonlinearity. 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 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:

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

When such a neuron is activated $A_i^l > 0$, so it updates the residual stream by pulling out the corresponding value v_i^l.

3.2 Contribution Score and Effective Score of Neurons

Inspired by Geva et al. (2022), in order to judge the importance of neurons in generating answers, we analyze their contributions to the output. The contribution score of a neuron to an FFN output is:

$$C_i^l := \frac{|A_i^l| \; ||v_i^l||}{\sum_{j=1}^{d_m} |A_j^l| \; ||v_j^l||} \quad (3)$$

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.

Whenever a neuron is activated, the associated column of the values (scaled by the neuron's value) is incorporated into the residual stream. The product of the value of the activated neuron A_i^l and the corresponding v_i^l is then transformed linearly and mapped to the vocabulary.

Following Geva et al. (2022); Voita et al. (2023), projecting the neuron to the vocabulary using embedding matrix $E_r \cdot A_i^l v_i^l$ can be viewed as obtaining the effective score given by the i-th neuron to the output reference token r for a given input. Specifically, a larger $E_r \cdot A_i^l v_i^l$ has a higher probability to produce a gold answer (r). A negative $E_r \cdot A_i^l v_i^l$ reduces the probability in generating r. In this way, we can quantify the effect of a neuron on the output distribution. We give detailed descriptions about the neuron projection to the vocabulary space in Appendix A.1.

3.3 Definition of Four Types of Neurons

For the set of all neurons N^l in the $l-$th layer, the activation value of one neuron n in one language $lang$ is A_{lang}^l. Note that some activation functions (e.g. GeLU) can result in negative activation values. The definition of ***all-shared neurons*** is:

$$N_{all}^l := \bigcap_{lang}^{langs} \{n \in N^l : A_{lang}^l > 0\} \quad (4)$$

where $langs$ is the sets of testing languages and N_{all} means these neurons are activated in all languages. For ***non-activated neurons*** which have activation value less than or equal to zero in all languages, the definition is:

$$N_{non}^l := \bigcap_{lang}^{langs} \{n \in N^l : A_{lang} \leq 0\} \quad (5)$$

Specific neurons are neurons only activated in one specific language. They can be denoted using:

$$N_{specific}^l := \bigcup_{lang_{k1}}^{langs} \left(\{n \in N^l : A_{lang_{k1}} > 0\} \bigcap_{\substack{lang_k \\ k \neq k1}}^{langs} \{n \in N^l : A_{lang_k} \leq 0\}\right) \quad (6)$$

The remaining neurons are ***partial-shared neurons*** as they are activated by inputs from multiple languages, but not all languages at the same time:

$$N_{partial}^l = N^l - N_{all}^l - N_{non}^l - N_{specific}^l \quad (7)$$

Unlike Bhattacharya and Bojar (2023); Tang et al. (2024), which exclusively on sub-word activation statistics and thus capture incomplete semantics, we examine the activation state of the last token. For each input text with tokens $x_1, x_2, ...x_S$, we use the activation state x_S to investigate the behavior of neurons, 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. They are:

`Natural Language Inference`. XNLI (Conneau et al., 2018) is a multilingual natural languages inference dataset. 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.

`Cross-lingual Knowledge Editing (KE)`. MzsRE (Wang et al., 2023) is a multilingual question-answering dataset. 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.

These tasks cover 10 diverse languages, including English (en), German (de), Spanish (es), French (fr), Portuguese (pt), Russian (ru), Thai (th), Turkish (tr), Vietnamese (vi), and Chinese (zh). Prompts are detailed in Appendix A.2.

4.2 Base LLMs

We mainly analyze the behavior of neurons in a foundation multilingual LLM `BLOOM` (Scao et al., 2022) and an instruction-finetuned model `BLOOMZ` (Muennighoff et al., 2023). We also include the analysis of other decoder-only models: `XGLM` (Lin et al., 2022), `LLAMA2-7b-chat` (Touvron et al., 2023), and an encoder-decoder model `mT0` (Muennighoff et al., 2023) in the Appendix A.3.2.

5 Behaviors of Four Types of Neurons

For each task we use parallel test texts from ten languages as inputs, and record the activation state of each neuron. Subsequently, we calculate the percentage of the four types of neurons compared to the total neurons.

We use BLOOMZ as the backbone LLM to investigate the behaviors of four types of neurons, and the results of three tasks are shown in Figures 2-4 respectively. The supplemental analysis of cross-lingual KE (ALL (Edit) → EN (test)) task is shown in Figure 11 (Appendix A.3.1).

5.1 Neuron Behaviors Across Tasks

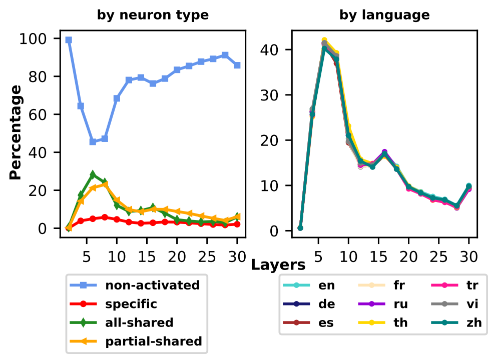


Figure 2: Neuron behavior pattern in the XNLI task. The left sub-figure shows the percentage of the four types of neurons across transformer layers (aka "by neuron type"). The right sub-figure shows aggregated behaviors of the activated neurons for each language across transformer layers (aka "by language").

5.1.1 Behavior Pattern of Four Types of Neurons

It can be observed from Figures 2-3 (left sub-figures) that there are more *non-activated neurons* on the whole than other types. However, **the neuron behavior is strongly task-related** as the pattern observed in the fact probing task differs significantly from the other two tasks. In the fact probing task, there are more *partial-shared neurons* (yellow line in Figure 4), whereas the other tasks involve far more *all-shared neurons* (green line in Figures 2-3).

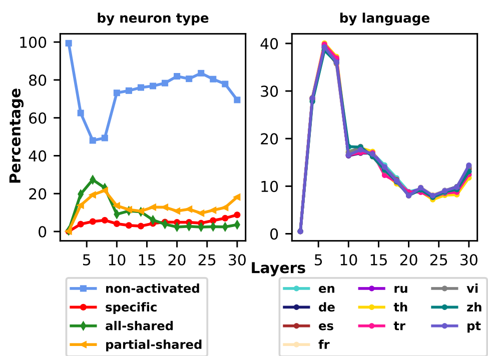


Figure 3: Neuron behavior pattern in the cross-lingual KE (EN (Edit) → ALL (test)) task.

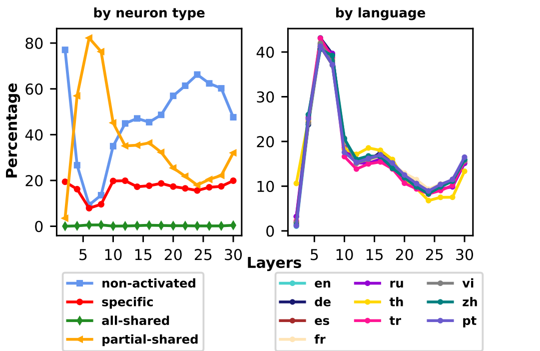


Figure 4: Neuron behavior pattern in the fact probing task.

For the neuron behaviors across layers, lower layers exhibit a higher prevalence of *all-shared neurons* compared to *specific neurons* and *partial-shared neurons* in XNLI and KE. The number of *all-shared neurons* peaks at a certain layer followed by a continuous decreasing pattern in these two tasks. Fewer *all-shared neurons* in the upper layers implies language-specific characteristics are retained there. Similarly, *partial-shared neurons* accumulate in the lower layer and it tends to outnumber all other neurons moving towards upper layers in all tasks. It could be observed that nearly 99% of the neurons are non-activated in the first layer for XNLI and KE tasks, which may be associated with the prompts, where the last token of input is punctuation (e.g., "?"). However, this phenomenon appears to be specific to BLOOMZ compared to other LLMs (Appendix A.3.2). We also investigate the impact of the number of languages on the percentage of four types of neurons in Appendix A.3.3. The comparison in Figure 16 indicates that the number of languages slightly affects the *all-shared neurons*.

5.1.2 Consistent Neuron Behavior Pattern

The percentage of activated neurons for each language exhibits a consistent pattern, as shown in Figures 2-4 (right sub-figures). At lower layers of an LLM, the number of activated neurons increases significantly, reaches the peak at around the 6-th layer, and then declines. It is not until at an upper layer (i.e., 28-th layer) that the number of active neurons commences to pick up its early increasing trend. Such a resurgence continues until it reaches the final layer of the LLM. It is a surprise to discover that the number of activated neurons is not influenced by the language of the inputs. This indicates **neurons in an LLM exhibit similar patterns across languages**.

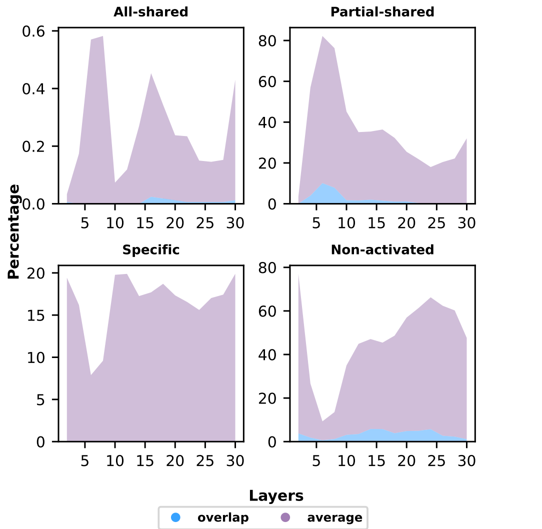


Figure 5: Behavior-repeating neurons in the fact probing task across the entire testset. "overlap" indicates the percentage of neurons that keep the same behaviors across all examples in the testset. "average" indicates the percentage of neurons for an input, which is averaged over the entire testset.

5.1.3 Neurons Behaviors Across Examples

As shown in Figures 2-4, the analysis is conducted by exploring the activation state under the same input and the corresponding translations in 10 languages. In this experiment, we now investigate how neurons behave across all examples for the fact probing task. Here, the objective is to understand if neurons that are specific to a designated behavior maintain this behavior across test samples representing different semantics. The results are shown in Figure 5 and in Appendix A.3.4 (for other tasks). To our surprise, almost no neuron (identified by its index) repeats its behavior (e.g., the "specific") across all examples. Different from the previous studies, which assumed that a neuron behaved consistently across examples, we reveal that **the behavior of a neuron is determined by the semantics of the inputs encountered, even within the same task**.

5.1.4 "Dead" Neuron Mystery

In the multilingual scenario, *non-activated neurons* comprise a significant portion, with more than 50% of neurons having a zero or negative activation value across layers for multilingual inputs (blue line in the left sub-figures) in Figures 2-4.

Do these "dead" neurons stay inactive in all test samples? In Figure 5 we see that less than 10% of *non-activated neurons* remain inactive. There is only a small proportion of persistently inactive neurons, which reflects the distributed nature of knowledge representation for LLMs. The majority of neurons are activated at some point depending on the input provided. That is why we should execute caution when pruning neurons, as we may damage the overall performance of an LLM.

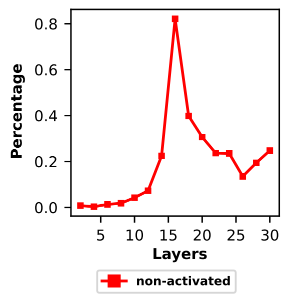


Figure 6: Neurons remaining "dead" in response to all tokens in an input sentence for the fact probing task.

Do non-activated neurons remain "dead" in response to each token in an input sentence? Previously, we performed analysis for the last token of an input. We analyze the behavior-repeating of *non-activated neurons* across tokens here for an input. Specifically, for each input sequence with tokens $x_1, x_2, ..., x_S$, we record *non-activated neurons* of each token, w.r.t the intersection of index. A neuron is counted when it stays inactive for every token of one test input sequence. As shown in Figure 6, less than 0.8% *non-activated neurons* remain "dead" for each token of input, further emphasizing the behavior-repeating of a neuron.

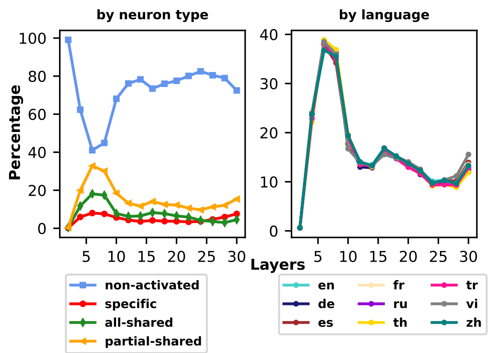


Figure 7: Neuron behavior pattern in the XNLI task with the BLOOM backbone.

5.2 Influence of Instruction Finetuning

Does instruction finetuning (IFT) have an impact on neuron behaviors? We compare the percentage of the four designated types of neurons of BLOOM and its IFT counterpart BLOOMZ.

The results from the XNLI task are shown in Figure 7 and Appendix A.3.5 (from other tasks). Compared to the results from BLOOMZ in Figure 2, *all-shared neurons* are under-represented in BLOOM (20% of BLOOM vs. 30% of BLOOMZ). Meanwhile, more *partial-shared neurons* are observed in BLOOM. IFT enhances the percentage of *all-shared neurons* and reduces the number of *partial-shared neurons*. We regard the increase of the number of *all-shared neurons* as an effect of multilingual representation alignment. **It appears that IFT contributes to multilingual alignment** based on the effects observed from this experiment. Whether this effect can be generalized to other LLMs and the rationale behind such effect warrants a future study.

Furthermore, we conduct ablation studies to investigate the impact of two key factors on the neuron behaviors: the size of the backbone model (Appendix A.3.6) and the number of multilingual demonstrations in the few-shot setting (Appendix A.3.7).

6 Neuron Attributions

In the preceding experiments, we examined the proportions of the four types of neurons across layers, tasks, and languages. In this section, we examine the relative contributions of each neuron type to task performance.

6.1 Neuron Contribution Score

As discussed in Section 3.2, the contribution score C_i^l of a neuron refers to its relative weight compared to the total sum of weights of all neurons, indicating the influence of each neuron on outputs. We examine the proportion of the four types of neurons among the top 5% contribution score under inputs in each language. The proportions of neurons in the cross-lingual KE task are depicted in Figure 8 (and the overall results of 10 languages are shown in Figure 27 in Appendix A.4.1).

It can be observed that ***all-shared neurons* are the top contributing neurons to the outputs at every layer**, regardless of their language inputs. This highlights their importance in the neural network. The group of *partial-shared neurons* is the second most influential group, demonstrating their impacts across the latter half of the model. It is not surprising that the *specific neurons* group has limited influence on cross-lingual KE outputs as they feature in a particular language type of inputs.

Furthermore, we evaluate the contribution proportion in the XNLI task, as depicted in Figure 26

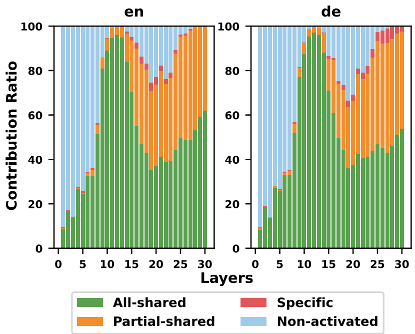
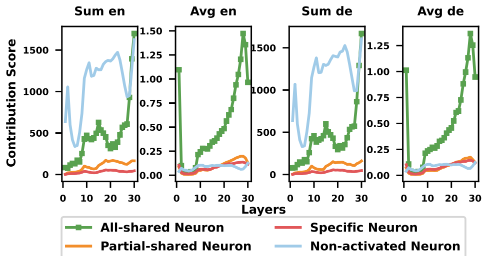
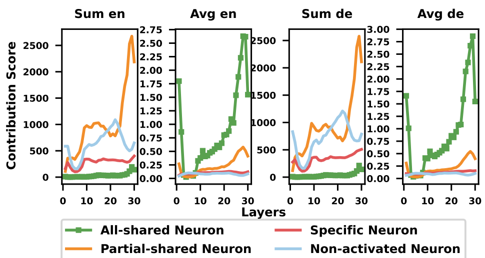


Figure 8: Contribution proportion of four types of neurons based on the cross-lingual KE (EN (Edit) → ALL (test)) task.

(Appendix A.4.1). Here, *all-shared neurons* constitute the highest proportion; however, *partial-shared neurons* show less influence compared to that observed in the cross-lingual KE task.



(a) XNLI task.



(b) Fact probing task.

Figure 9: Average and sum contribution score of the four types of neurons.

In order to more comprehensively analyze the overall contribution of four types of neurons, we extend the analysis from the top 5% to accommodate all neurons in this study. We analyze the average and the sum of contribution scores of all neurons in four categories. As shown in Figure 9, the average contribution of *all-shared neurons* significantly exceeds that of the other three types. In terms of the total contribution score, *all-shared neurons* in the upper layers achieve a value equal to that of *non-activated neurons* in Figure 9(a), despite having a significantly lower count than *non-activated neurons* (<10% vs. 80% in Figure 2). In the fact probing task, *partial-shared neurons* score the highest, while *all-shared neurons* score the lowest, primarily due to their respective counts (>20% vs. <1% in Figure 4). In summary, ***all-shared neurons* play a significant role in contributing to multiple tasks**. Future studies on neuron activation should consider their contribution as well as their frequency.

language	neuron type	max	min	avg
en	all-shared	1.85	-0.94	0.07
	partial-shared	0.22	-0.16	0.00
	specific	0.02	-0.02	0.00
	non-activated	0.04	-0.03	0.00
de	all-shared	1.03	-0.60	0.02
	partial-shared	0.13	-0.13	0.00
	specific	0.07	-0.03	0.00
	non-activated	0.02	-0.01	0.00

Table 1: Maximum, minimum, average effective score of four types of neurons on the cross-lingual KE task.

6.2 Neuron Effective Score

Since each neuron encodes information, we can explore their contributions to the correct answer. As discussed in Section 3.2, the projection to vocabulary $E_r \cdot A_i^l v_i^l$ can be viewed as the effective score given by the i-th neuron to the output reference token r for a given input.

In the cross-lingual KE (EN (Edit) → ALL (test)) task, we calculate the effective score of each type of neuron with the BLOOMZ backbone. The maximum, minimum, and average scores are shown in Table 1 (the overall results of ten languages are shown in Table 4), and the maximum effective scores across layers are shown in Figure 10. *All-shared neurons* achieve the highest maximum score and the lowest minimum score, indicating they are pushed (or eliminated) strongly by the activation function. In contrast to *all-shared neurons* and *partial-shared neurons*, where the maximum scores are substantially higher than the minimum scores (1.85 vs. -0.94 and 0.22 vs. -0.16 in English), for *specific neurons* and *non-activated neurons*, the score has a smaller span (± 0.07), suggesting ***all-shared neurons* have greater influence on the output distribution**.

As shown in Figure 10, the maximum effective score of *all-shared neurons* significantly exceeds that of the other three types. Moreover, *all-shared neurons* aggregate at the first layer and upper layers. This observation confirms a recent finding that early layers detect shallow patterns and upper layers are characterized by semantic patterns (Ferrando et al., 2023), supporting the notion that ***all-shared neurons* play a key role in generating prediction**. Details of effective score analysis in other tasks are depicted in Appendix A.4.2.

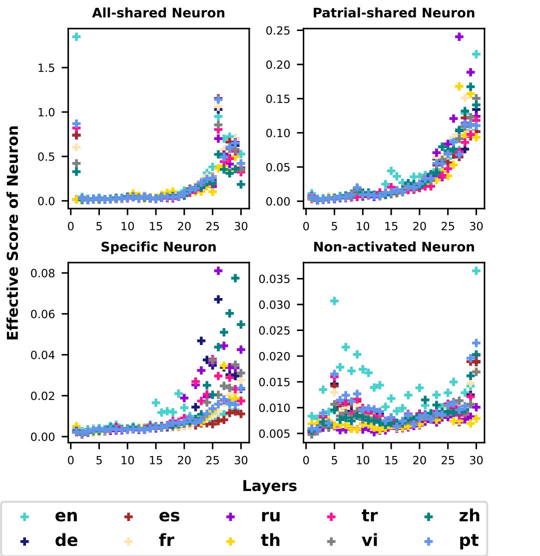


Figure 10: Maximum effective score of four types of neurons across layers based on the cross-lingual KE task.

accuracy	en	de	es	fr	ru	th	tr	vi	zh
baseline	53.8	41.8	50.3	49.0	47.6	40.9	34.9	50.5	51.1
w/o. all-shared	**16.7**	**3.50**	**10.1**	**10.0**	**6.6**	**9.0**	**1.4**	**12.1**	**14.5**
w/o. specific	53.7	41.7	50.3	48.9	47.4	40.6	35.3	50.4	49.3
w/o. partial-shared	52.9	40.4	49.7	47.6	49.2	40.3	36.1	50.0	50.0

Table 2: The accuracy when deactivating *all-shared neurons*, *specific neurons*, and *partial-shared neurons*, respectively. "w/o." stands for "without".

6.3 Effects on Accuracy

We have investigated neuron attribution using the contribution score and effective score, which are based on the internal states of an LLM. Direct impacts of neurons on the output performance (i.e., accuracy) provide insights from a pragmatic perspective. We explore the change of accuracy by intentionally deactivating three distinct types of neurons. This is performed by precisely identifying the type of neurons for an input and setting their activation values to zero. Take the XNLI task as an example, we record the results of an LLM with deactivated neurons in Table 2. **The most significant decrease in accuracy is observed when the *all-shared neurons* are deactivated** (e.g. 91.6% decrease in the German testset). Deactivating *specific* and *partial-shared neurons* also negatively impacts the accuracy, but at a smaller magnitude when compared to the effect by *all-shared neurons*. We also prove the importance of *all-shared neurons* with the LLAMA backbone in Table 5 (Appendix A.5.1). In order to prove the key role of *all-shared neurons* across tasks, we conduct the ablation experiments on the cross-lingual KE task in Appendix A.5.2.

accuracy	en	de	es	fr	ru	th	tr	vi	zh
baseline	53.8	41.8	50.3	49.0	47.6	40.9	34.9	50.5	51.1
co. specific	53.0	**42.7**	50.8	**50.1**	46.9	40.1	34.4	**51.0**	51.3
co. partial-shared	53.2	41.8	51.0	49.3	**49.4**	41.5	**36.8**	50.7	51.2
co. specific and partial-shared	52.6	42.1	**51.6**	49.4	48.2	**41.7**	36.4	50.7	**51.7**

Table 3: The accuracy when converting *specific neurons*, and *partial-shared neurons* to *all-shared neurons* type. "co." (covert) means setting the activation value of inactive portions of neuron types to above the activation threshold.

We are intrigued by the possibility of enhancing an LLM's performance by simply converting *specific neurons* and *partial-shared neurons* to *all-shared neurons*. To achieve this, we set the activation value of the inactive portion of these two types of neurons to just above the activation threshold. The results shown in Table 3 indicate that **increasing the number of *all-shared neurons* improves the accuracy** of an LLM in most cases. The results are consistent with the finding in Section 5.2 where IFT model BLOOMZ with better performance has more *all-shared neurons* than foundation model BLOOM. The occasional drawbacks in some cases (i.e., "en") require an in-depth study of the internal mechanism of an LLM involving more neuron components. We leave this to a future study.

7 Conclusion

In this paper, we propose a novel approach for analyzing neurons in FFN blocks by categorizing them into four distinct types: *all-shared neurons*, *partial-shared neurons*, *specific neurons* and *non-activated neurons*. We conduct a detailed analysis of LLM behaviors in multilingual tasks. Experimental results disclose novel insights relating to neuron behaviors: 1) We demonstrate that a neuron's activation pattern is influenced by the tasks it encounters, and even the behavior of a neuron changes with different inputs for the same task. 2) We show the importance of *all-shared neurons* in output generation in multilingual tasks from the neuron attribution study. 3) We prove that multilingual alignment can significantly enhance the accuracy of an LLM in multilingual tasks by increasing the percentage of *all-shared neurons* (i.e., via converting other types of neurons or via IFT). Future work will focus on exploring internal activation mechanisms underpinning the observed importance of *all-shared neurons* and multilingual alignment across a wider range of tasks and LLMs.

Limitations

In this paper, we develop a method to analyze 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 10 languages. Exploring the effects of incorporating a larger number of languages into the proposed method warrants further investigation. *All-shared neurons* in FFN blocks are identified to be of great importance, but how and why they work is still a mystery to disclose. Other network components, for example, induction 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.

Sid Black, Lee Sharkey, Léo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. 2022. Interpreting neural networks through the polytope lens. *CoRR*, abs/2211.12312.

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.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2019. How important is a neuron. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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. 2021a. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

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. 2021b. 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.

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*.

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.

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.

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.

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*.

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

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.

A Example Appendix

A.1 Detailed Interpretation of Projection in Vocabulary Space

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

$$h^l = x^l + FFN^l(x^l) \tag{8}$$

In order to analyze the behaviors 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:

$$\begin{aligned} & p(w|x^l + A_i^l v_i^l, E) \\ & = \frac{exp(E_w \cdot x^l + E_w \cdot A_i^l v_i^l)}{Z(E(x^l + A_i^l v_i^l))} \\ & \propto exp(E_w \cdot x^l) \cdot exp(E_w \cdot A_i^l v_i^l) \end{aligned} \tag{9}$$

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.

A.2 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 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.

A.3 Supplemental Results of Neuron Behavior Analysis

A.3.1 Neuron behaviors in additional cross-lingual KE task

We already show the neuron behaviors in in cross-lingual KE (EN (edit) → ALL (test)) task in Section 5.1.1. We supply the analysis in the cross-lingual KE (ALL (edit) → EN (test)) task in Figure 11.

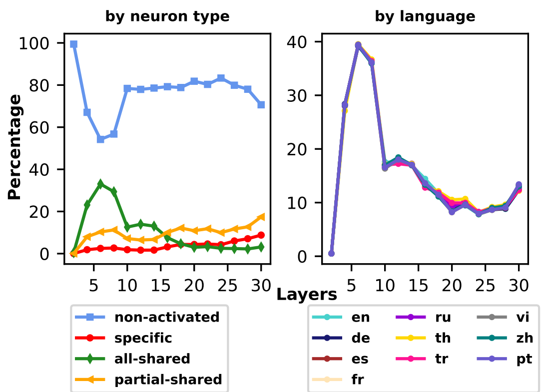


Figure 11: Neuron behavior pattern in cross-lingual KE (ALL (edit) → EN (test)) task.

A.3.2 Neurons Behaviors across LLMs

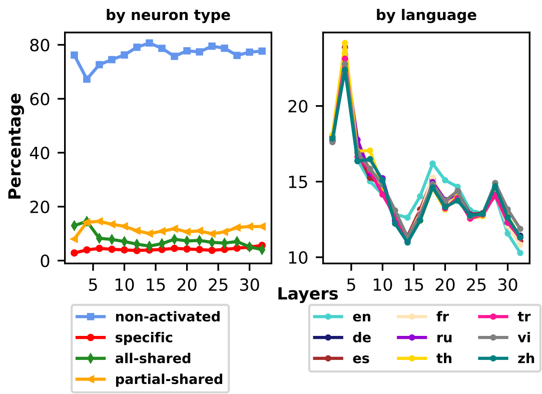


Figure 12: Neuron behavior pattern in the XNLI task with XGLM backbone.

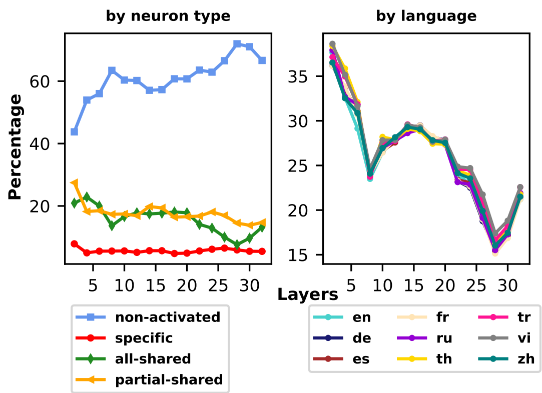


Figure 13: Neuron behavior pattern in the XNLI task with LLAMA2-7b-chat backbone.

Do the above-mentioned neuron behaviors change over different LLMs? We further study the neuron behaviors in other decoder-only multilingual LLMs (XGLM and LLAMA2) and an encoder-decoder multilingual LLM mT0. For the XNLI task, the results of XGLM and LLAMA2-7b-chat backbones are captured in Figure 12 and Figure 13, and the results of mT0-encoder and mT0-decoder are shown in Figure 14 and Figure 15. Furthermore, the percentage of four types of neurons for inputs in each language on each LLM demonstrates the similar pattern, indicating that neurons remain consistent behaviors across LLM backbones.

The number of active neurons increase in the lower layers, followed by a decrease moving onwards and a rise in the upper layers, despite the absolution values are different from those obtained for BLOOMZ. Moreover, it could be observed that encoder in mT0 involves more all-shared neurons compared to the proportion in decoder.

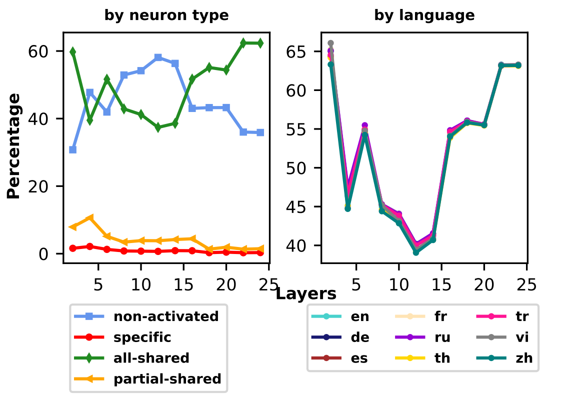


Figure 14: Neuron behavior pattern in the XNLI task in mT0-encoder.

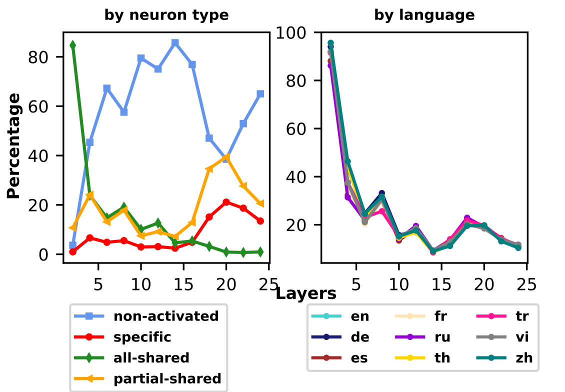


Figure 15: Neuron behavior pattern in the XNLI task in mT0-decoder.

A.3.3 Influence of the Quantity of Languages

We design an ablation study to investigate the influence of the number of languages (i.e., in a series of 3, 5, 7, 9). The results of the XNLI task in response to the quantitative change in languages are illustrated in Figure 16. The rise in the number of languages is associated with an observable increase in the percentage of *partial-shared neurons* and a slight decrease in the percentage of *all-shared neurons*. This suggests that the quantity of languages has a minor impact on the *all-shared neurons*.

A.3.4 Neurons Behaviors Across Examples for Other Tasks

We conduct additional experiments on the XNLI task and cross-lingual KE (EN (edit) $\rightarrow$ ALL (test)) task to investigate the behavior repeating for neurons. The results shown in Figures 17- 18 are consistent with the results of the fact probing task, where only a few neurons maintain the same behaviors across all examples in the testset.

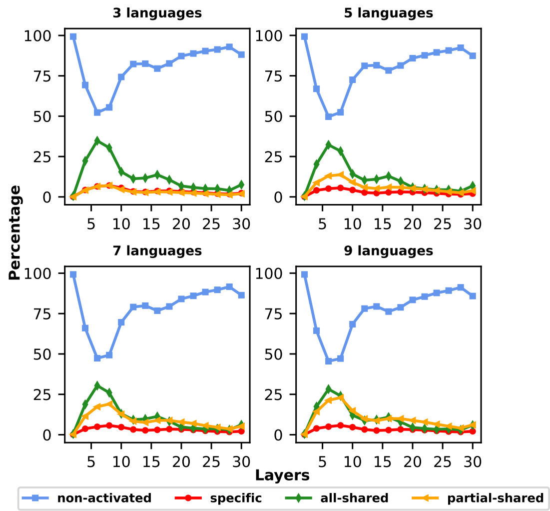


Figure 16: Effects of the number of languages based on the XNLI task.

A.3.5 Activation percentage of BLOOM

The results analyzed base on the foundation model BLOOM of fact probing task and cross-lingual KE task are shown in Figures 19 - 21.

A.3.6 Influence of Model Scale

We investigate neuron behaviors across the BLOOMZ series with 0.56b, 1b, 3b, 7.1b parameters in a XNLI task. As shown in the results captured in Figure 22, 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-7.1b may reflect on a higher level of sparsity for a larger LLM (consistent with Voita et al. (2023); Li et al. (2023)).

A.3.7 Neuron Behaviors 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 behaviors. We compare the results of an LLM with 0-shot, 2-shot, 4-shot, 6-shot examples in a cross-lingual KE (EN (edit) $\rightarrow$ ALL (test)) task. Four types of neurons in scope have almost identical behaviors across various few-shot examples (Figure 23). Although **in-context examples lead to no observable neuron behavioral changes**, more examples lead to better performances. Could ICL lead to a better

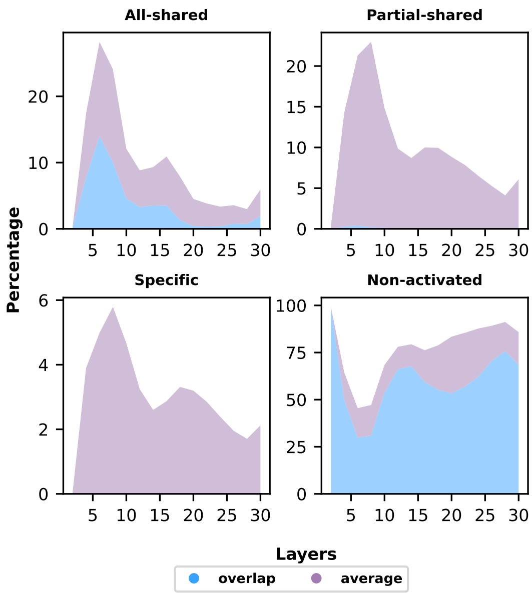


Figure 17: Behavior-repeating neurons in XNLI task across the entire testset.

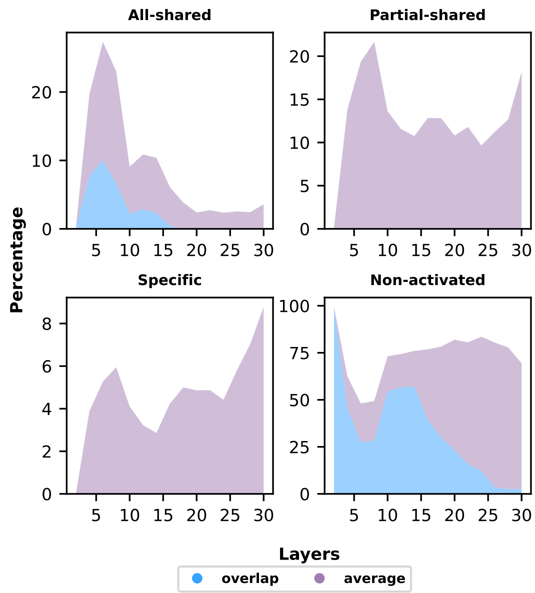


Figure 18: Behavior-repeating neurons in cross-lingual KE (EN (edit) $\rightarrow$ ALL (test)) task across the entire testset.

neuron activation composition instead of invoking more neurons? We leave this to a future study.

A.3.8 Activation Value across Layers

According to Eq. 1, neurons with larger activation values tend to contribute more to the output. By visualizing neuron activation for a series of thresholds $[0, 0.1, 0.2, 0.3, 0.4, 0.5]$, we can scrutinize the relative importance of various types of activated neurons. The percentage of various types of activated neurons for each activation threshold in the XNLI task are shown in Figure 24. When applying a threshold (>0), there are fewer all-shared, partial-shared, and *specific neurons* left in the lower layers (layers 0-10) compared to using a lower threshold (i.e., = 0). Under the same threshold scenario, more activated neurons appear in the second half of the model (layers 15-30). It is worth noting these neurons have a higher activation value than neurons activated from a lower threshold (i.e., = 0). Considering both the percentage of activated neurons and their corresponding activation value, it becomes apparent that the neurons in the upper layers contribute more to the output performance.

A.4 Supplemental Results of Neuron Attribution

A.4.1 Contribution Score of Different Tasks

The contribution score of the four types of neurons evaluated on the fact probing task, XNLI task, cross-lingual KE task are shown in Figures 25 - 28.

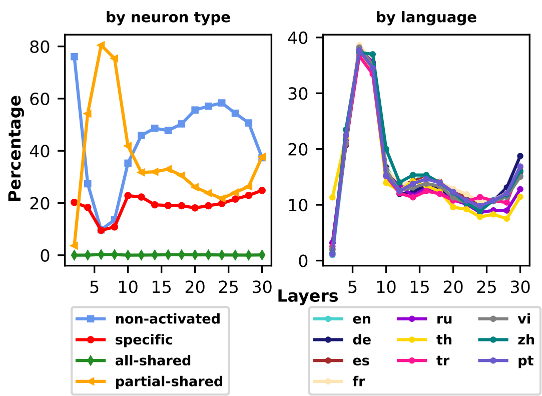


Figure 19: Neuron behavior pattern in the fact probing task with BLOOM backbone.

A.4.2 Effective Score of Different Tasks

The maximum, minimum, and average effective scores of four types of neurons in 10 languages evaluated on the cross-lingual KE (EN (edit) $\rightarrow$ ALL (test)) task are shown in Table 4. The maximum effective of four types of neurons across layers evaluated on the fact probing task, XNLI task, cross-lingual KE (ALL (edit) $\rightarrow$ EN (test)) task are shown in Figures 29 - 31.

A.5 Supplemental Results of Effects on Accuracy

A.5.1 Effects with LLAMA Backbone

In order to further prove the importance of *all-shared neurons* across LLMs, we conduct the ex-

	all-shared			partial-shared			specific			non-activated		
	max	min	mean	max	min	mean	max	min	mean	max	min	mean
en	1.85	-0.94	0.07	0.22	-0.16	0.00	0.02	-0.02	0.00	0.04	-0.03	0.00
de	1.03	-0.60	0.02	0.13	-0.13	0.00	0.07	-0.03	0.00	0.02	-0.01	0.00
es	1.15	-0.84	0.02	0.12	-0.11	0.00	0.01	-0.01	0.00	0.02	-0.02	0.00
fr	1.06	-0.78	0.01	0.15	-0.11	0.00	0.03	-0.04	0.00	0.02	-0.02	0.00
ru	0.70	-0.45	0.00	0.24	-0.13	0.00	0.08	-0.03	0.00	0.01	-0.01	0.00
th	0.50	-0.90	0.00	0.17	-0.10	0.00	0.03	-0.05	0.00	0.01	-0.01	0.00
tr	0.82	-0.51	0.03	0.12	-0.12	0.00	0.04	-0.03	0.00	0.02	-0.02	0.00
vi	0.86	-0.68	0.01	0.15	-0.11	0.00	0.04	-0.04	0.00	0.02	-0.02	0.00
zh	0.52	-0.42	0.00	0.17	-0.20	0.00	0.08	-0.07	0.00	0.02	-0.01	0.00
pt	1.14	-0.83	0.02	0.11	-0.15	0.00	0.02	-0.02	0.00	0.02	-0.02	0.00

Table 4: Maximum, minimum, average effective score of the four types of neurons on the cross-lingual KE (EN (edit) → ALL (test)) task.

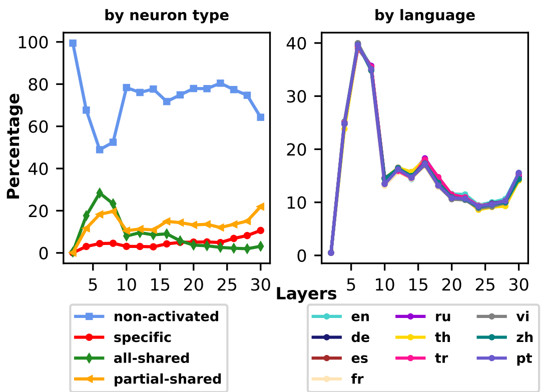


Figure 20: Neuron behavior pattern in the cross-lingual KE (ALL (edit) → EN (test)) task with BLOOM backbone.

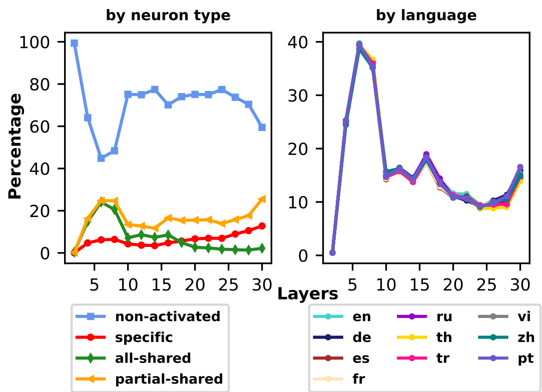


Figure 21: Neuron behavior pattern in the cross-lingual KE (EN (edit) → ALL (test)) task with BLOOM backbone.

accuracy	en	de	es	fr	ru	th	tr	vi	zh
baseline	59.1	47.6	50.1	47.0	49.1	41.4	40.2	51.6	46.1
w/o. all-shared	**3.0**	**3.6**	**4.4**	**1.9**	**4.7**	**6.9**	**3.6**	**13.5**	**4.8**
w/o. specific	59.2	47.3	49.9	47.0	49.1	41.9	40.1	51.4	46.2
w/o. partial-shared	59.1	48.4	51.5	47.9	49.7	42.9	41.5	50.8	48.0

Table 5: The effects of accuracy on the XNLI task with LLAMA2-7b-chat backbone, when deactivating *all-shared neurons*, *specific neurons*, and *partial-shared neurons*, respectively. "w/o." stands for "without".

periments with deactivating neurons on the XNLI task with LLAMA2-7b-chat backbone. The results in Table 5 show that there is more significant effect when *all-shared neurons* are deactivated. It demonstrates that *all-shared neurons* play a key role in predicting correct answers across LLMs.

A.5.2 Effects on the Cross-lingual KE Task

We further explore the influence of deactivating *all-shared neurons*, *specific neurons*, and *partial-shared neurons* on the cross-lingual KE (EN (edit) → ALL (test)) task with BLOOMZ backbone. The results in Table 6 are consistent with the results of XNLI task in Table 2, demonstrating the critical role of *all-shared neurons* for generating correct output.

accuracy	en	de	es	fr	pt	th	tr	ru	vi	zh
baseline	96.23	46.84	36.88	40.38	35.80	4.71	28.13	0.67	40.92	10.63
w/o. all-shared	**15.21**	**9.69**	**5.52**	**4.98**	**7.67**	**0.54**	**2.42**	**0.00**	**7.67**	**2.96**
w/o. specific	96.23	46.84	36.74	40.24	35.53	4.71	29.21	0.67	40.65	11.84
w/o. partial-shared	68.78	44.95	35.40	37.01	34.19	5.11	25.44	0.67	39.84	8.48

Table 6: The effects of accuracy on the cross-lingual KE (EN (edit) → ALL (test)) task with BLOOMZ backbone, when deactivating *all-shared neurons*, *specific neurons*, and *partial-shared neurons*, respectively. "w/o." stands for "without".

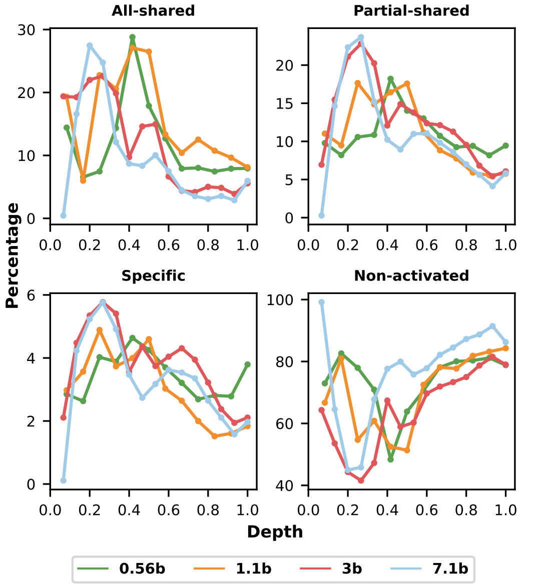


Figure 22: Neuron behavior pattern in a XNLI task with the model size as 0.56b, 1b, 3b, 7b.

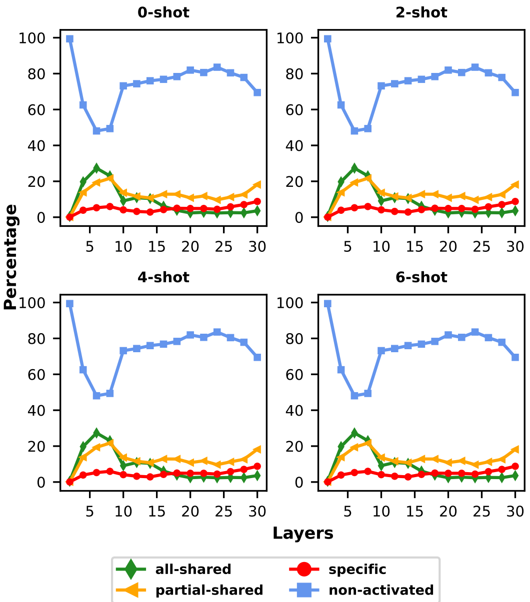


Figure 23: Neuron behavior pattern in cross-lingual KE (EN (edit) → ALL (test)) task with BLOOMZ backbone under the in-context learning.

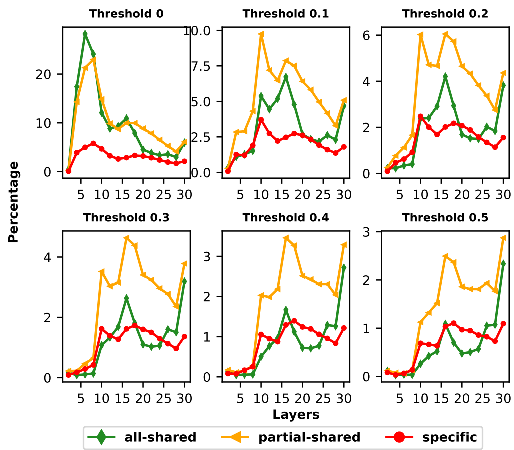


Figure 24: Neuron behaviors in a XNLI task with BLOOMZ backbone under the threshold in $[0, 0.1, 0.2, 0.3, 0.4, 0.5]$.

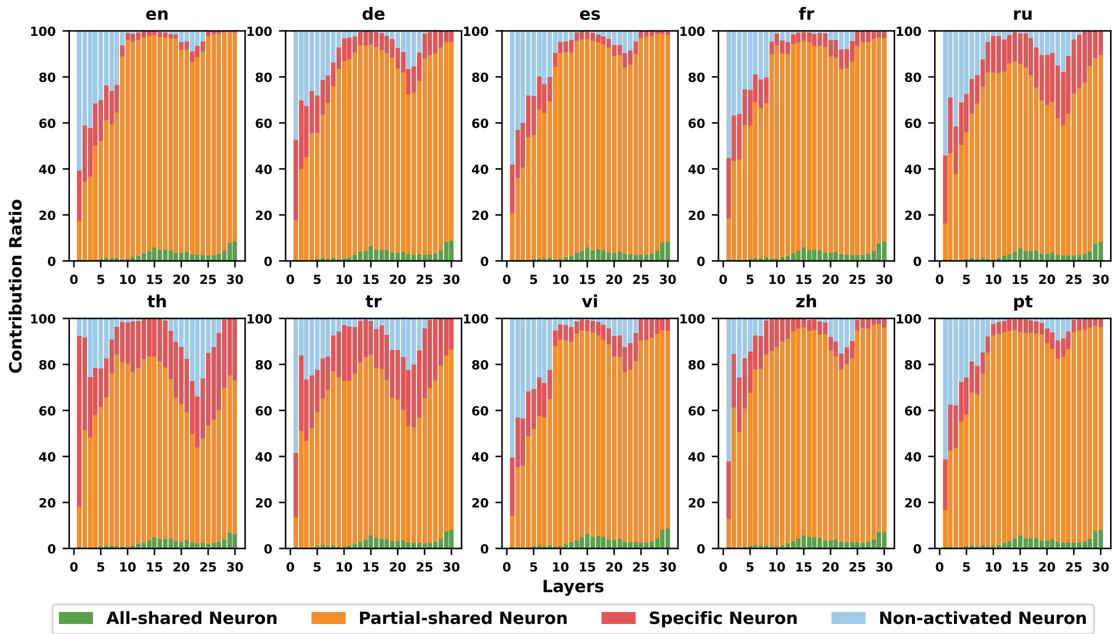


Figure 25: Contribution proportion of four types of neurons based on the fact probing task with BLOOMZ backbone.

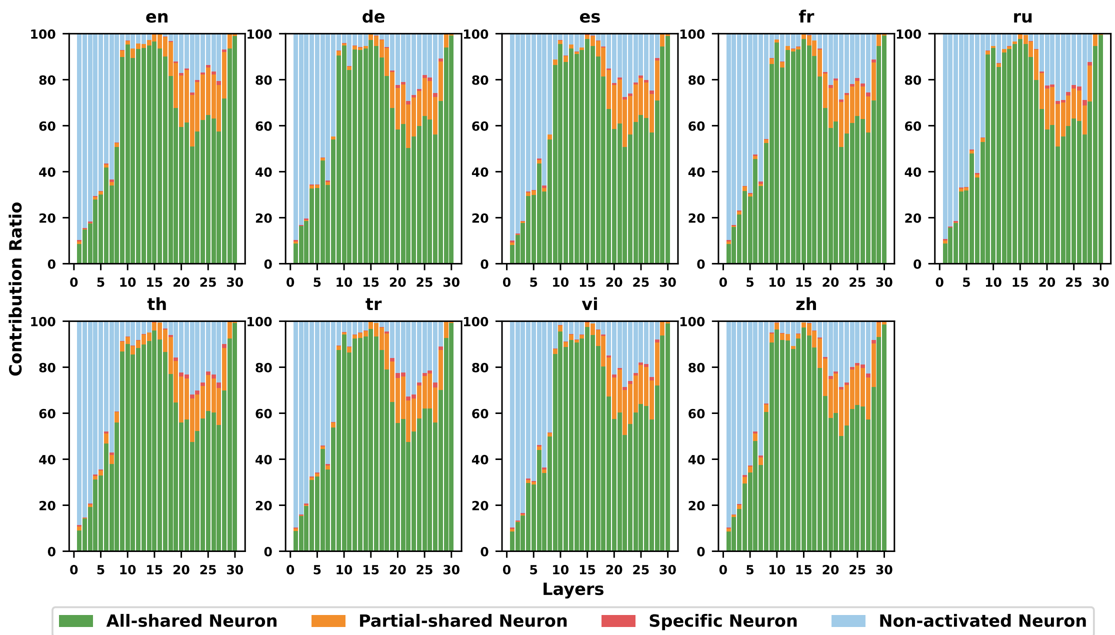


Figure 26: Contribution proportion of four types of neurons based on the XNLI task with BLOOMZ backbone.

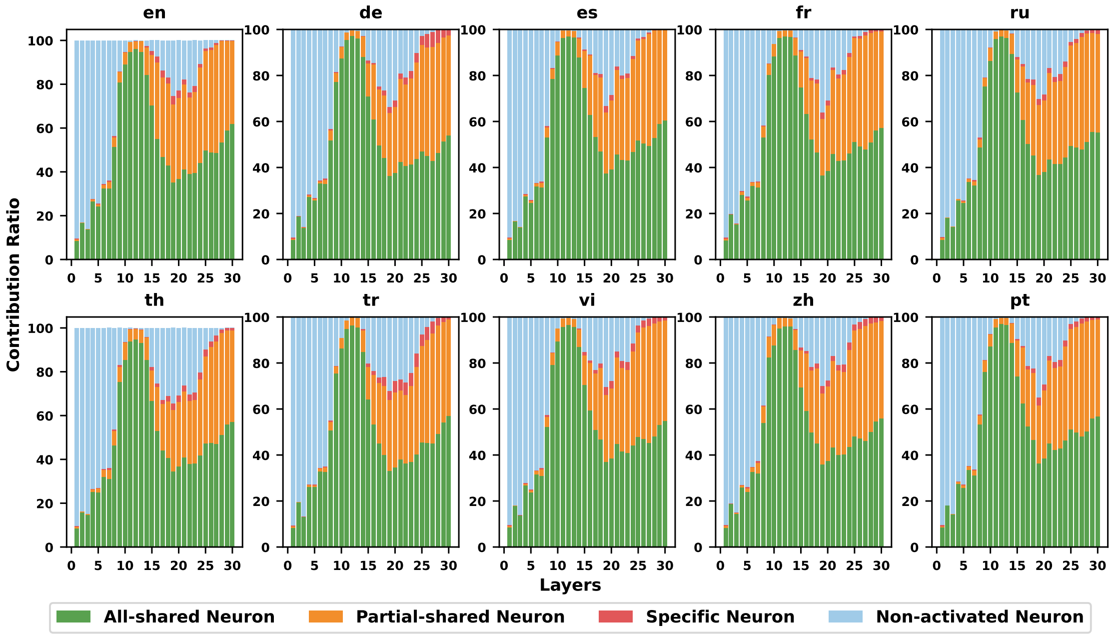


Figure 27: Contribution proportion of four types of neurons based on the cross-lingual KE (EN (edit) → ALL (test)) task with BLOOMZ backbone.

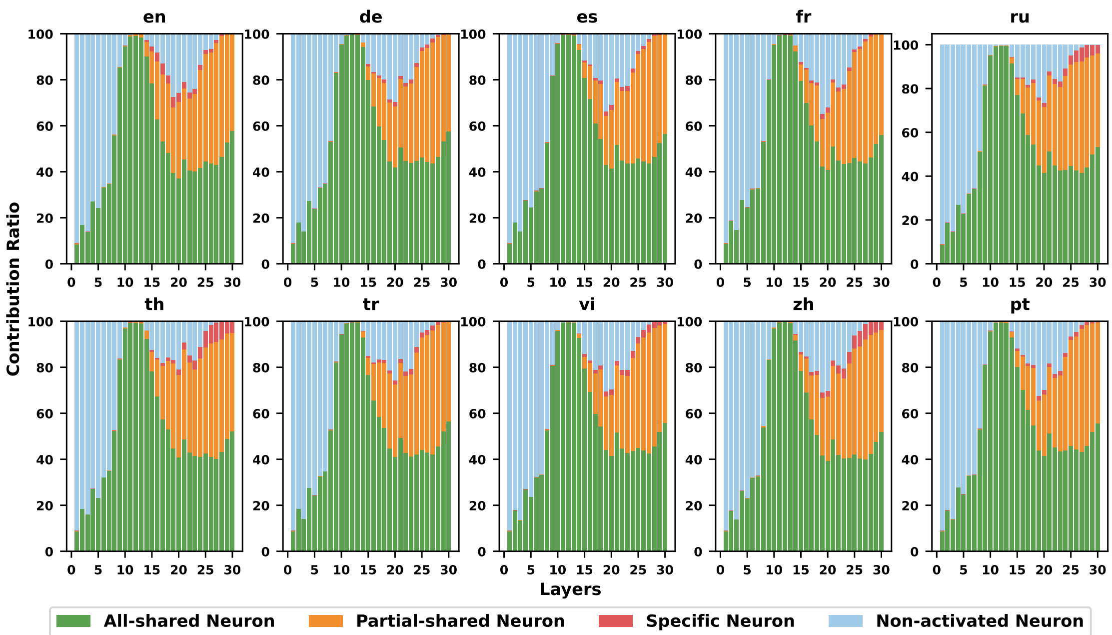


Figure 28: Contribution proportion of four types of neurons based on the cross-lingual KE (ALL (edit) → EN (test)) task with BLOOMZ backbone.

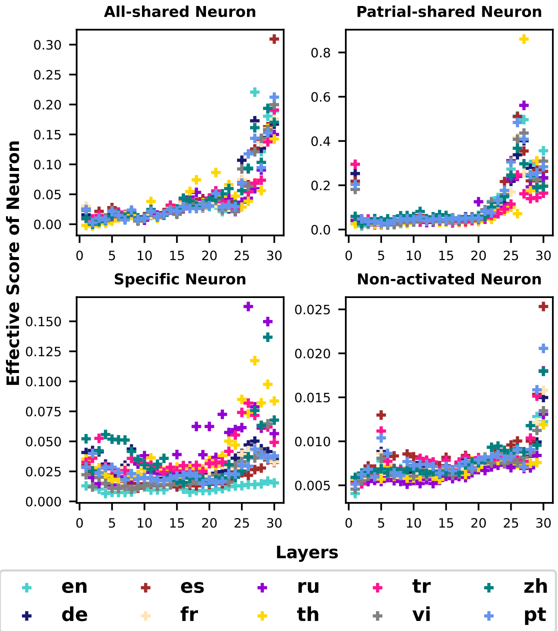


Figure 29: Effective score of four types of neurons based on the fact probing task with BLOOMZ backbone.

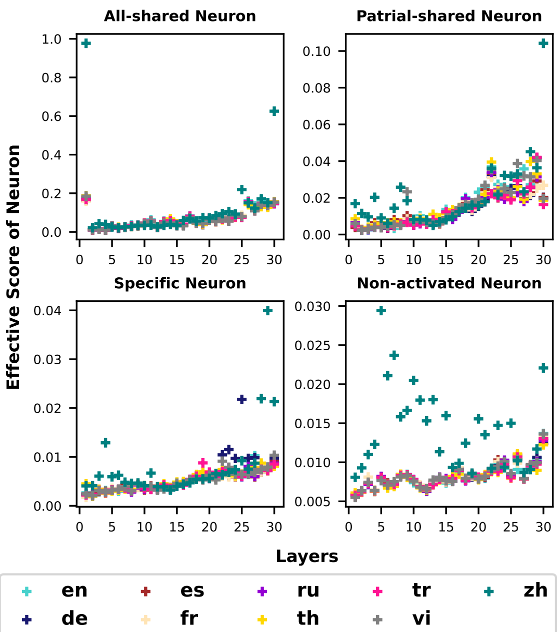


Figure 30: Effective score of four types of neurons based on the XNLI task with BLOOMZ backbone.

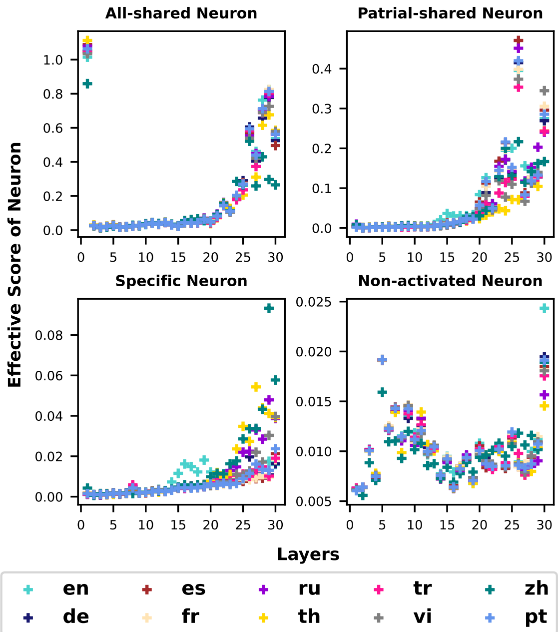


Figure 31: Effective score of four types of neurons based on the cross-lingual KE (ALL (edit) $\rightarrow$ EN (test)) task with BLOOMZ backbone.