where $x \in \mathbb{R}^{l \times d_{model}}$ is the embedding fed into the FFN, $W_{gate}, W_{up} \in \mathbb{R}^{d_{model} \times d_{inter}}$[2], $W_{down} \in \mathbb{R}^{d_{inter} \times d_{model}}$. The calculation of the importance of the k-th neuron in W_{up}, when processing the input c, as presented in Equation 1, can be equivalently transformed to

$$\text{Imp}(W_{up}[:, k]|c) = \|\hat{\text{FFN}}(x) - \text{FFN}(x)\|_2 = \left\|\big(h_{\text{ffn}} \cdot \texttt{Mask}[k]\big) W_{down}(x)\right\|_2, \tag{4}$$

where $h_{\text{ffn}} \in d_{inter}$ represents the embedding before W_{down}, and $\texttt{Mask}[k] \in d_{inter}$ is a vector with the k-th element equal to 1 and the rest equal to 0. To calculate Imp$(W_{up}[:, k]|c)$ for $k \in d_{inter}$ parallelly, we introduce a diagonal mask matrix of size (d_{inter}, d_{inter}), denoted as $\texttt{Mask}$. Therefore,

$$\text{Imp}(W_{up}|c) = \|(h_{\text{ffn}} \cdot \texttt{Mask}) W_{down}(x)\|_2. \tag{5}$$

Furthermore, we observe that deactivating the k-th neuron of W_{down} is equivalent to deactivating the k-th neuron in W_{up}, as they both result in $h_{\text{ffn}}[k] = 0$. Hence, we can also derive Imp$(W_{down}|c)$ by employing Equation (5).

Self-Attention Network When processing input c, the self-attention network in a certain layer is

$$\text{Attention}(x) = \text{Softmax}\big(\frac{W_Q(x) W_K^T(x)}{\sqrt{d}}\big) W_V(x), \tag{6}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_{mid}}$. [3] Since $W_V(x)$ is not in the non-linear softmax calculation, we can calculate Imp$(W_V|c)$ by applying Equation (5). For W_Q, we obtain Imp$(W_Q[:, k]|c)$ by deactivating its k-th neuron, specifically, $\hat{W}_Q \leftarrow W_Q[:, k] = 0$. Firstly, we calculate the difference in attention weight before and after deactivation, prior to scaling and softmax,

$$\Delta_k(x) = W_Q(x) W_K^T(x) - \hat{W}_Q(x) W_K^T(x) = W_Q(x)[:, k] W_K(x)[k, :] \in \mathbb{R}^{l \times l}. \tag{7}$$

Next, as the changes in attention exhibit a positive correlation with the changes in the output of this layer, the importance of $W_Q[:, k]$ in processing c, as defined in Equation 1, can be approximated as

$$\begin{aligned} \text{Imp}(W_Q[:, k]|c) &\approx \|\widehat{\text{attention}}(x) - \text{attention}(x)\|_2 \\ &\approx \left\|\text{softmax}\big(\frac{W_Q(x) W_K^T(x) - \Delta_k(x)}{\sqrt{d}}\big) - \text{softmax}\big(\frac{W_Q(x) W_K^T(x)}{\sqrt{d}}\big)\right\|_2. \end{aligned} \tag{8}$$

This process can also be calculated in parallel, specifically,

$$\begin{aligned} \Delta(x) &= W_Q(x) W_K^T(x) - \hat{W}_Q(x) W_K^T(x) \\ &= W_Q(x).resize(l, 1, d_{mid}) \times W_K(x).resize(1, l, d_{mid}) \in \mathbb{R}^{l \times l \times d_{mid}}. \end{aligned} \tag{9}$$

Therefore, the importance of W_Q in processing input c is calculated by

$$\text{Imp}(W_Q|c) \approx \left\|\text{softmax}\big(\frac{W_Q(x) W_K^T(x) - \Delta(x)}{\sqrt{d}}\big) - \text{softmax}\big(\frac{W_Q(x) W_K^T(x)}{\sqrt{d}}\big)\right\|_2. \tag{10}$$

Similarly, since W_K is symmetrical to W_Q, Imp$(W_K|c)$ can be calculated in the same way.

2.3 Detection of Language-Specific Neurons

We then apply PLND to selected languages and models to validate its effectiveness in detecting language-specific neurons and to further investigate the relationships between languages.

Experimental Setup. We test two open-source models that perform well on multilingual tasks, including *Vicuna-7b-v1.5*[4] (Chiang et al., 2023) and *Mistral-7b-Instruct-v0.2* (Jiang et al., 2023). For simplicity, we abbreviate them as Vicuna and Mistral hereafter to represent the two models respectively. We select the text summarization task with the XLSum (Hasan et al., 2021) dataset as the reference task to evaluate multilingual performance as it requires the model to comprehend the

[2] $W(\cdot)$ represents the linear matrix product of the input x and the parameter W, i.e., $W(x) := xW$.

[3] In some models like Vicuna and Mistral, $d_{model} = d_{mid}$, but we use different notations to avoid ambiguity.

[4] We do not directly utilize Llama2-chat as it does not follow multilingual instructions, consistently responding in English regardless of the language of the query.

Table 1: Multilingual performance on XLSum when deactivating language-specific neurons ("Lang-Spec") and an equivalent number of randomly selected neurons ("Random").

Model	Method	Fr	Zh	Es	Ru	Avg.
Vicuna	Original	14.2	61.1	10.4	20.8	26.6
	Deactivate Random	14.1	61.6	10.4	20.8	26.7
	Deactivate Lang-Spec	**0.83**	**0.00**	**0.24**	**0.42**	**0.37**
Mistral	Original	15.2	56.4	10.6	21.0	25.8
	Deactivate Random	15.4	55.9	10.2	21.2	25.7
	Deactivate Lang-Spec	**0.21**	**0.39**	**0.15**	**0.07**	**0.21**

input text and generate a coherent fragment. We adopt 4 high-resource languages including French (Fr), Chinese (Zh), Spanish (Es), and Russian (Ru), as their initial performance on those languages is already quite reasonable for observing the multilingual processing mechanism. Furthermore, we utilize OSCAR (Caswell et al., 2020) corpus which contains web crawling texts for each language to compile a language-specific corpus without task-specific considerations. More details are presented in Appendix B.

Existence of Language-Specific Neurons Using PLND, we feed a corpus in a specific language to LLMs and identify neurons that are consistently activated, which are responsible for processing queries in that language. To ascertain whether these neurons are genuinely language-specific, we assess the performance of LLMs in corresponding languages when these neurons are deactivated versus when the same number of randomly sampled neurons are deactivated.

Table 1 demonstrates the decline of multilingual capabilities when deactivating language-specific neurons. Although just deactivating around 0.13% neurons, LLMs lose their multilingual capabilities and fail to generate meaningful content. In contrast, deactivating the same number of randomly selected neurons does not yield any difference. Therefore, the detected neurons are language-specific and related to handling corresponding multilingual inputs.

2.4 Analysis of Language-Specific Neurons

We further investigate the degree of overlap among their language-specific neurons. Our findings reveal that in both Mistral and Vicuna, English shows limited overlap with other languages, indicating many language-specific neurons, while languages within the same family, such as Spanish, French, and English, demonstrate more overlap. More details are illustrated in Appendix C.

In addition, we examine two more types of multilingual LLMs, including BLOOMZ (Muennighoff et al., 2023), a *hyper-multilingual* LLM claiming to support 46 languages, and Chinese Llama (Cui et al., 2023), a *bilingual* LLM focusing on English and Chinese. We find that language-specific neurons in BLOOMZ follow patterns similar to Mistral and Vicuna. However, in Chinese LLama, Chinese dominates as the primary language for reasoning and knowledge extraction across all languages, with notably absent language-specific neurons. Details are shown in Appendix D.

Given the certain overlap ratio of language-specific neurons from other languages with those of English, as illustrated in the first column of Figure 5 and Figure 6, we conduct supplementary experiments to demonstrate that these neurons are not language-agnostic neurons crucial for general comprehension and logical reasoning (Liang et al., 2024; Tang et al., 2024). Instead, these overlapping neurons represent only a subset of language-specific neurons, while the language-agnostic neurons responsible for essential understanding and reasoning are those not identified as language-specific. Further elaboration and detailed results are presented in Appendix E.

3 Multilingual Workflow (MWork) of LLMs

3.1 MWork

By classifying the hidden representations of each layer in LLMs into English or non-English (as shown in Figure 1), we can observe the shift from non-English to English-centric, and back to non-English with the progression through the layers. This motivates us to hypothesize a three-stage multilingual workflow: *understanding* the original non-English queries and interpreting them in

English, *task-solving* in English, and *generating* back to the original language. Nevertheless, the presence of certain non-English tokens during the English-centric task-solving stage inspires us to further investigate this stage.

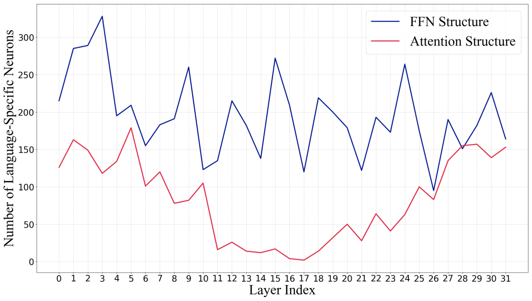

Figure 3: Number of language-specific neurons when processing multilingual queries.

With the proposed PLND method, we extract language-specific neurons from attention and feed-forward structures when processing various multilingual queries. We plot the average number of activated language-specific neurons of Mistral when processing each query in Figure 3. Notably, the number of language-specific neurons decreases within the self-attention structure in the task-solving layer but remains consistent across the layers of the feed-forward structure. This decline implies a reliance on the English language for reasoning while extracting multilingual knowledge to support query processing, which is also consistent with (Geva et al., 2021)'s interpretation of the feed-forward structure as key-value memories for knowledge extraction. Therefore, we further decompose the task-solving layer into two parts: *reasoning* in English and *extracting knowledge* in a multilingual context.

Considering the above insights, we propose the MWork hypothesis for explaining LLM's multilingual workflow: LLMs first *understand* user input by unifying diverse linguistic features. They then engage in the *task-solving* phase, employing English for reasoning and leveraging multilingual knowledge through self-attention and feed-forward structures, respectively. Finally, the models *generate* responses aligned with the query's original language.

3.2 Verification Experiment Setup

To verify MWork, we selectively deactivate language-specific neurons from each component. Then its functionality can be verified if this deactivation results in minimal impact on English performance while exhibiting a notable decline in multilingual performance for the corresponding task.

Dataset To comprehensively understand how LLMs work with different abilities, we employ four kinds of tasks including MGSM (Shi et al., 2022) for reasoning task, XQuAD (Artetxe et al., 2020) for understanding task, X-CSQA (Lin et al., 2021) for knowledge question answering task, and XLSum (Hasan et al., 2021) for generation task. Detailed information regarding these datasets and the testing prompts can be found in Appendix F. We adopt 6 languages including English (En), German (De), French (Fr), Chinese (Zh), Spanish (Es), and Russian (Ru), as their initial performance on those languages is already quite reasonable for observing the multilingual processing mechanism. For XLSum, we randomly sample 500 data points from the whole test set for each language taking into consideration its long inference time, while for other tasks, we employ the entire test set. We evaluate the vanilla performance of Vicuna and Mistral on these datasets for later comparison as presented in Appendix G. For reasoning, understanding, and knowledge question answering tasks, we adopt accuracy as the metric. As for the generation tasks, we adopt ROUGE-L as the metric.

Deactivation Strategy We primarily consider two aspects when selecting the deactivation settings: (1) language-specific neurons versus randomly chosen neurons, and (2) the position of neurons, which encompasses four structures. Note that for a fair comparison, we ensure the numbers of deactivated neurons in all settings are the same. More detailed settings are explained from Section 3.3 to Section 3.6. For the concrete numbers of different layers, we tune hyperparameters by XQuAD in Chinese. Details are explained in Appendix H.

Notations Tables 2 to 5 present the results of deactivating certain neurons, where "Under" denotes the understanding layers, "S-ATTN" and "S-FFN" correspond to the self-attention and the feed-forward structures within the task-solving layers respectively, "Gen" refers to the generation layers. The term "Random" is used to describe deactivating randomly chosen neurons, whereas "Lang-Spec" refers to the deactivation of language-specific neurons. We also present the gap between the original performance (as shown in Table 11) and performance after deactivation (as shown in Table 14 to Table