Table 2: Results of the **understanding** task, where '✗' indicates that chosen neurons in the corresponding layer are deactivated, and '✓' signifies they are activated. Δ is defined as the difference between the reduction in performance in English, denoted as Δ_{Eng}, and the reduction in performance in non-English languages, denoted as $\Delta_{\text{n-Eng}}$.

Model	**Deactivating Method**					**Performance**				
	Under	S-ATTN	S-FFN	Gen	Neuron	Eng	n-Eng	Δ_{Eng}	$\Delta_{\text{n-Eng}}$	$\Delta \uparrow$
Vicuna	✗	✓	✓	✓	Random	57.8	53.9	+0.3	−0.1	+0.4
	✗	✗	✗	✗	Random	57.9	54.2	+0.4	+0.3	+0.1
	✓	✗	✗	✓	Lang-Spec	40.9	38.6	−15.9	−15.3	−0.6
	✓	✓	✓	✗	Lang-Spec	57.9	52.8	−0.4	−1.1	+0.7
	✗	✓	✓	✓	Lang-Spec	56.5	46.0	−0.5	−7.9	+7.4
Mistral	✗	✓	✓	✓	Random	58.1	55.5	+1.0	−0.2	+1.2
	✗	✗	✗	✗	Random	57.6	55.5	+0.5	−0.2	+0.7
	✓	✗	✗	✓	Lang-Spec	53.2	47.0	−3.9	−8.7	+4.8
	✓	✓	✓	✗	Lang-Spec	56.4	54.6	−0.7	−1.0	+0.3
	✗	✓	✓	✓	Lang-Spec	56.2	48.3	−0.9	−7.4	+6.5

17) for English (Δ_{Eng}) and averaged non-English languages ($\Delta_{\text{n-Eng}}$), respectively. A single metric Δ is then introduced as $\Delta_{\text{Eng}} - \Delta_{\text{n-Eng}}$, where a high value indicates such deactivation operation does not bring much impact to the English performance but lead to performance drop in non-English. Therefore, this provides a direct single indicator that the deactivated neurons are language-specific and hold a significant responsibility in executing the corresponding task.

3.3 Verify the Understanding Stage in `MWork`

Deactivating Method Table 2 shows the results of the understanding task following the deactivation of five distinct sets of neurons: (i) neurons randomly selected from the understanding layers; (ii) neurons randomly chosen across all layers; (iii) language-specific neurons within the task-solving layers; (iv) language-specific neurons in the generation layers; (v) language-specific neurons in the understanding layers. As mentioned above, in order to verify the functionality of the understanding layer (setting v), we compare it with deactivating other types of layers, specifically setting iii for the task-solving layer and setting iv for the generation layer. Full results are listed in Appendix I.

Findings We find that by deactivating randomly sampled neurons, no matter in the understanding layer or all layers, the performance of LLMs in both English and non-English languages is almost unaffected compared to other deactivating methods. Note that in some cases, deactivating randomly sampled neurons may even increase the performance because irrelevant neurons are removed, which also aligns with the finding from (Sharma et al., 2023). When assessing the differential impact on English and non-English language performance after the deactivation, specifically the difference calculated as $\Delta_{\text{Eng}} - \Delta_{\text{n-Eng}}$, it is evident that the deactivation of random neurons within the understanding layer amplifies this effect. This observation lends partial support to the hypothesized role of the understanding layer in language processing.

Furthermore, we find that deactivating language-specific neurons in the understanding layer influences the performance in English a little while significantly decreasing the performance in non-English languages. When deactivating language-specific neurons in the task-solving layer, both English and non-English languages are significantly reduced while deactivating language-specific neurons in the generation layer influences a little for both English and non-English languages. Therefore, we prove that the first several layers are responsible for understanding because deactivated neurons just disable LLMs on the NLU task in non-English languages. Furthermore, disabling language-specific neurons in the task-solving layer shows that LLMs rely on English, as performance drops across all languages.

3.4 Verify the Reasoning Structure in `MWork`

Deactivating Method Table 3 shows the result of the reasoning task, where we deactivate 6 sets of neurons. We adhere to the previous logic of selecting deactivation settings, with the exception that

Table 3: Results of the **reasoning** task. Disabling all language-specific neurons, except for those involved in self-attention structure within the task-solving layer, greatly reduces performance.

Model	**Deactivating Method**					**Performance**				
	Under	S-ATTN	S-FFN	Gen	Neuron	Eng	n-Eng	Δ_{Eng}	$\Delta_{\text{n-Eng}}$	$\Delta \uparrow$
Vicuna	✓	✗	✓	✓	Random	20.0	11.3	−0.4	−1.8	+1.4
	✓	✗	✗	✓	Random	18.4	12.2	−2.0	−1.0	−1.0
	✗	✗	✗	✗	Random	19.6	12.5	−0.8	−0.7	−0.1
	✓	✗	✗	✓	Lang-Spec	7.2	3.4	−13.2	−9.8	−3.4
	✗	✓	✓	✗	Lang-Spec	18.1	8.3	−2.3	−4.9	+2.6
	✗	✓	✗	✗	Lang-Spec	19.0	7.8	−1.4	−5.4	+4.0
Mistral	✓	✗	✓	✓	Random	40.8	23.4	−5.2	−2.9	−2.3
	✓	✗	✗	✓	Random	39.2	24.0	−6.8	−2.3	−4.5
	✗	✗	✗	✗	Random	45.2	26.8	−0.8	+0.5	−1.3
	✓	✗	✗	✓	Lang-Spec	38.2	18.4	−7.8	−7.9	+0.1
	✗	✓	✓	✗	Lang-Spec	44.0	18.1	−2.0	−8.2	+6.2
	✗	✓	✗	✗	Lang-Spec	46.2	18.3	+0.2	−8.0	+8.2

Table 4: Results of the **knowledge** question answering task. The highest performance reduction difference (Δ) is achieved by disabling all language-specific neurons in the feed-forward structure within the task-solving layer.

Model	**Deactivating Method**					**Performance**				
	Under	S-ATTN	S-FFN	Gen	Neuron	Eng	n-Eng	Δ_{Eng}	$\Delta_{\text{n-Eng}}$	$\Delta \uparrow$
Vicuna	✓	✓	✗	✓	Random	57.5	39.5	−0.3	+0.0	−0.3
	✓	✗	✗	✓	Random	56.0	38.7	−1.8	−0.8	−1.0
	✗	✗	✗	✗	Random	57.7	39.6	−0.1	+0.1	−0.2
	✓	✗	✓	✓	Lang-Spec	33.7	30.3	−24.1	−9.2	−14.9
	✓	✓	✗	✓	Lang-Spec	57.5	37.5	−0.3	−2.0	+1.7
Mistral	✓	✓	✗	✓	Random	61.0	37.0	−0.3	−0.5	+0.2
	✓	✗	✗	✓	Random	60.7	36.3	−0.6	−1.2	+0.6
	✗	✗	✗	✗	Random	61.8	37.4	+0.1	−0.1	+0.2
	✓	✗	✓	✓	Lang-Spec	51.2	28.9	−10.1	−8.6	−1.5
	✓	✓	✗	✓	Lang-Spec	61.2	35.1	−0.1	−2.4	+2.3

we do not conduct an independent experiment on deactivating neurons in the understanding layer, as its functionality has already been verified. Details are listed in Appendix I.

Findings We find that deactivating randomly sampled neurons in task-solving layers disables the capabilities of LLMs in reasoning to a greater extent than deactivating randomly sampled neurons in all layers, which verifies the function of the task-solving layer. Furthermore, comparing three deactivating language-specific neuron methods, we find that deactivating the task-solving layer decreases performance in both English and non-English. On the contrary, when we only deactivate language-specific neurons not in the task-solving layer, non-English is influenced more seriously than English. Moreover, eliminating interference from the feed-forward layer achieves better results, which verifies the function of attention structure in the task-solving layer.

3.5 Verify the Knowledge Extraction Structure in `MWork`

Deactivating Method Table 4 shows the result of the knowledge question answering task, where we deactivate 5 sets of neurons. Similarly, we exclude the deactivation of neurons in layers that have already been verified and instead concentrate on the self-attention structure and feed-forward structure in the task-solving layer. Details are listed in Appendix I.

Findings Likewise, targeted deactivation of language-specific neurons within the feed-forward structure of the task-solving layer predominantly affects non-English languages. This implies that

Table 5: Results of the **generation** task. The highest performance reduction difference (Δ) is achieved by disabling all language-specific neurons in the generation layer.

Model	**Deactivating Method**					**Performance**				
	Under	S-ATTN	S-FFN	Gen	Neuron	Eng	n-Eng	Δ_{Eng}	$\Delta_{\text{n-Eng}}$	$\Delta\uparrow$
Vicuna	✓	✓	✓	✗	Random	13.2	26.8	+0.1	+0.1	+0.0
	✗	✗	✗	✗	Random	13.0	26.7	−0.1	+0.0	−0.1
	✓	✓	✓	✗	Lang-Spec	13.1	25.7	+0.0	−1.1	+1.1
Mistral	✓	✓	✓	✗	Random	13.6	25.9	+0.1	+0.1	+0.0
	✗	✗	✗	✗	Random	13.6	25.7	+0.1	−0.2	+0.3
	✓	✓	✓	✗	Lang-Spec	13.8	24.3	+0.3	−1.5	+1.8

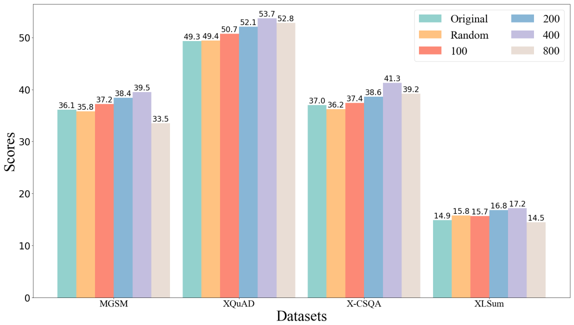

Figure 4: Enhancement results on high-resource languages, while the number is average among languages.

Table 6: Enhancement is achieved by fine-tuning Mistral-7b-v0.1 model utilizing 400 documents from each language correspondingly. The results are averaged across four tasks. Performance on English ("En") is obtained by averaging the results from four fine-tuned models.

Method	**En**	**Vi**	**Th**	**Ar**	**Sw**
Original	41.1	32.7	25.6	21.7	15.1
Random	40.8	32.7	25.2	21.2	15.1
Lang-Spec	**44.6**	**34.9**	**28.5**	**23.4**	**16.9**

processing multilingual queries necessitates accessing the multilingual information embedded within the relevant structures. However, disabling the self-attention structure compromises the ability to solve tasks across all languages.

3.6 Verify the Generation Structure in `MWork`

Deactivating Method Table 5 shows the result of the generation task, where we deactivate 3 sets of neurons. Since all previous layers have been verified, we solely deactivate neurons in the generation layer and compare them with randomly selected neurons. Details are listed in Appendix I.

Findings Similar to other tasks, the disabling of language-specific neurons within the generation layer diminishes their capacity to generate content in the respective languages. By selectively deactivating neurons that are not associated with English, we do not completely eliminate the models' multilingual generation abilities. However, as demonstrated in Table 1, the complete deactivation of all language-specific neurons results in the total loss of the LLMs' multilingual generation capabilities.

4 Multilingual Enhancement with `MWork`

We have verified `MWork` for explaining the multilingual working mechanism of LLMs in the above section via deactivating certain neurons. While opposite to employing deactivation, we can also enhance their multilingual ability, especially the understanding and generating ability, by fine-tuning these language-specific neurons. With language-specific neurons comprising only around 0.1% of all parameters, the need for training documents to improve multilingual capabilities can be significantly reduced to just a few hundred. Additionally, fine-tuning only the language-specific neurons for a particular language does not impact performance in other languages, allowing us to enhance specific languages while preserving performance in others.

`MWork` helps with enhancing multilingual ability by hundreds of documents. We employ *Mistral-7b-v0.1* for enhancement to eliminate the interference of instruction fine-tuning, and select causal language modeling as our training task. We create a dataset comprising $\{100, 200, 400, 800\}$