How Does Alignment Enhance LLMs' Multilingual Capabilities? A Language Neurons Perspective

Shimao Zhang[1*†], Zhejian Lai[1*], Xiang Liu[1*], Shuaijie She[1], Xiao Liu[2], Yeyun Gong[2], Shujian Huang[1‡], Jiajun Chen[1]

[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2] Microsoft Research Asia
{smzhang, laizj, liuxiang, shesj}@smail.nju.edu.cn, {huangsj, chenjj}@nju.edu.cn,
{xiao.liu.msrasia, yegong}@microsoft.com

Abstract

Multilingual Alignment is an effective and representative paradigm to enhance LLMs' multilingual capabilities, which transfers the capabilities from the high-resource languages to the low-resource languages. Meanwhile, some research on language-specific neurons provides a new perspective to analyze and understand LLMs' mechanisms. However, we find that there are many neurons that are shared by multiple but not all languages and cannot be correctly classified. In this work, we propose a ternary classification methodology that categorizes neurons into three types, including *language-specific* neurons, *language-related* neurons, and *general* neurons. And we propose a corresponding identification algorithm to distinguish these different types of neurons. Furthermore, based on the distributional characteristics of different types of neurons, we divide the LLMs' internal process for multilingual inference into four parts: (1) multilingual understanding, (2) shared semantic space reasoning, (3) multilingual output space transformation, and (4) vocabulary space outputting. Additionally, we systematically analyze the models before and after alignment with a focus on different types of neurons. We also analyze the phenomenon of "Spontaneous Multilingual Alignment". Overall, our work conducts a comprehensive investigation based on different types of neurons, providing empirical results and valuable insights to better understand multilingual alignment and multilingual capabilities of LLMs.

Code —
https://github.com/NJUNLP/Language-Neurons-Alignment
Extended version — https://arxiv.org/abs/2505.21505

Introduction

By training on the extensive corpus, large language models (LLMs) demonstrate outstanding language capabilities (Yang et al. 2024; Liu et al. 2024; Grattafiori et al. 2024; Zhang et al. 2025). However, due to the unbalanced pretraining corpus across different languages, LLMs have very uneven performance on high-resource languages and low-resource languages (Huang et al. 2023; Zhu et al. 2023; Zhang et al. 2024; Fan et al. 2025). Therefore, researchers have conducted comprehensive explorations to further enhance the multilingual performance of LLMs. A straightforward approach is increasing the proportion of non-English texts during pretraining (Ni et al. 2021; Yang et al. 2024) or performing continual pretraining with multilingual texts (Liu, Winata, and Fung 2021; Ji et al. 2024). But these approaches often entail high computational costs and substantial amounts of multilingual data.

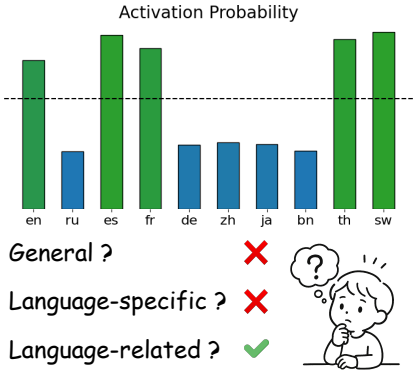


Figure 1: A neuron's activation probability across different languages. This neuron that exhibits high activation probabilities across *multiple* (**green**) but *not all* (**blue**) languages can't be correctly categorized under the existing classification methodology. The dashed line denotes the threshold that determines whether a neuron exhibits high activation in a given language.

Considering LLMs' great performance on high-resource languages, multilingual alignment has emerged as a representative paradigm for enhancing multilingual reasoning by transferring knowledge from high-resource to low-resource languages (Zhao et al. 2024a; She et al. 2024). A representative example is MAPO (Multilingual-Alignment-as-Preference Optimization) (She et al. 2024), which improves multilingual alignment by utilizing a well-trained multilingual translation model to compute alignment scores based

*Equal contribution.
†Work done during his internship at MSRA.
‡Corresponding author.

on the conditional generation probability of translating non-English responses into English.

Many studies conduct systematic mechanism analyses of the multilingual alignment and LLMs' multilingual capabilities. Zhao et al. (2024b) split the multilingual processing workflow into three parts: multilingual understanding, resolving tasks, and generating outputs in the target language. This three-stage inference workflow clearly demonstrates how LLMs leverage English as a pivot language to handle multilingualism using a unified pattern.

Inspired by the neurobiological underpinnings of human language faculties, Tang et al. (2024) categorize neurons in LLMs into two primary types: language-specific neurons and general neurons. Notably, these language-specific neurons are primarily situated in the model's top and bottom layers (Tang et al. 2024), which is consistent with the three-stage multilingual workflow of Zhao et al. (2024b).

However, we identify a key limitation in the existing neuron classification methodology: the identification of language-specific neurons focuses on the processing a particular language and thus neglects the inter-language alignment; while general neurons encode universal knowledge which is independent of a specific language. This leads to a critical question: what is the mechanism of sharing neurons between some languages? Furthermore, it is not well understood how multilingual alignment enhances the LLMs' multilingual reasoning capabilities from the perspective of language neurons.

In this work, we comprehensively investigate the multilingual alignment of LLMs with MAPO (She et al. 2024) as a representative multilingual alignment algorithm. As shown in Figure 1, we observed the existence of neurons that exhibit high activation probability on multiple but not all languages. These neurons are neither language-specific nor general for all languages. Hence, we refer to them as **language-related neurons**.

Moreover, we redefine **language-specific neurons** to restrict each neuron to be activated for only one language. The **general neurons** are defined as the neurons that are effective for all languages. To facilitate subsequent analysis, we propose a corresponding identification algorithm for distinguishing these different types of neurons.

Then we analyze the models before and after alignment, focusing on the changes in different types of neurons. Based on the distributional characteristics of these neurons, we divide LLMs' internal process for multilingual inference into four parts, with different parts exhibit distinct dependencies on different types of neurons. We demonstrate that multilingual alignment significantly enhances the activation of the corresponding types of neurons across the relevant layers. Additionally, we analyze the "spontaneous multilingual alignment" (Zhang et al. 2024) phenomenon in LLMs, providing insights into the roles of general neurons and language-related neurons shared across languages. For further analysis, we also provide observations about the uniqueness of English and the neuron distributions. Overall, based on different types of neurons, we present empirical results and valuable insights that contribute to a deeper understanding of multilingual alignment and the multilingual reasoning capabilities of LLMs.

Related Work

Multilingual Alignment

Conducting pretraining or continual pretraining on the multilingual corpus is a straightforward and effective method to enhance LLMs' multilingual capabilities (Ni et al. 2021; Ji et al. 2024). However, these methods typically require substantial investments in time, data, and computational resources. Thus, many researchers perform multilingual alignment to improve LLMs' multilingual performance by transferring the capabilities from high-resource languages to low-resource languages (Eronen, Ptaszynski, and Masui 2023; Zhao et al. 2024c,a; She et al. 2024), which efficiently and effectively improves the model performance in low-resource language scenarios. Furthermore, Zhang et al. (2024) first finds the "Spontaneous Multilingual Alignment" phenomenon in LLMs, which demonstrates that conducting multilingual alignment based on a small number of languages effectively improves the alignment even between English and many languages unseen during alignment.

Mechanistic Interpretability

In addition to enhancing LLMs' multilingual performance, research on the underlying mechanisms of multilingual capabilities in LLMs is still ongoing. It is crucial for us to understand and explain the LLMs and related methods explicitly. Typically, the existing approaches primarily perform mechanistic interpretability analyses by observing the internal states of the model (Nostalgebraist 2020; Zhang et al. 2024; Zhao et al. 2024b; Mousi et al. 2024). Overall, neuron states and latent intermediate logits are both important objects of observation. For latent logits, Wendler et al. (2024) utilizes logit lens (Nostalgebraist 2020) to directly project the logits in the intermediate layers to the vocabulary space, which reveals the latent participation of English in the intermediate layers. For neuron states, Hu et al. (2024) analyzes the neuron activation overlap to measure the extent of shared neuron activation across different languages.

Language-Specific Neurons

Many studies have revealed the language-related and universal components in LLMs. At the layer level, the multilingual processing of LLMs is considered to involve three stages (Zhao et al. 2024b; Wendler et al. 2024): converting multilingual inputs into a shared semantic space, intermediate-layer reasoning, and outputting in the target language. The top and bottom layers of the model handle multilingual processing, while the intermediate layers perform inference in similar patterns across different languages. This demonstrates a distinct division of labor within the model at the layer level regarding language specificity.

Furthermore, many studies investigate the finer-grained methods for language-specific neuron identification (Kojima et al. 2024; Tang et al. 2024; Tan, Wu, and Monz 2024). Tang et al. (2024) categorizes activated neurons into language-specific neurons and general neurons. They detect language-specific neurons by calculating language activation probabil-

ity entropy on massive text. However, we find that some neurons are activated by multiple languages (i.e., not language-specific), yet are not universally activated across all languages (i.e., not general). Simply categorizing activated neurons into two classes blurs this distinction. Thus, we propose a new method to identify neurons, which categorizes activated neurons into three types: language-specific neurons, language-related neurons, and general neurons.

Methodology

Preliminary Study

Most LLMs are pretrained mainly on the high-resource language corpus, which leads to LLMs' unstable and unbalanced performance in multilingual scenarios. As a representative multilingual alignment algorithm, Multilingual-Alignment-as-Preference Optimization (MAPO) (She et al. 2024) effectively and efficiently improves the LLMs' multilingual performance. Additionally, it is also important for us to understand and analyze the mechanism of LLMs' multilingual capabilities and multilingual alignment. Moreover, some studies on the identification of the language-specific and general neurons in LLMs (Tang et al. 2024; Kojima et al. 2024). It is found that LLMs' capabilities of processing a particular language mainly come from a small subset of neurons (Tang et al. 2024).

However, there are still many important questions waiting for further investigation. On the one hand, many methods overlook neurons activated by multiple languages yet not general, namely language-related neurons lie between language-specific and general categories. On the other hand, research from the perspective of language neurons on the underlying mechanisms of LLMs' multilingual alignment and multilingual reasoning capabilities remains quite limited, which is essential for better understanding and improving the multilingual performance of LLMs.

Multilingual Alignment

MAPO is a typical multilingual alignment algorithm to align the reasoning capabilities of non-English language responses with those of English, which serves as the pivot language. Specifically, for a given query X in a target (non-English) language and its corresponding English variant X_{En}, we collect their respective responses Y and Y_{En}. An off-the-shelf translation model, parameterized by θ, is deployed to estimate the conditional generation probability $P(Y \mid Y_{En}; \theta)$ by force-decoding Y conditioned on Y_{En}. A higher conditional probability is interpreted as stronger alignment between the target language response and its English counterpart. This probability is then used as an alignment score, denoted $r_\theta(X, Y)$.

This alignment score can be integrated into preference optimization algorithms. In PPO (Schulman et al. 2017), $r_\theta(X, Y)$ can be directly employed as the reward score. In DPO (Rafailov et al. 2023), for each target language, n distinct outputs are generated. Based on the alignment score, these n outputs are used to form $\binom{n}{2}$ preference pairs (Y_w, Y_l), where Y_w is deemed superior to Y_l due to a higher alignment score. The model is optimized by Eq. 1 to Eq. 3:

$$A = \beta \log \frac{\pi_\theta(Y_w|X)}{\pi_{ref}(Y_w|X)} \tag{1}$$

$$B = \beta \log \frac{\pi_\theta(Y_l|X)}{\pi_{ref}(Y_l|X)} \tag{2}$$

$$L_{DPO}(\pi_\theta; \pi_{ref}) = -E_{(X,Y_w,Y_l)\sim D} \left[\log \sigma \left(A - B\right)\right] \tag{3}$$

Language Neurons Identification

Following Tang et al. (2024), the neurons in our work are defined as a linear transformation of a single column in a weighted matrix W followed by a non-linear activation, SiLU (Shazeer 2020). For the j-th neuron in the i-th layer, its activation probability when processing responses in language k is computed as:

$$p^k_{i,j} = \mathbb{E}\left(\mathbb{I}\left(\mathrm{SiLU}(x^i W^i)_j > 0\right) \mid \text{language } k\right) \tag{4}$$

We define language-related neurons as those neurons exhibiting high activation probabilities for multiple but not all languages. However, methods that only focus on language specificity are defective in detecting the above neurons. To better detect the different types of neurons, we trade off two intrinsic properties of each neuron: (1) *Language-specificity*, which is quantified by the entropy of its activation probability distribution across languages; (2) *Effectiveness*, which measures the extent to which the neuron participates when the model solves tasks. And it is quantified by the neuron's maximum activation probability across different languages in our work. By doing so, we are better able to simultaneously identify neurons that are relatively more specific and relatively more generalized. They are combined into a unified metric as formulated in Eq 5:

$$\mathrm{score}_{i,j} = -\sum_{k=1}^{l} p'^k_{i,j} \log p'^k_{i,j} \; - \; \lambda \max_{1\le k\le l} p^k_{i,j}, \tag{5}$$

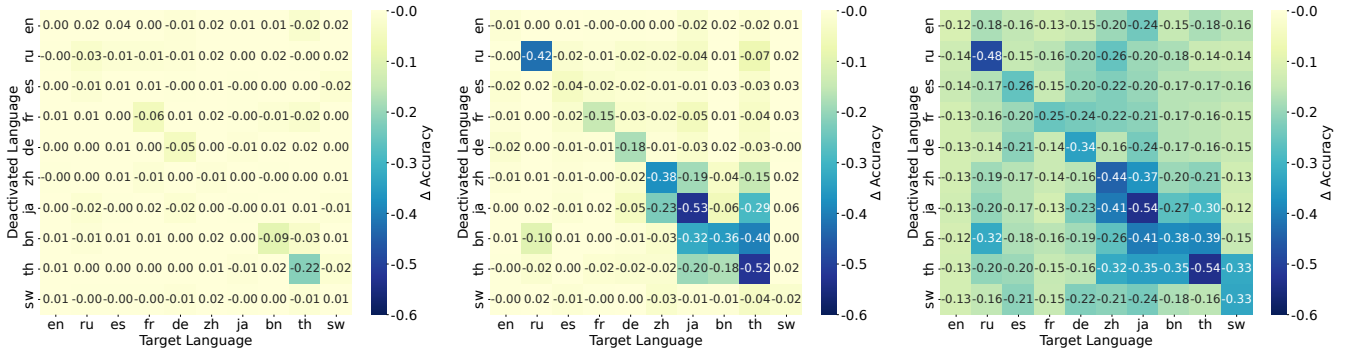
where $p'_{i,j}$ represents the probability distribution $p_{i,j}$ after normalization and λ is a balancing coefficient. Specifically, we automatically determine a value of λ such that general neurons are not identified as language-specific or language-related neurons shared by all l languages. Following Tang et al. (2024), neurons with scores falling in the lowest 1% are selected.

Furthermore, to identify how many languages each selected neuron is related to, we introduce a threshold τ:

$$N_{i,j} = \sum_{k=1}^{l} \mathbb{I}\big(p^k_{i,j} > \tau\big). \tag{6}$$

Following Tang et al. (2024), the threshold τ is set to the top 5% of all activation probabilities. A neuron is considered as a **language-specific neuron** if $N_{i,j} = 1$, and as a **language-related neuron** if $1 < N_{i,j} < l$. Meanwhile, a neuron is considered **general neuron** if it exhibits high activation probabilities across all l languages.

Finally, given our focus on multilingual reasoning tasks, we select neurons exclusively based on responses from multilingual reasoning datasets, rather than relying on multilingual plain text (Tang et al. 2024).



(a) Specific Neurons (b) Specific & Related Neurons (c) Specific & Related & General Neurons

Figure 2: Accuracy changes of MistralMathOctopus on MGSM after deactivating language-specific neurons or language-specific & language-related neurons or language-specific & language-related & general neurons. For comparison, the results of Tang et al. (2024) are provided in appendix.

Experiments

Experimental Setup

Following She et al. (2024), we conduct our experiments and analyzes on the mathematical reasoning tasks in different languages. In this section, we introduce our experimental settings in detail.

Models We include two different models in our experiments and analyses. Following She et al. (2024), we conduct our experiments on MistralMathOctopus-7B [1] and MetaMathOctopus-7B [2]. MistralMathOctopus is obtained by fine-tuning MetaMath-Mistral (Yu et al. 2023) with MGSM8KInstruct (Chen et al. 2023). MetaMathOctopus is obtained by fine-tuning MetaMath (Yu et al. 2023) with MGSM8KInstruct. Considering limited computational resources and reproducibility, we directly utilize the publicly released base models. Our analyses are mainly based on MistralMathOctopus in the main text and we report more results in the appendix.

Datasets We conduct experiments on two representative mathematical reasoning benchmarks, MGSM (Shi et al. 2022) and MSVAMP (Chen et al. 2023). MGSM is a widely used benchmark for multilingual mathematical reasoning evaluation. MSVAMP is an out-of-domain test set in contrast to MGSM, which evaluates robustness and generalization (Zhu et al. 2024; She et al. 2024).

Languages Following She et al. (2024), we choose the following 10 different languages for analysis. As a pivot language, English (en) is used as the alignment target. We also choose Chinese (zh), Russian (ru), German (de), French (fr), Spanish (es), Japanese (ja), Swahili (sw), Thai (th) and Bengali (bn) as 9 representative non-English languages.

Implementations Due to limited computational resources, our exploration focuses on the most effective DPO variant of MAPO (She et al. 2024). We select 1, 4, and 8 tasks from the NumGLUE (Mishra et al. 2022), an arithmetic reasoning benchmark, and translate questions into 9 languages, consistent with the MGSM, thereby creating a multilingual seed dataset. To construct preference pairs, we sample responses using the corresponding base models and employ NLLB-200-distilled-600M[3] as the translation model to obtain alignment scores. Finally, for each model and each target language (excluding English), we gain 10,000 preference pairs. Training is conducted using LoRA (Hu et al. 2022). During the neuron selection stage, we perform force-decoding on the responses of the MGSM or MSVAMP dataset to obtain the activation probabilities of neurons for each language. Additional implementation details are provided in appendix.

Language Neurons Identification

Based on the neuron identification algorithm, we identify the language-specific neurons, language-related neurons, and general neurons in the model. To further validate the effectiveness of our algorithm, we report the changes in Accuracy (defined as producing numerically correct answers in the correct language) after deactivating the identified neurons across different languages. Following Tang et al. (2024), we also report changes in the perplexity (PPL) scores of LLMs in Appendix. Experiments are conducted on the base model, with results presented in Figure 2. We report the results of two settings, deactivating language-specific neurons and deactivating language-specific & language-related neurons. The higher PPL change (darker cells) indicates the stronger reliance on the neurons be deactivated.

It can be found that **the performance for each language mostly relies on both its language-specific neurons and language-related neurons rather than other neurons**, as the diagonal elements in each row show the highest changes in PPL. However, **different languages vary in the extent to which they rely on their corresponding language-specific and language-related neurons**, reflecting differences in cross-lingual alignment across different languages.

[1] https://huggingface.co/kevinpro/MistralMathOctopus-7B
[2] https://huggingface.co/kevinpro/MetaMathOctopus-7B
[3] https://huggingface.co/facebook/nllb-200-distilled-600M

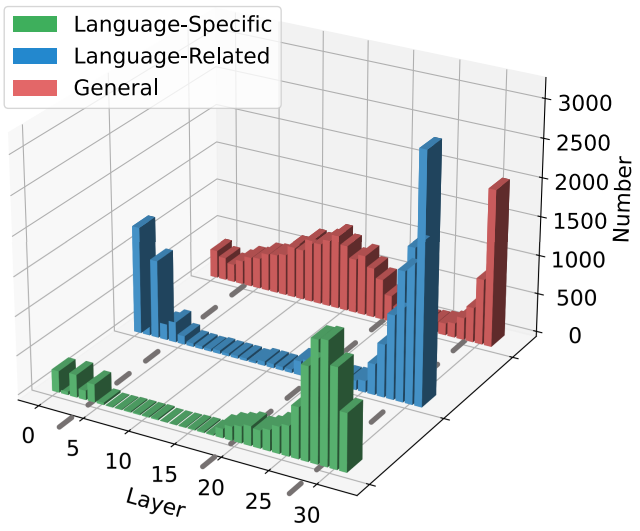


Figure 3: Layer-wise distribution of the different types of neurons of MistralMathOctopus on MGSM.

Notably, deactivating both language-specific and language-related neurons leads to a more pronounced effect compared to deactivating only language-specific neurons. Suggesting that for a given language, in addition to its language-specific neurons, **shared language-related neurons also significantly contribute to its performance**.

Also, deactivating all the language-related neurons of one language doesn't cause significant impacts on the model's performance in other languages, which indicates that the language-related neurons for a specific language are relatively evenly dispersed across multiple other languages. The above findings confirm the validity of the language neurons identified by our method and further provide insights into the characteristics of language neurons.

Layer-wise Functionality Analysis

Based on the identified neurons, we perform layer-wise functional analyses of all layers in the LLMs. We begin by analyzing the distributions of different types of neurons in the base model. And we report the results in Figure 3.

Some works have divided the LLM's multilingual inference process into three stages (Wendler et al. 2024; Zhao et al. 2024b). However, through the analysis of the distribution of different types of neurons, we find that although the combined function of the last stage is still to produce language-specific symbols, **the behavior of language-specific neurons first reaches its peaks and then declines, which differs substantially from the other two types**. So we further divide the final stage into two parts, leading to a four-stage interpretation of the LLMs' internal process for multilingual inference:

1. **Multilingual Understanding:** In the initial layers, the number of both language-specific and language-related neurons peaks, while the number of general neurons is relatively low. The model maps multilingual inputs into a unified semantic space at this stage.
2. **Shared Semantic Space Reasoning:** In the intermediate layers, the model engages in reasoning within a shared semantic space across different languages. During this stage, both language-specific and language-related neurons are largely absent, whereas general neurons become dominant.
3. **Multilingual Output Space Transformation:** The model transfers features into the multilingual output space in this stage in preparation for generating the final output. In this part, the number of both language-specific and language-related neurons reaches a peak again, while the number of general neurons drops to the lowest point.
4. **Vocabulary Space Outputting:** In the last layer, the model maps vectors of different languages into a shared vocabulary space to generate outputs. The total number of language-specific neurons and language-related neurons increases, but there is an opposite trend between language-specific and language-related neurons. Additionally, unlike assumptions in previous work (Wendler et al. 2024; Zhao et al. 2024b), the number of general neurons reaches maxima at this stage, consistent with the behavior in the early layers. We therefore speculate that this may be related to the shared vocabulary across different languages.

Meanwhile, the distribution of different types of neurons aligns with the conclusions from the existing studies mentioned above. Overall, the number of neurons varies correspondingly with the different inference stages of LLMs.

Layer-wise Neuron Changes Analysis

We further analyze the changes in different types of neurons before and after alignment. Based on the four functional stages above, we quantify the layer-wise changes (Δ) in the number of different types of neurons in Figure 4.

During the multilingual understanding stage, the number of both language-specific and language-related neurons in-

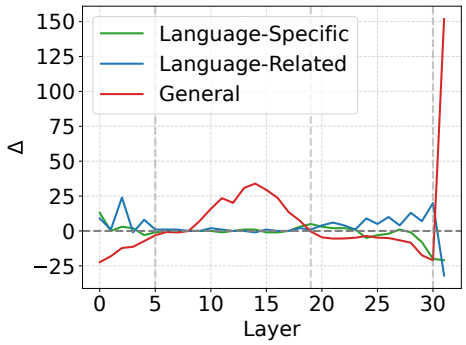


Figure 4: Layer-wise changes in the number of different types of neurons of MistralMathOctopus on MGSM.

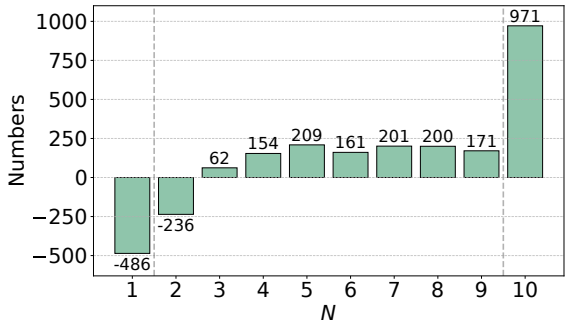


Figure 5: Changes in the number of neurons shared by N languages after alignment.

creases, while general neurons decrease. In the subsequent shared semantic space reasoning stage, general neurons increase substantially, whereas both language-specific neurons and language-related neurons remain stable and nearly absent.

In the third stage, as general neurons decrease, the sum of language-specific neurons and language-related neurons increase overall. Additionally, we notice that language-specific and language-related neurons show **opposite trends**. This reflects that the aligned model **relies more on shared language-related neurons than language-specific neurons in the third stage**, which is difficult to observe under the previous taxonomies.

Finally, in the last stage, the number of general neurons increases significantly in the aligned model, accompanied by a reduction in both language-specific and language-related neurons. This **contradicts the conventional three-stage assumption**, indicating that general neurons also play a significant role in the final vocabulary outputting stage. Additionally, combined with the observations in Figure 5, the model relies on a smaller set of language neurons with a higher degree of sharing in this stage. We also report the results of different checkpoints during the alignment process in appendix.

Overall, we find the *union set* of language-specific and language-related neurons and *general neurons* exhibit generally opposite trends across layers, which corresponds to the characteristics of LLMs at different stages. Especially, at the last stage, general neurons play an more important role. Multilingual alignment facilitates more effective activation of the appropriate neurons at each stage, thereby improving the model's capability to handle multilingual tasks.

Macroscopic Analysis of Neurons

We further conduct macroscopic analysis for different types of neurons. In our neuron identification algorithm, the number of languages that share a specific neuron is an attribution characterizing all types of activated neurons. Since our study involves 10 languages, the valid range of N is from 1 to 10. Among these, values of N from 2 to 9 correspond to language-related neurons. As special cases in our work, $N = 1$ represents language-specific neurons, while $N = 10$ corresponds to general neurons.

We report the changes in the number of neurons after multilingual alignment for each value of N ranging from 1 to 10 in Figure 5. The results show a decrease in the number of language-specific neurons, while an increase in the number of language-related neurons, which are shared across multiple languages. This indicates that **multilingual alignment encourages LLMs to develop and utilize more shared language-related neurons, rather than language-specific neurons**, which are applicable to only a single language. Meanwhile, during the alignment, the model improves its task-relevant common knowledge. Therefore, the number of general neurons also increases significantly. Language-related neurons and language-specific neurons exhibit overall opposite trends, further highlighting the importance of a more fine-grained analysis of the two categories within our new taxonomy. In addition, we also report the overall neuron numbers for each value of N before and after alignment in Appendix.

Spontaneous Multilingual Alignment Analysis

The "spontaneous multilingual alignment" phenomenon is first revealed and discussed by Zhang et al. (2024), which shows that conducting alignment in a small number of languages significantly improves multilingual alignment even between English and many languages unseen during the alignment process. We further analyze this phenomenon in our experiments. As shown in Table 1, spontaneous multilingual alignment also emerges under the multilingual alignment method employed in our study. Besides the languages used for alignment, LLMs exhibit notable performance gains in other unaligned languages. To understand how multilingual alignment generalizes to other languages, we analyze the changes in different types of neurons before and after multilingual alignment based on our method.

Taking the case of "zh/de ⇒ en" as a representative example, we report the average results in Table 2. For the trained languages, the number of language-specific neurons decreases, while the number of language-related neurons increases. This indicates the aligned languages tend to utilize more language-related neurons shared with other languages rather than exclusive language-specific neurons. Moreover,

MGSM	bn	th	sw	ja	zh	ru	de	es	fr	en	Avg.
base	43.6	53.2	50.4	55.6	59.6	59.2	61.2	62.8	56.8	75.6	57.8
zh/de ⇒ en	46.4	55.6	59.2	56.8	64.0	71.2	66.8	71.2	69.2	75.2	63.6
sw/th ⇒ en	48.8	58.8	59.2	56.4	68.4	68.4	69.2	69.6	70.4	77.6	64.7

Table 1: Accuracy of the MistralMathOctopus base model and aligned model on MGSM. "X/Y ⇒ T" indicates that languages X and Y are used for multilingual alignment.

Language	Language-Specific	Language-Related
Trained	-37	+232
Others	-36	+205

Table 2: Average results of neuron count changes across multiple languages. "Trained" indicates the trained languages in the spontaneous multilingual alignment experiment. "Others" indicates other languages except the trained languages. We round the results to the nearest integer.

Language	Language-Specific	Language-Related
English	46	603
non-English	613	2006

Table 3: Average number of different types of neurons for English and non-English languages of MistralMathOctopus on MGSM. We round the results to the nearest integer.

we extend this analysis to languages other than the trained languages and observe a similar phenomenon. These findings indicate **multilingual alignment facilitates the use of language-related neurons while reducing the reliance on language-specific neurons in both trained and other unseen languages**. We hypothesize that the new language-related neurons shared with trained languages contribute to the performance improvement on other unseen languages.

Further Analysis

Uniqueness of English Since current LLMs are primarily pretrained on English data, English is often regarded as playing a special role within LLMs (Wendler et al. 2024). In our work, we observe that English exhibits markedly different characteristics compared to other languages. Based on the identified neurons in our work, in Figure 2, deactivating the language neurons of English has a negligible impact on the model's performance in English, which is entirely different from the behavior observed in other languages. This is consistent with the results of Tang et al. (2024). Furthermore, we quantify the sum of language-specific neurons and language-related neurons for English and non-English languages based on the MistralMathOctopus base model (Table 3).

Our analysis reveals that English has significantly fewer neurons than other languages, both in terms of language-specific and language-related neurons. We hypothesize that this is due to that English actually possesses numerous language-related neurons. And since English serves as a pivot language, these language-related neurons are likely shared with almost all other languages. It causes them to be also regarded as general neurons.

Variable (%)	Language-Specific	Language-Related
Domain	80.7	92.3
Alignment	95.6	92.1

Table 4: Overlap ratio of different types of neurons across *different domains* and *before and after alignment*. Following She et al. (2024), MSVAMP is regarded as an out-of-domain dataset. The results of MistralMathOctopus on MGSM are used as the fiducial value.

Stability of Neuron Distributions We discuss the stability of neuron distributions across *different data domains*, as well as *before and after alignment*. To quantify the stability of neuron distributions, we compute the neuron overlap ratio in both settings, with the results summarized in Table 4. We can find that although the exact positions of a few neurons may vary across different settings, the positional distribution of most neurons remains stable. This also indicates good reliability and generalization of the language neurons identified under fixed hyperparameters.

Conclusion

In this work, we systematically investigate the multilingual alignment from the perspective of language neurons. First, based on the defects of the existing binary classification methodology, we propose a ternary classification methodology, which defines the language-specific neurons, language-related neurons, and general neurons. Then we propose a corresponding language neuron identification algorithm, which detects the above different types of neurons in LLMs.

Furthermore, we examine the multilingual alignment mechanism by analyzing the roles of different types of neurons. Based on their distributional characteristics, we categorize LLMs' internal process into four functional parts. Our analysis reveals that multilingual alignment enhances the model's utilization of the corresponding types of neurons across different functional parts. Meanwhile, we find that alignment promotes a greater reliance on shared language-related neurons across languages, rather than on language-specific neurons. This also corresponds to the phenomenon of spontaneous multilingual alignment.

Overall, we provide further analysis and valuable insights to better understand multilingual alignment and multilingual capabilities of LLMs.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120), research project of Nanjing University-China Mobile Joint Institute (NJ20250038), the Fundamental Research Funds for the Central Universities (No. 2024300507, 2025300390).

References

Chen, N.; Zheng, Z.; Wu, N.; Gong, M.; Zhang, D.; and Li, J. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Eronen, J.; Ptaszynski, M.; and Masui, F. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3): 103250.

Fan, Y.; Mu, Y.; Wang, Y.; Huang, L.; Ruan, J.; Li, B.; Xiao, T.; Huang, S.; Feng, X.; and Zhu, J. 2025. SLAM: Towards Efficient Multilingual Reasoning via Selective Language Alignment. *arXiv preprint arXiv:2501.03681*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, P.; Liu, S.; Gao, C.; Huang, X.; Han, X.; Feng, J.; Deng, C.; and Huang, S. 2024. Large Language Models Are Cross-Lingual Knowledge-Free Reasoners. *arXiv preprint arXiv:2406.16655*.

Huang, H.; Tang, T.; Zhang, D.; Zhao, W. X.; Song, T.; Xia, Y.; and Wei, F. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Ji, S.; Li, Z.; Paul, I.; Paavola, J.; Lin, P.; Chen, P.; O'Brien, D.; Luo, H.; Schütze, H.; Tiedemann, J.; et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Kojima, T.; Okimura, I.; Iwasawa, Y.; Yanaka, H.; and Matsuo, Y. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Z.; Winata, G. I.; and Fung, P. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. *arXiv preprint arXiv:2105.03953*.

Mishra, S.; Mitra, A.; Varshney, N.; Sachdeva, B.; Clark, P.; Baral, C.; and Kalyan, A. 2022. NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3505–3523. Dublin, Ireland: Association for Computational Linguistics.

Mousi, B.; Durrani, N.; Dalvi, F.; Hawasly, M.; and Abdelali, A. 2024. Exploring Alignment in Shared Cross-lingual Spaces. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6326–6348. Bangkok, Thailand: Association for Computational Linguistics.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3977–3986.

Nostalgebraist. 2020. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

She, S.; Zou, W.; Huang, S.; Zhu, W.; Liu, X.; Geng, X.; and Chen, J. 2024. MAPO: Advancing Multilingual Reasoning through Multilingual-Alignment-as-Preference Optimization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10015–10027. Bangkok, Thailand: Association for Computational Linguistics.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Tan, S.; Wu, D.; and Monz, C. 2024. Neuron Specialization: Leveraging intrinsic task modularity for multilingual machine translation. *arXiv preprint arXiv:2404.11201*.

Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, X.; Wei, F.; and Wen, J.-R. 2024. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5701–5715. Bangkok, Thailand: Association for Computational Linguistics.

Wendler, C.; Veselovsky, V.; Monea, G.; and West, R. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 15366–15394.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yu, L.; Jiang, W.; Shi, H.; Yu, J.; Liu, Z.; Zhang, Y.; Kwok, J. T.; Li, Z.; Weller, A.; and Liu, W. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Zhang, S.; Gao, C.; Zhu, W.; Chen, J.; Huang, X.; Han, X.; Feng, J.; Deng, C.; and Huang, S. 2024. Getting More from Less: Large Language Models are Good Spontaneous Multilingual Learners. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8037–8051. Miami, Florida, USA: Association for Computational Linguistics.

Zhang, S.; Liu, X.; Zhang, X.; Liu, J.; Luo, Z.; Huang, S.; and Gong, Y. 2025. Process-based Self-Rewarding Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 18097–18110. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Zhao, J.; Zhang, Z.; Gao, L.; Zhang, Q.; Gui, T.; and Huang, X. 2024a. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024b. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Zhao, Y.; Zhang, W.; Wang, H.; Kawaguchi, K.; and Bing, L. 2024c. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *arXiv preprint arXiv:2402.18913*.

Zhu, W.; Huang, S.; Yuan, F.; She, S.; Chen, J.; and Birch, A. 2024. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.