Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. 2023. Pixel representations for multilingual translation and data-efficient cross-lingual transfer. *arXiv preprint arXiv:2305.14280*.

Pratyusha Sharma, Jordan T Ash, and Dipendra Misra. 2023. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *arXiv preprint arXiv:2312.13558*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

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*.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual bert and the impact of fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227.

Eshaan Tanwar, Manish Borthakur, Subhabrata Dutta, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *CoRR*, abs/2306.05179.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. *arXiv preprint arXiv:2402.14700*.

Zhong Zhang, Bang Liu, and Junming Shao. 2023b. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024b. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *arXiv preprint arXiv:2402.18913*.

A English and Non-English Tokens

We employ `cld3` package to detect the language of each token in the embeddings of each layer, which is a language detection library based on the Compact Language Detector 3 model developed by Google. Furthermore, if the detection result is reliable, i.e., cld3.get_language(token).is_reliable $== True$, we adopt the detection results, otherwise the token is categorized as a non-word.

B Multilingual Corpus

Note that our selection criterion for the number of documents is based on achieving substantial coverage of each language's vocabulary, ensuring that the selected contexts provide a representative sample of the language, as shown in Table 7.

Table 7: Corpus details across languages are tailored to encompass the majority of each language's vocabulary, where "corpus size" indicates the number of contexts selected, "corpus vocab" represents the vocabulary coverage within the selected contexts, "vocab size" refers to the number of vocabularies of that language.

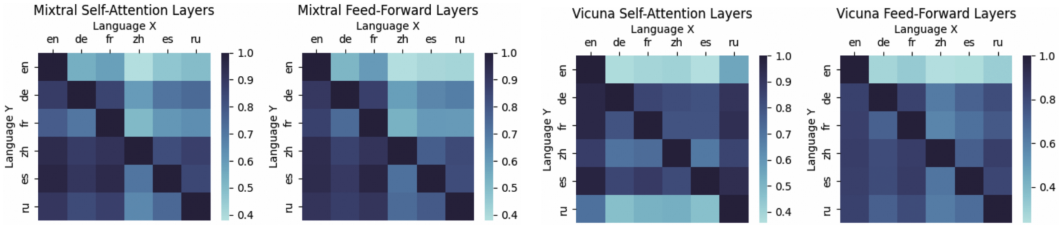
Language	En	De	Fr	Zh	Es	Ru
Corpus Size	180k	30k	50k	20k	20k	20k
Corpus Vocab	249k	154k	134k	198k	90k	144k
Vocab Size	273k	148k	135k	329k	93k	150k

C Interrelation of Language-Specific Neurons Across Languages

Using neurons identified by `PLND`, we investigate the relationships between languages via the degree of overlap among their language-specific neurons, defined as

$$\text{overlap}(x, y) = \frac{|\mathcal{N}_x \cap \mathcal{N}_y|}{|\mathcal{N}_y|}, \tag{11}$$

where $\mathcal{N}_{language}$ represents the set of detected language-specific neurons. Figure 5 shows the neuron overlapping ratio $\text{overlap}(x, y)$ of any two languages in different structures of two models.



(a) Mistral-7B-Instruct-v0.2. (b) Vicuna-7b-v1.5.

Figure 5: Overlapping ratio of language-specific neurons in self-attention and feed-forward structures.

We can observe that in both Mistral and Vicuna, the intersection with English from other languages is relatively limited (i.e., the first row of each figure), suggesting that English possesses a predominant number of language-specific neurons. Additionally, there is a pronounced tendency for languages belonging to the same family to demonstrate a higher degree of overlap with each other, such as Spanish, French, and English.

D Analysis on Different Multilingual LLMs

We further 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.

Hyper-Multilingual LLMs Figure 6 illustrates the degree of neuron overlap among languages within both the self-attention and feed-forward structures of BLOOMZ. In contrast to the findings shown in Figure 5, there is a marked reduction in overlap, indicating that individual languages maintain a higher degree of independence and do not extensively share neurons with one another.

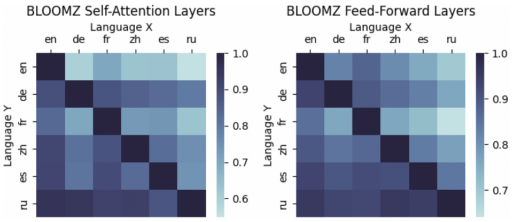


Figure 6: Overlapping ratio of language-specific neurons in BLOOMZ

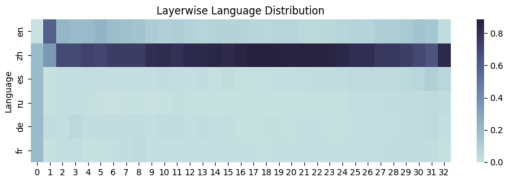


Figure 7: Ratio of languages among layers in Chinese Llama given non-English instructions.

Bilingual LLMs We employ Chinese Llama (Cui et al., 2023), which extends existing vocabulary and incorporate secondary pre-training using Chinese data and fine-tune the model with Chinese instruction datasets. However, this intensive training can lead to a degradation in performance for languages other than Chinese. As depicted in Figure 7, Chinese predominates as the primary language for reasoning processing and knowledge extraction across all languages. Consequently, the absence of language-specific neurons results in the transformation of it into a Chinese-centric LLM.

E Language-Agnostic Neurons

We initially implement a radical deactivation approach, wherein we specifically deactivate overlapping elements between each language and English. These elements precisely correspond to the intersecting neurons in the first column of Figure 5. Presented below are the comprehensive findings pertaining to Mistral. Our evaluation is centered around the reasoning task, which is recognized as the most indicative and challenging assessment for the model. We compare under the optimal "deactivating" method, which involves deactivating all language-specific neurons except those in S-ATTN.

Table 8: Performance of deactivating language-specific neurons without overlapped between English.

Language	Eng	non-Eng	Δ_{Eng}	$\Delta_{\text{non-Eng}}$	$\Delta\uparrow$
All language-specific neurons	46.2	18.3	$+0.2$	-8.0	$+8.2$
LSN without overlapped between English	45.8	20.2	-0.2	-6.1	$+5.9$

As evident by Table 8, the performance of English remains stable, contrasting sharply with the significant decline in the performance of multilingual. Removing overlapped neurons, as opposed to deactivating all language-specific neurons, leads to a less pronounced drop, yet the impact remains noteworthy. This demonstrates that overlapped neurons are not language-agnostic; they are not utilized for general comprehension and logical reasoning. Otherwise, the fundamental reasoning capacity and performance in multilingual contexts would remain unaffected. In addition, we retained the language-specific neurons that overlapped in all languages, meaning that we removed them from the language-specific neurons to be deactivated. Detailed results follow.

Table 9: Performance of deactivating language-specific neurons without all languages overlapped.

Language	Eng	non-Eng	Δ_{Eng}	$\Delta_{\text{non-Eng}}$	$\Delta\uparrow$
All language-specific neurons	46.2	18.3	$+0.2$	-8.0	$+8.2$
LSN without all languages overlapped	45.6	18.7	-0.4	-7.6	$+7.2$

The neurons that overlap across all languages only account for 0.02% of the total number of neurons. From the results in Table 9, we can see that the performance is almost the same as deactivating all language-specific neurons. This further proves that these neurons are not language-agnostic neurons, but only a subset of language-specific neurons.