Unraveling Babel: Exploring Multilingual Activation Patterns of LLMs and Their Applications

Weize Liu[1], **Yinlong Xu**[1], **Hongxia Xu**[1,2*], **Jintai Chen**[3], **Xuming Hu**[4,5*], **Jian Wu**[1,6]
[1]Liangzhu Laboratory, Zhejiang University [2]AI Research Center, WeDoctor Cloud
[3]University of Illinois Urbana-Champaign
[4]The Hong Kong University of Science and Technology (Guangzhou)
[5]The Hong Kong University of Science and Technology
[6]School of Public Health, Zhejiang University
{weizeliu1115,xuminghu97}@gmail.com einstein@zju.edu.cn

Abstract

Recently, large language models (LLMs) have achieved tremendous breakthroughs in the field of NLP, but still lack understanding of their internal neuron activities when processing different languages. We designed a method to convert dense LLMs into fine-grained MoE architectures, and then visually studied the multilingual activation patterns of LLMs through expert activation frequency heatmaps. Through comprehensive experiments on different model families, different model sizes, and different variants, we analyzed the similarities and differences in the internal neuron activation patterns of LLMs when processing different languages. Specifically, we investigated the distribution of high-frequency activated experts, multilingual shared experts, whether multilingual activation patterns are related to language families, and the impact of instruction tuning on activation patterns. We further explored leveraging the discovered differences in expert activation frequencies to guide sparse activation and pruning. Experimental results demonstrated that our method significantly outperformed random expert pruning and even exceeded the performance of unpruned models in some languages. Additionally, we found that configuring different pruning rates for different layers based on activation level differences could achieve better results. Our findings reveal the multilingual processing mechanisms within LLMs and utilize these insights to offer new perspectives for applications such as sparse activation and model pruning.

1 Introduction

Large language models (LLMs), proficient in utilizing a wide range of linguistic structures, have attained notable advancements in the field of natural language processing (Zhao et al., 2023). However, how an LLM uses multiple languages within one single structure remains elusive. Previous studies in neuroscience suggest that although there is considerable overlap in the brain regions involved in processing different languages, there are discernible differences (Crinion et al., 2006; Videsott et al., 2010; Friederici, 2011). Specifically, some certain regions appear to be specialized for particular languages, while some regions are language-agnostic. In the field of NLP, several recent studies (Tang et al., 2024; Zhao et al., 2024; Kojima et al., 2024) have investigated language-specific neurons within LLMs. But the internal mechanisms of LLMs in processing different languages and how to leverage these mechanisms remain insufficiently explored.

We still lack an intuitive understanding of the internal neuron activity of LLMs when processing different languages—it remains like a black box. This is due to the difficulty of decomposing LLMs into recognizable components. This presents a significant obstacle for us to better utilize LLMs. Therefore, we aim to investigate the differences and connections in internal neuron activations of LLMs when processing different languages.

We refer to the different activation patterns exhibited by internal neurons of LLMs when processing different languages as multilingual activation patterns. To open the black box and intuitively understand and explain the internal multilingual activation patterns of LLMs, we devised a method that involves converting dense LLMs into fine-grained MoE architectures and then calculating the activation frequencies of experts. Subsequently, we visualize the multilingual activation patterns of LLMs by heatmaps.

We conducted a comprehensive experimental study on the multilingual activation patterns of different LLMs, including various model families (Llama 2, Llama 3, Mistral), different sizes (7B–70B), and different variants (pre-trained and instruction tuned variants). We analyzed the distri-

* Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11855–11881
November 12-16, 2024 ©2024 Association for Computational Linguistics

bution of high-frequency activated experts, multilingual shared experts, whether the activation patterns of different languages are related to language families, and the impact of instruction tuning on activation patterns.

Considering that dense LLMs can be transformed into MoE architectures (Zhang et al., 2022), based on obtaining multilingual activation patterns, we further propose to leverage the activation frequency differences of various experts to guide sparse activation and pruning, aiming to minimize the amount of computation and inference cost, while maintaining model performance as much as possible. We advocate using only high-frequency experts for inference, excluding the parameters of other experts to achieve sparse activation and pruning. To this end, we propose two specific pruning schemes: pruning based on frequency thresholds and pruning based on frequency sorting. In experiments across various models and metrics, our method significantly outperformed random expert pruning and even exceeded the performance of the original unpruned models in some languages. This further validates the effectiveness of the multilingual activation patterns discovered by our method, providing a new feasible path for model pruning. Meanwhile, we found that equal proportional pruning at each layer is inferior to that of unequal pruning, which verifies the inter-layer differences in activation levels of LLMs. Therefore, we suggest configuring different pruning rates for different layers based on differences in activation levels.

Overall, we have not only analyzed the multilingual activation patterns within various LLMs, but also explored their connections with language families and instruction tuning. Furthermore, based on the identified multilingual activation patterns, we provide a new and effective approach to achieve sparse activation and pruning.

2 Related Work

Initially, in the task of neural machine translation (NMT) using small language models, some works (Lin et al., 2021b; Zhang et al., 2021; Xie et al., 2021) explored language-specific components. Armengol-Estapé et al. (2022) studied the exceptional performance of GPT-3 in Catalan, despite the small proportion of Catalan in the training corpus. Bhattacharya and Bojar (2023) evaluated the language specificity of the detector in the XGLM model using a parallel corpus of Czech and English, by conceptualizing the FFN as a system of detector, selector, and combiner. Several recent works (Tang et al., 2024; Zhao et al., 2024; Kojima et al., 2024) investigated the existence of language-specific neurons in LLMs and the modification of these neurons. Zhang et al. (2024) identified a core region corresponding to linguistic competence in LLMs through fine-tuning with various languages. In terms of MoE, the LLaMA-MoE model (Team, 2023) explored transforming Llama 2 7B into an MoE model and its continued training. However, exploring the multilingual activation patterns within LLMs and how to leverage these patterns remains an area that has not been sufficiently investigated.

3 Exploring Multilingual Activation Patterns in LLMs

Zhang et al. (2022) proposed MoEfication, exploring the conversion of the feed-forward networks (FFNs) in pre-trained Transformers into MoE structures without altering the original model parameters, and maintaining the performance on downstream tasks by conditionally selecting experts. Inspired by this, we propose to study the internal neuron activation patterns of LLMs when processing different languages by visualizing the expert activation frequency and its variations after converting the model to an MoE structure.

Expert Construction Our first step is to split the parameters of the FFNs into different experts. The FFNs in Llama/Mistral models comprise three layers: up-projection, gate-projection, and down-projection, as illustrated in the schematic diagram presented in Appendix A. Based on the intuition that the neurons with similar parameters exhibit similar activation patterns, we adopt the parameter clustering split method to cluster the parameters of each FFN layer, dividing them into different experts. Specifically, we perform balanced K-Means clustering (Malinen and Fränti, 2014) on the parameters of the up-projection layer, dividing it into 256 clusters. Then, we divide the neurons and their parameters of down-projection and gate-projection layers into different clusters based on the clustering results of the up-projection layer. Dividing neurons into different experts can significantly reduce the computational burden in subsequent experiments and enable us to directly observe the internal multilingual activation patterns of LLMs through heatmap visualization. To achieve fine-grained pa-

rameter splits while maintaining the visualization effects of heatmaps, we divide each FFN layer into 256 experts.

Cross-layer Expert Selection Next, we design the cross-layer expert selection method to identify experts with higher activation levels and frequencies as the language-specific experts for each language. LLMs contain multiple FFN layers, but the MoEfication method is limited to selecting experts within a single FFN layer, which fails to reflect the differences in activation levels across layers of varying depth within the LLM. Therefore, we need to extend the MoEfication method to cross-layer expert selection. For each input token, we use the sum of the activation values of all neurons for each expert as the score of this expert, representing the activation level. Given the direct incomparability of activation magnitudes across different FFN layers, to facilitate a global comparison of activation levels across layers, we perform a Z-score normalization on the scores of all experts within each layer. Subsequently, we rank the scores of all experts across all layers. We select approximately the top 10% of experts (for the 7B/8B models, we select the top 800, and for the 70B model, we select the top 2000) as the activated experts for a given input token, increasing their activation counts by 1. By normalizing and then performing cross-layer comparison, we better identified the experts whose activation values stood out relative to other experts.

Expert Activation Patterns After extensive testing with a multitude of tokens, we calculate the activation frequency for each expert (activation count/total number of tokens). Finally, the activation frequencies of all experts across n layers are compiled into an $n \times 256$ activation matrix for heatmap visualization.

4 Experiments

Models. Our experiments involve four models: Llama 2 7B (Touvron et al., 2023), Llama 3 8B, Llama 3 70B, and Mistral-7B-v0.3 (Jiang et al., 2023), along with their respective instruction tuning variants, to conduct a comprehensive study on the effects of different model families, model sizes, and instruction tuning.

Data. We selected the nine most widely spoken languages from the 46 languages contained in the ROOTS corpus (Laurençon et al., 2022) for experiments. The data sources are in Appendix B. The language families and genera of these languages are presented in Table 1. For each language, we test its activation pattern using 10,000 data samples, with a maximum input length of 200 tokens per sample.

Family	Genus	Language
Indo-European	Germanic	English (en)
Indo-European	Romance	French (fr)
Indo-European	Romance	Spanish (es)
Indo-European	Romance	Portuguese (pt)
Indo-European	Indic	Bengali (bn)
Indo-European	Indic	Urdu (ur)
Indo-European	Indic	Hindi (hi)
Afro-Asiatic	Semitic	Arabic (ar)
Sino-Tibetan	Chinese	Chinese (zh)

Table 1: The language families and genera. The ISO 639-1 language codes for each language are shown in parentheses in the Language column.

4.1 Multilingual Activation Patterns of LLMs

We present the activation pattern heatmaps of some models and languages in Figure 1, with the heatmaps of the remaining languages and models provided in Appendix C. We define layers closer to the input as shallow layers, with the shallowest layer being layer 0, and layers closer to the output as deep layers, with the deepest layer being layer 31/79. From the figure, it can be observed that for the Llama family models, the activation frequency differences between different experts in the shallow and middle layers are relatively small. But starting from a certain layer, the sparsity of expert activations significantly increases. Most experts exhibit lower activation frequencies, but a few experts have very high frequencies, indicating that significant differences in activation levels occur between different experts. Unlike the Llama family models, the Mistral model exhibits a distinct light-colored band in the middle layers, indicating that there is a portion of the middle layers with generally lower activation levels. By examining the heatmaps of other languages in Appendix C, we can observe that the positions of these light-colored bands in the Llama and Mistral models are consistent across all languages. However, the activation patterns for the same model also vary across different languages. For example, in Figure 1, it is evident that the activation frequency of shallow and intermediate experts is generally higher for

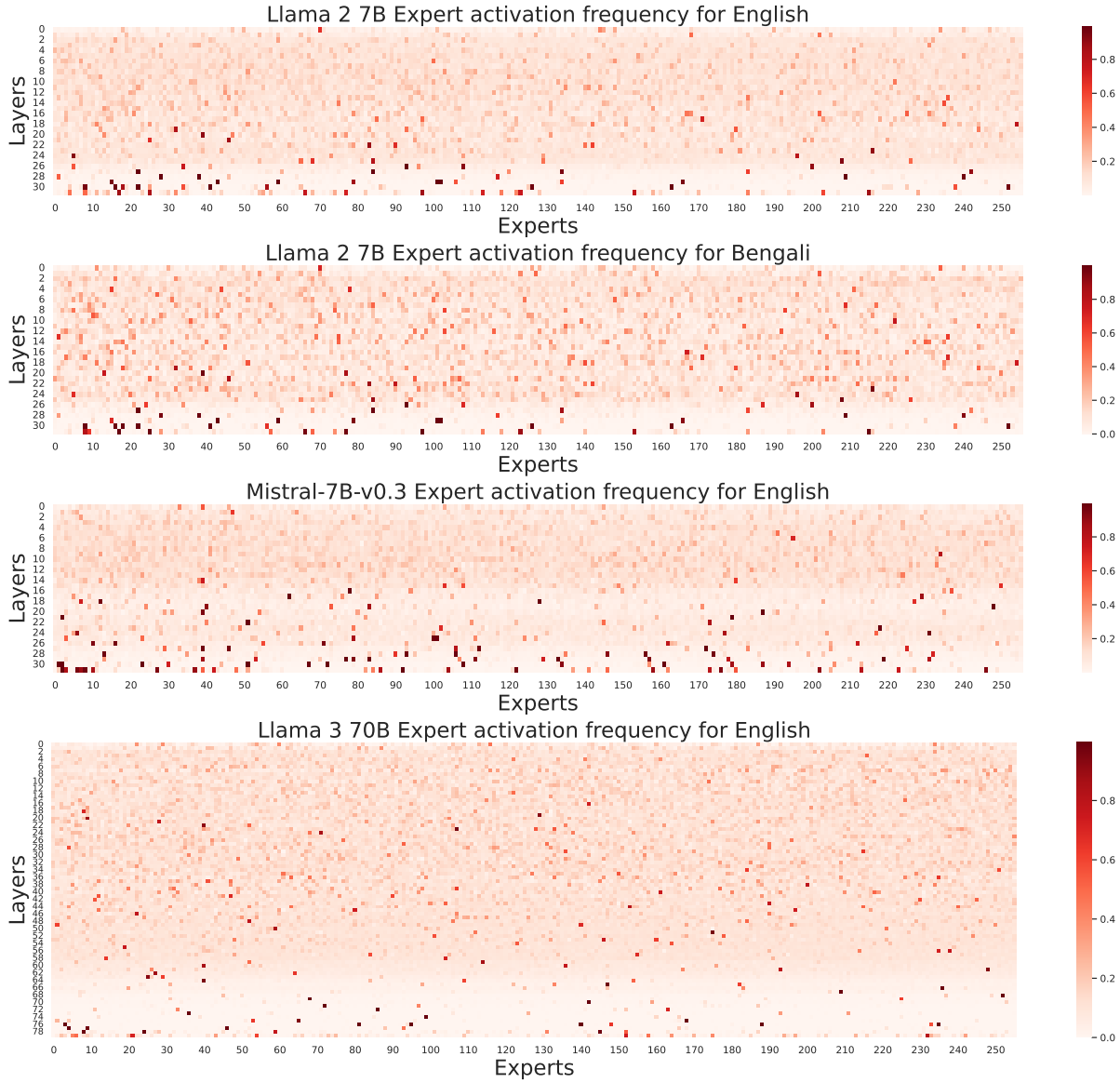


Figure 1: Heatmaps of activation patterns for some models and languages. Each heatmap is 32*256 (number of layers * number of experts), with darker colors indicating higher activation frequencies.

Bengali compared to English. By observing the activation patterns of different languages, we find that regardless of the model, the activation frequency of shallow- and middle-layer experts for English, French, Portuguese, and Spanish is significantly lower than that for Bengali, Urdu, and Hindi. This led us to speculate whether the activation patterns of different languages are related to their respective language families.

4.2 Are there connections at the language family level in the activation patterns?

To study the relationship between the multilingual activation patterns of LLMs and the language families and genera shown in Table 1, we calculated the similarity between the activation pattern matrices of different languages. We employed Euclidean distance, Kullback-Leibler (KL) divergence, and Pearson correlation coefficient as three measures to comprehensively reflect the similarity between different activation pattern matrices. The results of Llama 2 7B and Mistral-7B-v0.3 are illustrated in Figure 2, with similar results for other models listed in Appendix D. The results indicate that for all models, the activation patterns of three languages belonging to the Romance genus (French (fr), Spanish (es), Portuguese (pt)) and three languages of the Indic genus (Bengali (bn), Urdu (ur), Hindi (hi)) exhibit high similarity within the same genus. However, there is a significant difference in the activation patterns between these two genera. The activation pattern of English belonging to the Germanic genus is closer to that of the three languages belonging to the Romance genus, although

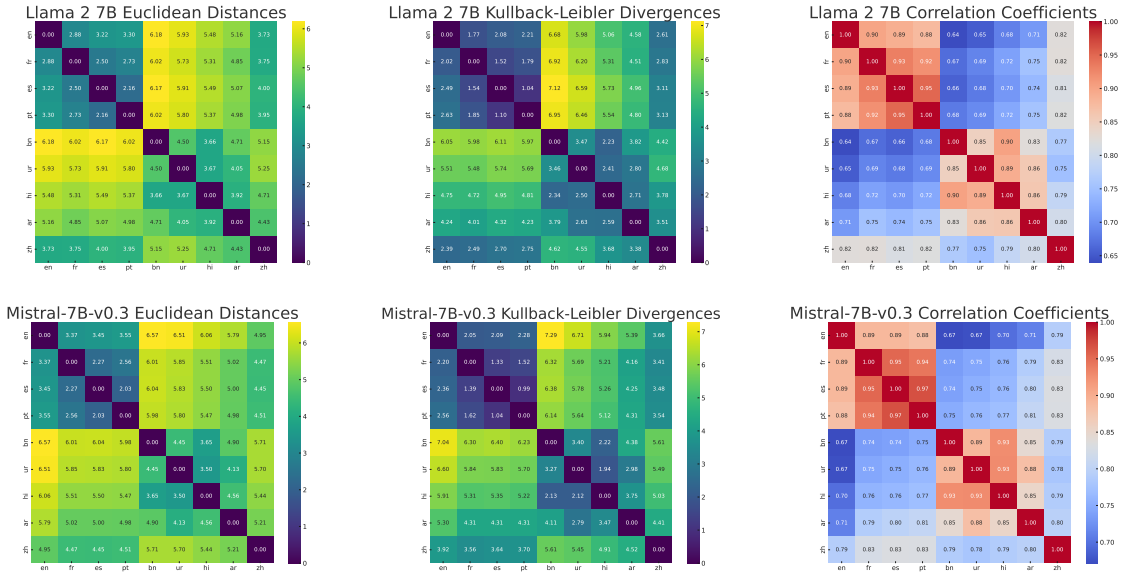


Figure 2: Heatmaps of similarity between activation pattern matrices for different languages in Llama 2 7B and Mistral-7B-v0.3. Each value in the Euclidean distance heatmaps represents the square root of the sum of the squares of the differences between corresponding elements of two matrices. Each value in the Kullback-Leibler (KL) divergence heatmaps represents the cumulative sum of the KL divergences computed row-wise between two matrices. Each value in the correlation coefficients heatmaps represents the mean of the Pearson correlation coefficients calculated row-wise between two matrices. The smaller the Euclidean distance/KL divergence, the more similar the two matrices are. The larger the correlation coefficient, the more similar the two matrices are.

there are still differences between them. This may be because English and Romance languages both use the Latin alphabet system and have a large number of loanwords and cognates (usually with the same or similar forms and meanings). Similarly, the activation pattern of Arabic is more similar to the languages of the Indic genus, possibly because Arabic has a large number of loanwords in Bengali, Urdu, and Hindi. Particularly, Urdu uses a modified Arabic alphabet system, with many words and expressions being very similar to Arabic, and both are written from right to left. In summary, we can confirm that the activation patterns of LLMs for different languages are closely related to the language families and genera to which these languages belong. Additionally, they may also be related to the alphabet systems and surface form similarities.

4.3 Multilingual shared experts

Based on the previous results, we can observe that some experts are frequently activated across different languages. We are also interested in understanding the distribution of these commonly highly activated experts. We define experts whose activation frequency for a given language is greater or equal to 0.05 as high-frequency experts of that language; otherwise, they are considered low-frequency experts. In Figure 3, we display the number of languages in which each expert is a high-frequency expert for Llama 2 7B and Mistral-7B-v0.3. For experts who are high-frequency experts in all 9 languages, we call them multilingual shared experts. For Llama 2 7B, the density of multilingual shared experts is significantly high in the middle layers (2–26 layers), while both shallower (0–2 layers) and deeper (26–32 layers) have sparse dark regions and more light regions. The results for Llama 3 8B and Llama 3 70B are provided in Appendix E. Their overall trends are consistent with those of Llama 2 7B. As observed in Section 4.1, the Mistral model has a low-activation band in the middle layers, but multilingual shared experts are still concentrated in the middle layers, while the shallow and deep layers are relatively sparse. This suggests that a large amount of neurons in the middle layers is language-independent and likely serves non-language-specific functions, such as various types of knowledge and abstract concepts that are independent of any particular language. In contrast, the activation patterns in the shallower and deeper layers are more closely related to language, exhibiting different activation patterns for different languages, indicating that these layers may undertake more language-specific processing functions. This may be because, in neural networks, the shallow layers typically learn low-level features of the

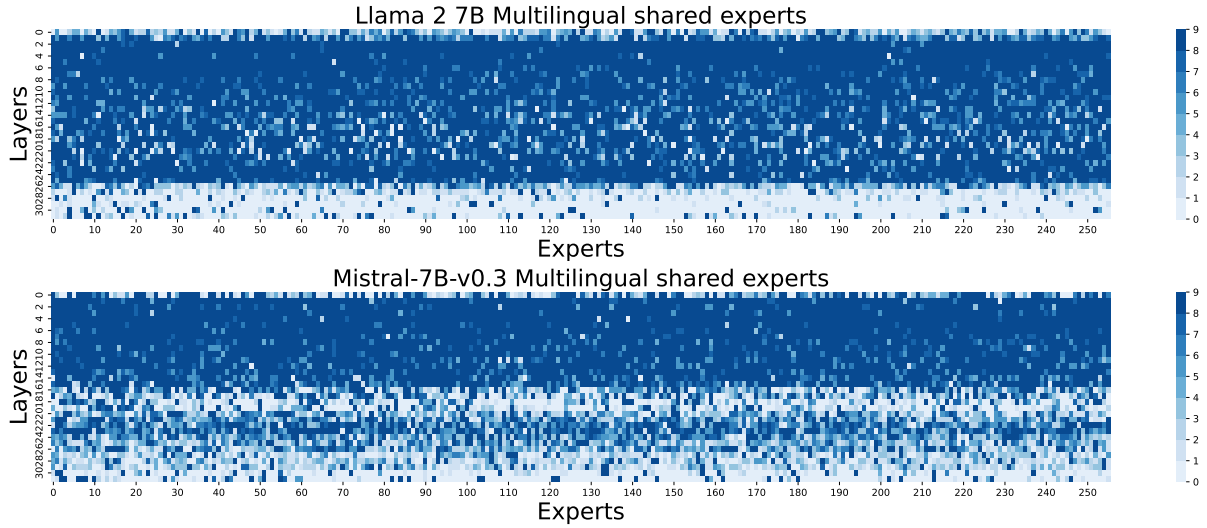


Figure 3: The heatmaps of multilingual shared experts for Llama 2 7B and Mistral-7B-v0.3. The color shade of each cell indicates how many languages the expert is a high-frequency expert in.

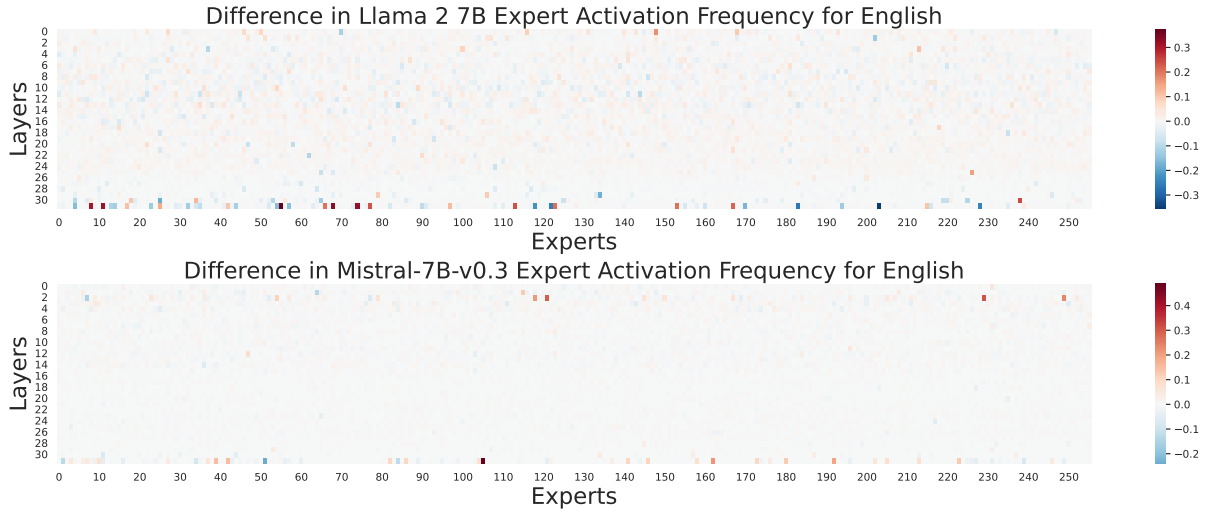


Figure 4: Changes in expert activation frequency for the instruction tuning variants of Llama 2 7B and Mistral-7B-v0.3 compared to the original pre-trained models.

input data, such as vocabulary, grammatical structures, and linguistic rules of different languages. Thus, the shallow layers contain a significant number of language-specific experts. In layers closer to the output, neurons need to generate outputs in specific languages. This necessitates adaptation to the linguistic structures of specific languages, resulting in a substantial number of language-specific experts in the deeper layers.

4.4 Impact of instruction tuning on activation patterns

In the previous experiments, we primarily focused on pre-trained base models rather than instruction-tuned models to minimize interference from other factors. But we also want to understand how the multilingual activation patterns of LLMs change after instruction tuning. Specifically, does instruction tuning exhibit certain patterns in its impact on multilingual activation? To this end, we tested the changes in expert activation frequency of the instruction tuning variants of the four models mentioned above compared to the original pre-trained model using the same expert split. The results for English using Llama 2 7B and Mistral-7B-v0.3 are shown in Figure 4, while the results for other models and languages are provided in Appendix F. The values in the heatmap represent the frequency of expert activations for the instruction tuning variant minus the frequency of expert activations for the pre-trained model. We found that in all models and languages, the changes in the last layer are significantly larger than those in other layers. In some experts, the activation frequencies increase (red), while in others they decrease (blue). We hope these findings can enhance our understanding of instruction tuning, aiding in more efficient and effective instruction tuning in the future.

Language	Expert activation frequency $\geq 5\%$			Expert activation frequency $\geq 1\%$			Origin
	Proportion	Random	Experts	Proportion	Random	Experts	
English	77.4%	3329.60 $\pm$ 4646.53	12.91	91.4%	13.57 $\pm$ 1.32	8.01	6.89
French	76.7%	37.49 $\pm$ 2.61	16.54	93.3%	11.43 $\pm$ 0.23	7.62	5.83
Spanish	76.3%	52.42 $\pm$ 4.56	21.52	93.2%	12.29 $\pm$ 2.35	8.96	6.71
Portuguese	76.0%	3445.22 $\pm$ 4804.65	22.68	93.4%	11.11 $\pm$ 0.40	8.34	6.19
Bengali	67.3%	274466.25 $\pm$ 387831.01	14.23	90.2%	20.18 $\pm$ 16.63	4.89	3.43
Urdu	68.5%	416160.50 $\pm$ 568821.95	13.24	90.3%	11.82 $\pm$ 2.85	7.18	4.54
Hindi	70.6%	22575.46 $\pm$ 23516.65	8.24	90.6%	6.31 $\pm$ 0.36	4.13	3.51
Arabic	71.4%	512693.39 $\pm$ 725010.36	13.52	90.8%	10.11 $\pm$ 0.46	7.14	4.58
Chinese	76.5%	19964.56 $\pm$ 28191.94	17.19	92.9%	10.32 $\pm$ 0.65	8.54	5.57
Average	73.41%	139191.65	15.56	91.79%	11.90	7.20	5.25

Table 2: The perplexity results of Llama 2 7B. The smaller the value, the better the model performance.

Language	Expert activation frequency $\geq 5\%$			Expert activation frequency $\geq 1\%$			Origin
	Proportion	Random	Experts	Proportion	Random	Experts	
English	77.4%	38.6 $\pm$ 3.5	45.2	91.4%	49.3 $\pm$ 2.3	51.5	53.7
French	76.7%	25.7 $\pm$ 0.7	26.5	93.3%	41.9 $\pm$ 0.6	**42.9**	42.9
Spanish	76.3%	16.3 $\pm$ 12.7	30.1	93.2%	41.5 $\pm$ 1.0	42.9	43.0
Portuguese	76.0%	18.6 $\pm$ 13.2	29.4	93.4%	35.9 $\pm$ 1.6	**36.9**	36.6
Chinese	76.5%	13.5 $\pm$ 11.5	30.3	92.9%	34.9 $\pm$ 0.9	**37.2**	36.9
Average	76.6%	22.5	32.3	92.8%	40.7	42.3	42.6

Table 3: Accuracy (%) of Llama 2 7B on the X-CSQA dataset.

Language	Expert activation frequency $\geq 5\%$			Expert activation frequency $\geq 1\%$			Origin
	Proportion	Random	Experts	Proportion	Random	Experts	
English	77.4%	2.0 $\pm$ 1.4	4.4	91.4%	11.5 $\pm$ 1.0	13.6	16.0
French	76.7%	1.5 $\pm$ 1.0	2.4	93.3%	8.6 $\pm$ 1.6	9.2	13.6
Spanish	76.3%	1.1 $\pm$ 1.2	3.6	93.2%	6.8 $\pm$ 0.6	**12.4**	10.0
Average	76.8%	1.5	3.5	92.6%	8.9	11.7	13.2

Table 4: Accuracy (%) of Llama 2 7B on the MGSM dataset.

4.5 Can expert activation frequencies guide sparse activation and model pruning?

After obtaining the multilingual activation patterns of LLM, we can observe differences in the activation frequencies of different experts, which reflects a certain sparsity in model activation. Building upon our prior application of the MoEfication to convert dense LLMs into fine-grained MoE architectures, during the inference process, we can reduce the amount of computation and lower FLOPs by activating only a subset of high-frequency experts, thereby significantly decreasing inference costs. Furthermore, we wonder whether it is possible to perform language-specific model pruning based on the different experts that are frequently activated by each language. Therefore, we further explored two pruning methods: (1) For each language, inference is conducted using only the high-frequency experts whose activation frequency is greater than or equal to a certain threshold. (2) Sort the experts in each layer by activation frequency, and only using the top $n\%$ of experts based on activation frequency.

Evaluation. We evaluate the performance changes of the models before and after pruning using perplexity (PPL) and accuracy on two datasets: X-CSQA (Lin et al., 2021a; Talmor et al., 2019) (commonsense question answering task) and MGSM (Cobbe et al., 2021; Shi et al., 2022) (grade-school math problems). For the perplexity test, we use 1,000 data samples from the ROOTS corpus that are different from the samples used in the expert activation pattern tests, with a maximum input length of 200 tokens per sample. Since the X-CSQA dataset does not publicly provide test set labels, we use 1,000 instances from the dev set for testing, with each question having five options. For each dataset, we selected several languages

	Expert activation frequency $\geq 0.5\%$			Expert activation frequency $\geq 0.1\%$			
Language	Proportion	Random	Experts	Proportion	Random	Experts	Origin
English	90.1%	7.7 ± 5.8	70.3	93.3%	7.9 ± 11.1	**76.7**	76.0
French	87.7%	3.3 ± 2.4	66.2	93.2%	11.9 ± 16.8	**67.0**	66.9
Spanish	87.5%	8.2 ± 11.2	67.1	93.1%	15.8 ± 21.3	**69.4**	68.2
Portuguese	87.5%	25.2 ± 18.5	65.9	93.2%	46.0 ± 7.5	**68.7**	66.5
Urdu	85.7%	0.3 ± 0.4	55.0	91.7%	12.9 ± 12.4	57.6	58.4
Hindi	86.0%	3.4 ± 2.4	57.0	90.5%	14.0 ± 12.6	57.2	60.0
Arabic	84.8%	0.0 ± 0.0	63.0	91.7%	11.8 ± 3.6	63.2	64.3
Chinese	87.0%	3.3 ± 4.2	59.9	92.8%	22.2 ± 15.7	**64.1**	63.3
Average	87.0%	6.4	63.1	92.4%	17.8	**65.5**	65.5

Table 5: Accuracy (%) of Llama 3 70B on the X-CSQA dataset.

	Expert activation frequency $\geq 0.5\%$			Expert activation frequency $\geq 0.1\%$			
Language	Proportion	Random	Experts	Proportion	Random	Experts	Origin
English	90.1%	1.6 ± 1.4	62.0	93.3%	8.3 ± 5.5	79.6	81.6
French	87.7%	2.7 ± 2.7	54.8	93.2%	5.9 ± 4.2	62.0	63.2
Spanish	87.5%	1.5 ± 1.0	66.4	93.1%	8.1 ± 7.9	75.2	75.6
Bengali	87.9%	0.0 ± 0.0	30.0	91.8%	4.1 ± 3.9	35.2	36.8
Average	88.3%	1.5	53.3	92.9%	6.6	63.0	64.3

Table 6: Accuracy (%) of Llama 3 70B on the MGSM dataset.

that overlap with those studied in our research and for which the original model can produce reasonable outputs.[1]

4.5.1 Pruning based on frequency thresholds.

To comprehensively study the impact of different model variants and sizes, we conducted experiments on Llama 2 7B, Llama 2-Chat 7B, and Llama 3 70B. Tables 2 to 6 present partial results, with additional results provided in Appendix G. During inference, for each language, we only use the experts whose activation frequency for that language is greater than or equal to a specified threshold, excluding the parameters of other experts. For Llama 2 7B and Llama 2-Chat 7B, we experimented with thresholds of 5% and 1%, which reduced the FFN layer parameters by approximately 25% and 10%, respectively. For Llama 3 70B, due to the generally low activation frequencies of experts, we experimented with thresholds of 0.5% and 0.1%, which reduced the FFN layer parameters by approximately 13% and 8%, respectively. Based on the model structure calculations, reducing the FFN layer parameters by 25% can decrease total inference FLOPs by approximately 18%, while a 13% reduction can lower FLOPs by approximately 9%.

To demonstrate the effectiveness of our method, we compare it with random expert selection. Specifically, we conduct experiments using the same proportion of randomly selected experts at each layer. In each table, the Proportion column represents the ratio of the experts used to the total number of experts. The Random column shows the results of using only randomly selected experts. We experimented with three different random seeds and reported the means and standard deviations. The Experts column displays the results of using only experts whose activation frequency is greater than or equal to the specified threshold. The Origin column presents the results of using the original model without pruning. We bold the results where the pruned models perform better than or equal to the original models.

We can observe that the model performance significantly declines when only using randomly selected experts. Additionally, the standard deviation of results from randomly selected experts is quite large, especially when using a higher pruning ratio. This indicates that the selection of different experts greatly affects the model performance. Pruning based on our method, which considers whether the expert activation frequency exceeds a threshold, can maintain the model performance as much as possible, far surpassing the random selection of experts. Surprisingly, when using our method with a lower threshold for pruning, all three models even outperformed the original models in some languages. This further demonstrates the correct-

[1]For instance, the Llama 2 7B and Llama 2-Chat 7B models are almost incapable of correctly answering questions in Urdu, Hindi, and Arabic, so we did not conduct experiments on these languages.

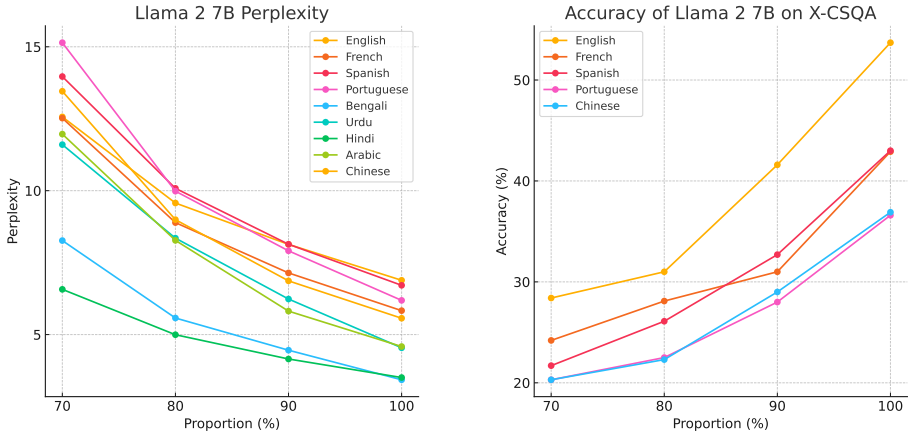


Figure 5: Results of Llama 2 7B pruning based on frequency sorting.

ness of the expert activation frequency differences identified by our method, providing a new feasible path for model pruning.

4.5.2 Pruning based on frequency sorting.

We also experimented with sorting the experts in each layer by activation frequency and using only the top 70%, 80%, and 90% of the experts based on activation frequency. The results for Llama 2 7B are shown in Figure 5, and those for Llama 3 70B in Figure 17. We found that the model performance deteriorates rapidly with increasing pruning ratios. With similar pruning ratios, the perplexity is comparable to pruning based on frequency thresholds, but the performance on the X-CSQA dataset is significantly worse than pruning based on frequency thresholds. The effect of equal proportional pruning at each layer is inferior to that of unequal pruning, reflecting the inter-layer differences in activation levels of LLMs. This validates our earlier finding that the sparsity of expert activation levels varies across different layers. Therefore, we recommend configuring different pruning rates for different layers based on the differences in activation levels.

5 Conclusion

In this study, we investigated the multilingual activation patterns in various LLMs from the perspective of MoE models. We also explored their connections with language families and instruction tuning, as well as the potential for guiding sparse activation and model pruning. These findings can assist us in better developing and utilizing the multilingual capabilities of LLMs. We hope these findings inspire new research in related fields.

6 Limitations

Despite achieving some meaningful conclusions in our research, there are still some limitations.

The limitations of pruning ratio. Section 4.1 and previous studies (Mirzadeh et al., 2023) have demonstrated that models like Llama 2 exhibit low activation sparsity in most layers, thus the pruning ratio we use is not very high. However, we believe that we offer new perspectives for applications such as model pruning, providing a foundation for further exploration in the future.

The limitations of interpretability by using a simplified model. Some work (Friedman et al., 2023) on mechanistic interpretability may present the concern that "even if the simplified representations can accurately approximate the full model on the training set, they may fail to accurately capture the model's behavior out of distribution." Thus, in order to verify the effectiveness of the high-frequency activated experts identified by our method for each language, we test the capability of the model pruned based on expert activation frequency and compare it with pruning done at the same ratio randomly. When analyzing the expert activation frequency of LLMs across different languages, we used the ROOTS corpus derived from Wikipedia. However, to evaluate the performance of the pruned models, we employed three distinct tasks for a comprehensive evaluation: the perplexity on another portion of the ROOTS corpus, the accuracy on the X-CSQA dataset and the MGSM dataset. We can see that the tests on the X-CSQA dataset and the MGSM dataset essentially evaluate the model's out-of-distribution capability. The re-

sults from these two datasets indicate that our models pruned based on expert activation frequency maintain performance comparable to the original models, and even exceed the original models in some languages. To a certain extent, this provides evidence that our approach enables simplified models to capture the out-of-distribution behaviors of the original models, thereby mitigating concerns.

In the future, we plan to extend our experiments to a broader array of models and languages. In addition, we will explore how to further leverage these insights to utilize and enhance the multilingual capabilities of LLMs.

7 Ethical Considerations

The use of AI assistants. We employed ChatGPT to assist us in polishing our paper and writing code.

Acknowledgements

This research was partially supported by the Key R&D Program of Zhejiang under grant No. 2024SSYS0026.

References

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Sunit Bhattacharya and Ondřej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 120–126, Singapore. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jenny Crinion, Robert Turner, Alice Grogan, Takashi Hanakawa, Uta Noppeney, Joseph T Devlin, Toshihiko Aso, Shinichi Urayama, Hidenao Fukuyama, Katharine Stockton, et al. 2006. Language control in the bilingual brain. *Science*, 312(5779):1537–1540.

Angela D Friederici. 2011. The brain basis of language processing: from structure to function. *Physiological reviews*.

Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. 2023. Interpretability illusions in the generalization of simplified models. In *Forty-first International Conference on Machine Learning*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021)*. To appear.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021b. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305.

Mikko I Malinen and Pasi Fränti. 2014. Balanced k-means for clustering. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, pages 32–41. Springer.

Seyed Iman Mirzadeh, Keivan Alizadeh-Vahid, Sachin Mehta, Carlo C del Mundo, Oncel Tuzel, Golnoosh Samei, Mohammad Rastegari, and Mehrdad Farajtabar. 2023. Relu strikes back: Exploiting activation sparsity in large language models. In *The Twelfth International Conference on Learning Representations*.

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*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. 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*.

LLaMA-MoE Team. 2023. Llama-moe: Building mixture-of-experts from llama with continual pretraining.

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*.

Gerda Videsott, Bärbel Herrnberger, Klaus Hoenig, Edgar Schilly, Jo Grothe, Werner Wiater, Manfred Spitzer, and Markus Kiefer. 2010. Speaking in multiple languages: Neural correlates of language proficiency in multilingual word production. *Brain and language*, 113(3):103–112.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *Ninth International Conference on Learning Representations 2021*.

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*, pages 877–890, Dublin, Ireland. Association for Computational Linguistics.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

A The FFN Structure

When calculating the scores for each expert, the activation values that we use are the hidden representations before the down-projection layer, as indicated by the red section in Figure 6.

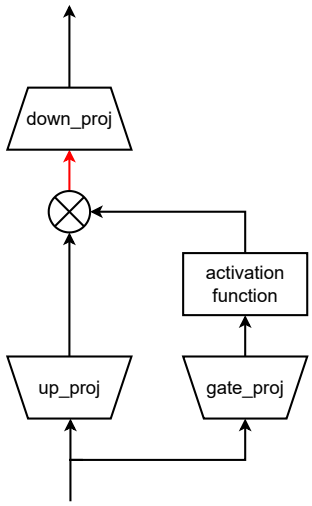


Figure 6: Schematic diagram of the FFN structure in Llama/Mistral model series.

B Data Sources

We obtained the ROOTS corpus data from Hugging Face, with their website URL presented in Table 7.

C Expert Activation Pattern Heatmaps

In Figures 7 to 10, we present the activation pattern heatmaps for the remaining models and languages.

D Similarity between Different Activation Matrices

Figure 11 presents the heatmaps of similarity between activation pattern matrices for different languages in Llama 3 8B and Llama 3 70B. It can be seen that these results are generally consistent with the patterns observed in Figure 2.

E Multilingual shared expert Heatmaps

In Figure 12, we display the number of languages in which each expert is a high-frequency expert for Llama 3 8B and Llama 3 70B.

Language	URL
English	https://huggingface.co/datasets/bigscience-data/roots_en_wikipedia
French	https://huggingface.co/datasets/bigscience-data/roots_fr_wikipedia
Spanish	https://huggingface.co/datasets/bigscience-data/roots_es_wikipedia
Portuguese	https://huggingface.co/datasets/bigscience-data/roots_pt_wikipedia
Bengali	https://huggingface.co/datasets/bigscience-data/roots_indic-bn_wikipedia
Urdu	https://huggingface.co/datasets/bigscience-data/roots_indic-ur_wikipedia
Hindi	https://huggingface.co/datasets/bigscience-data/roots_indic-hi_wikipedia
Arabic	https://huggingface.co/datasets/bigscience-data/roots_ar_wikipedia
Chinese	https://huggingface.co/datasets/bigscience-data/roots_zh-cn_wikipedia

Table 7: Wikipedia dataset URLs for various languages.

F Heatmaps of the impact of instruction tuning

In Figures 13 to 16, we present the heatmaps of changes after instruction tuning for the remaining models and languages.

G Pruning results

Tables 8, 9, and 10 present the remaining results of pruning based on frequency thresholds. Figure 17 illustrates the pruning based on frequency sorting results for Llama 3 70B.

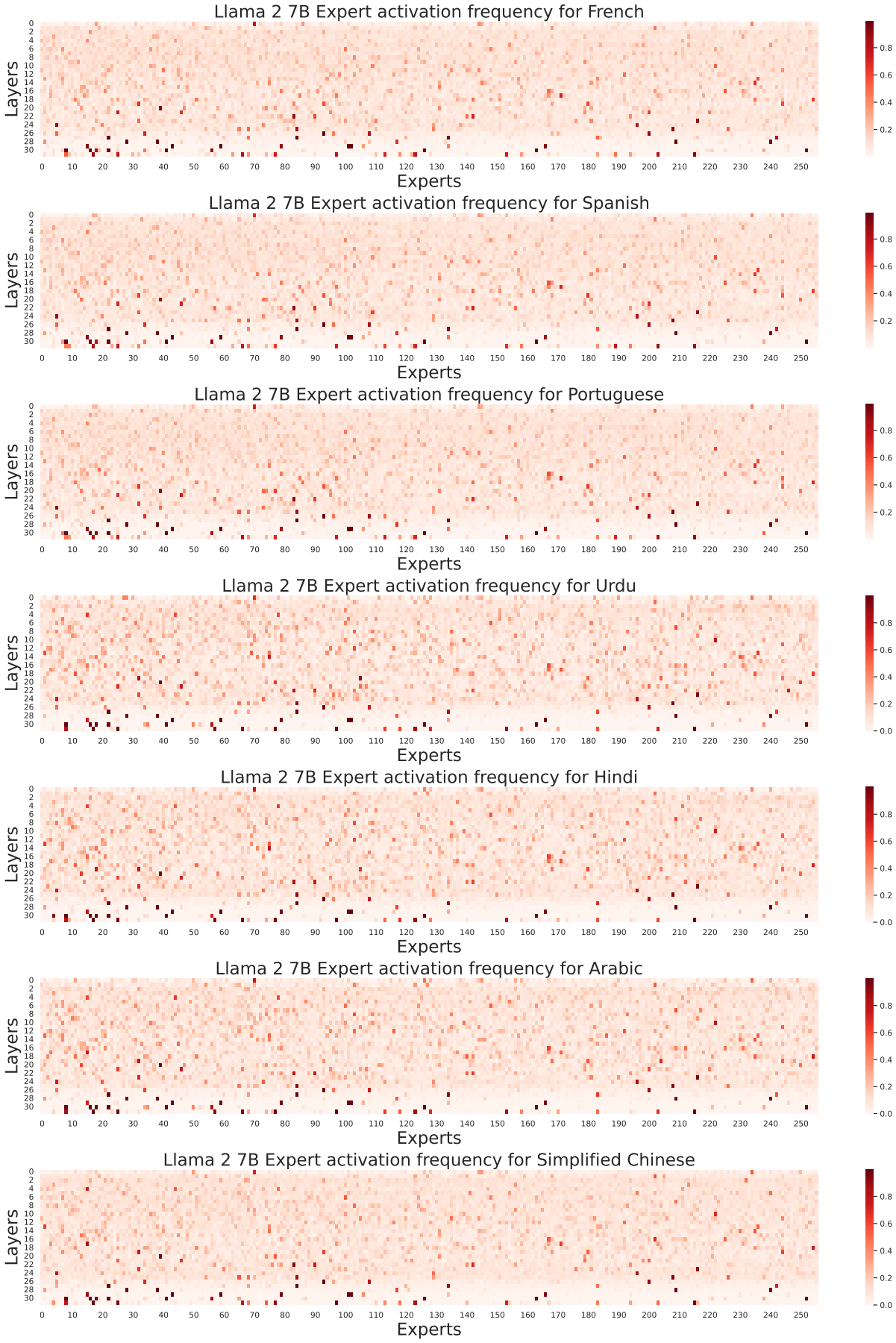
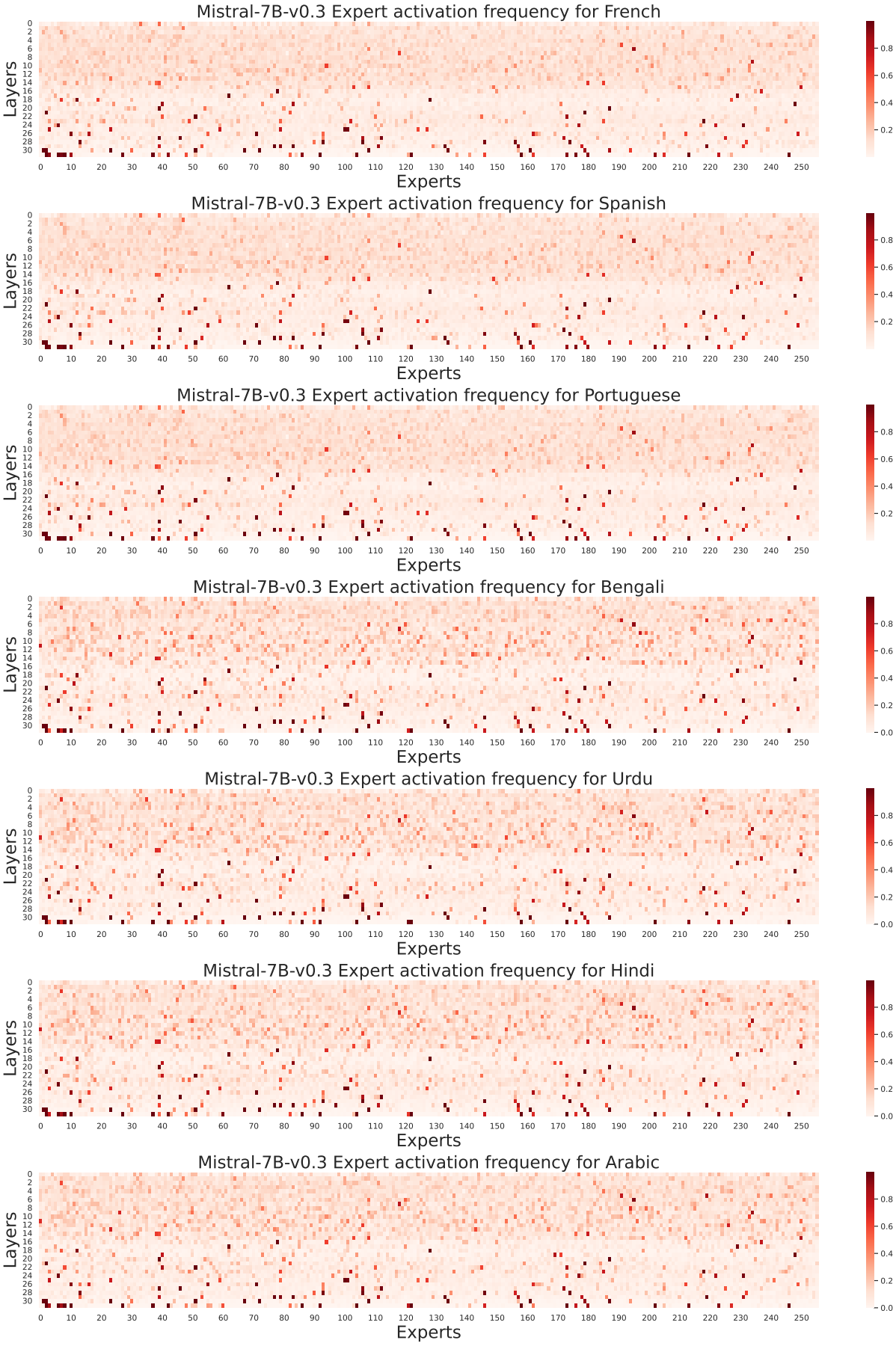


Figure 7: Heatmaps of activation patterns across languages for Llama 2 7B.


Mistral-7B-v0.3 Expert activation frequency for French
Mistral-7B-v0.3 Expert activation frequency for Spanish
Mistral-7B-v0.3 Expert activation frequency for Portuguese
Mistral-7B-v0.3 Expert activation frequency for Bengali
Mistral-7B-v0.3 Expert activation frequency for Urdu
Mistral-7B-v0.3 Expert activation frequency for Hindi
Mistral-7B-v0.3 Expert activation frequency for Arabic
Layers
Experts

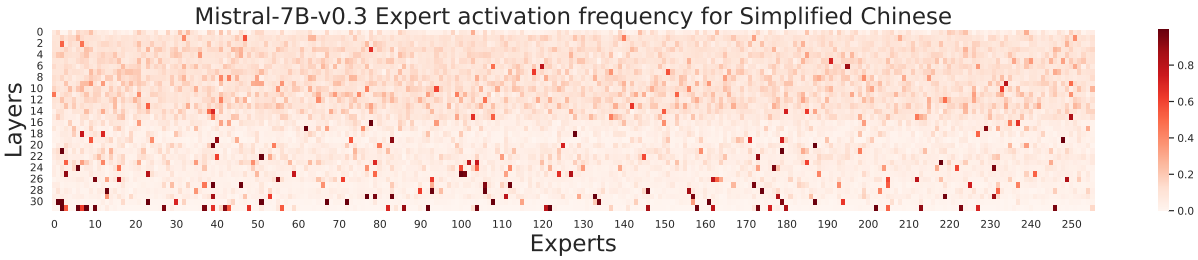
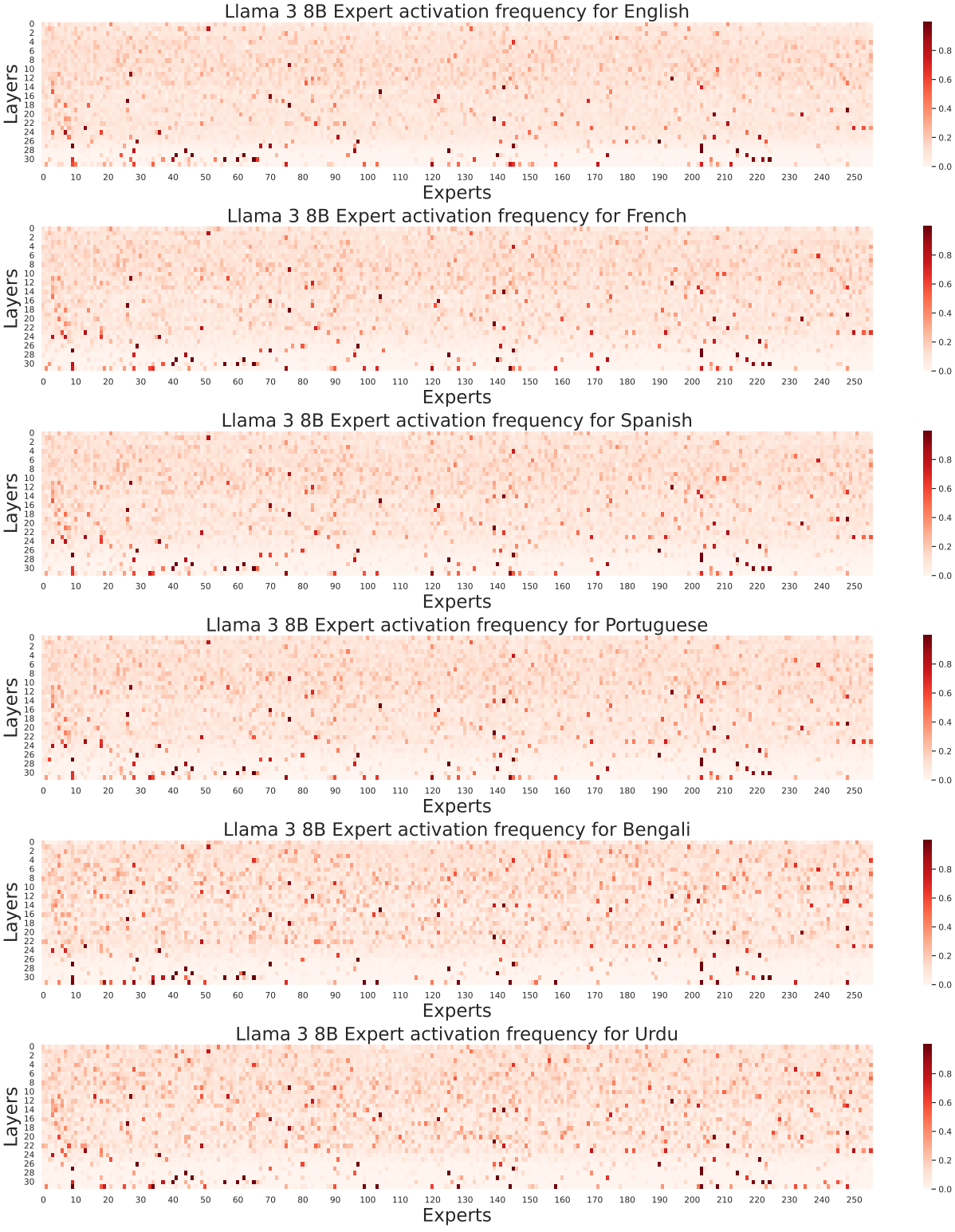


Figure 8: Heatmaps of activation patterns across languages for Mistral-7B-v0.3.



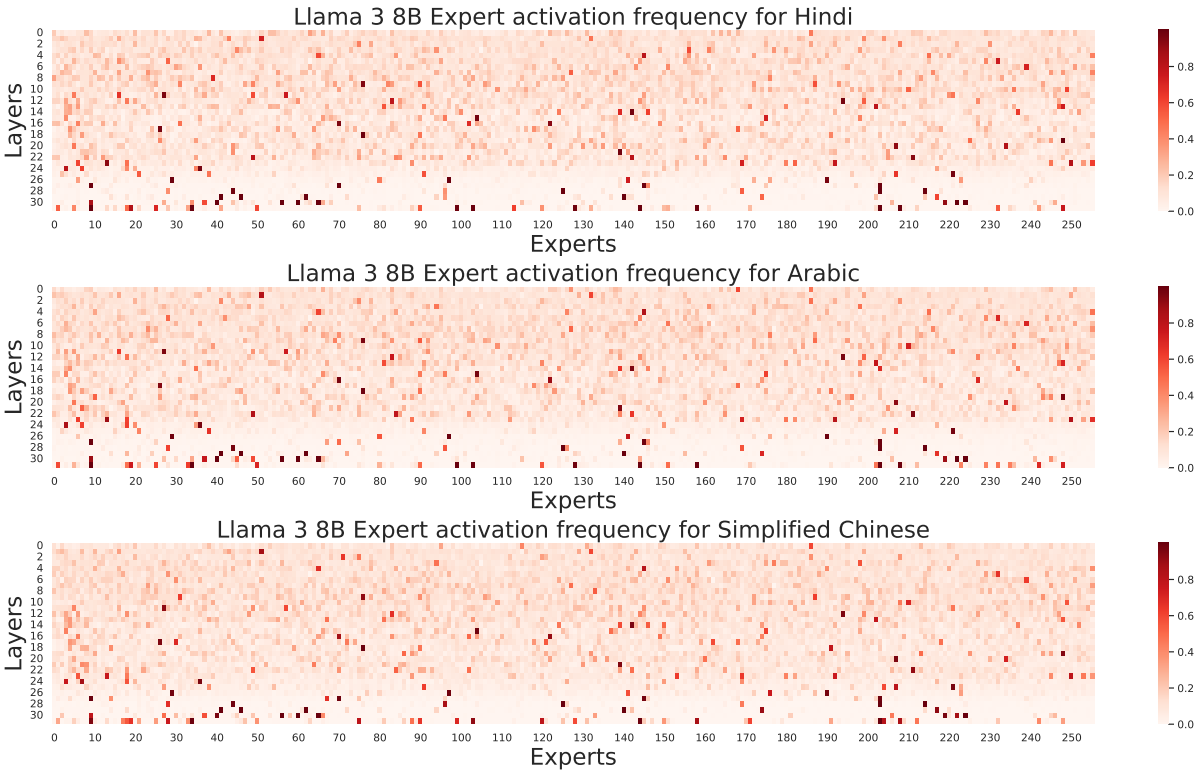
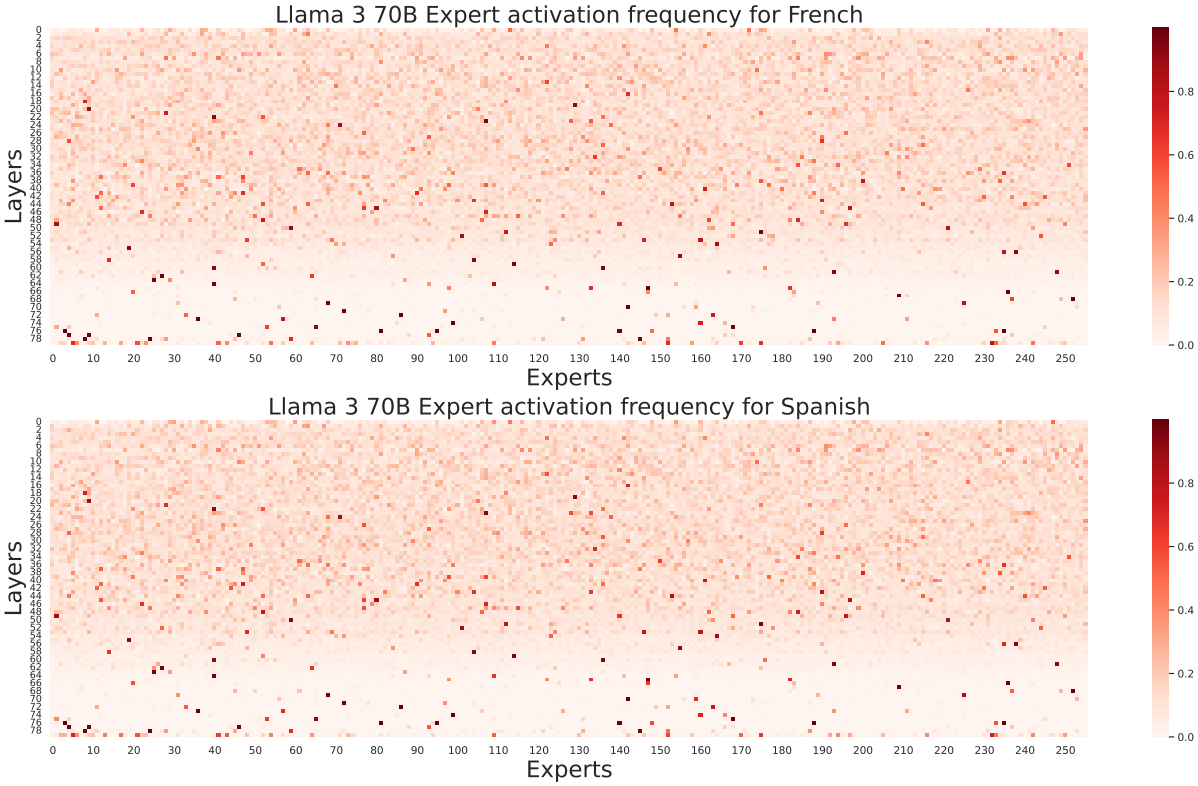
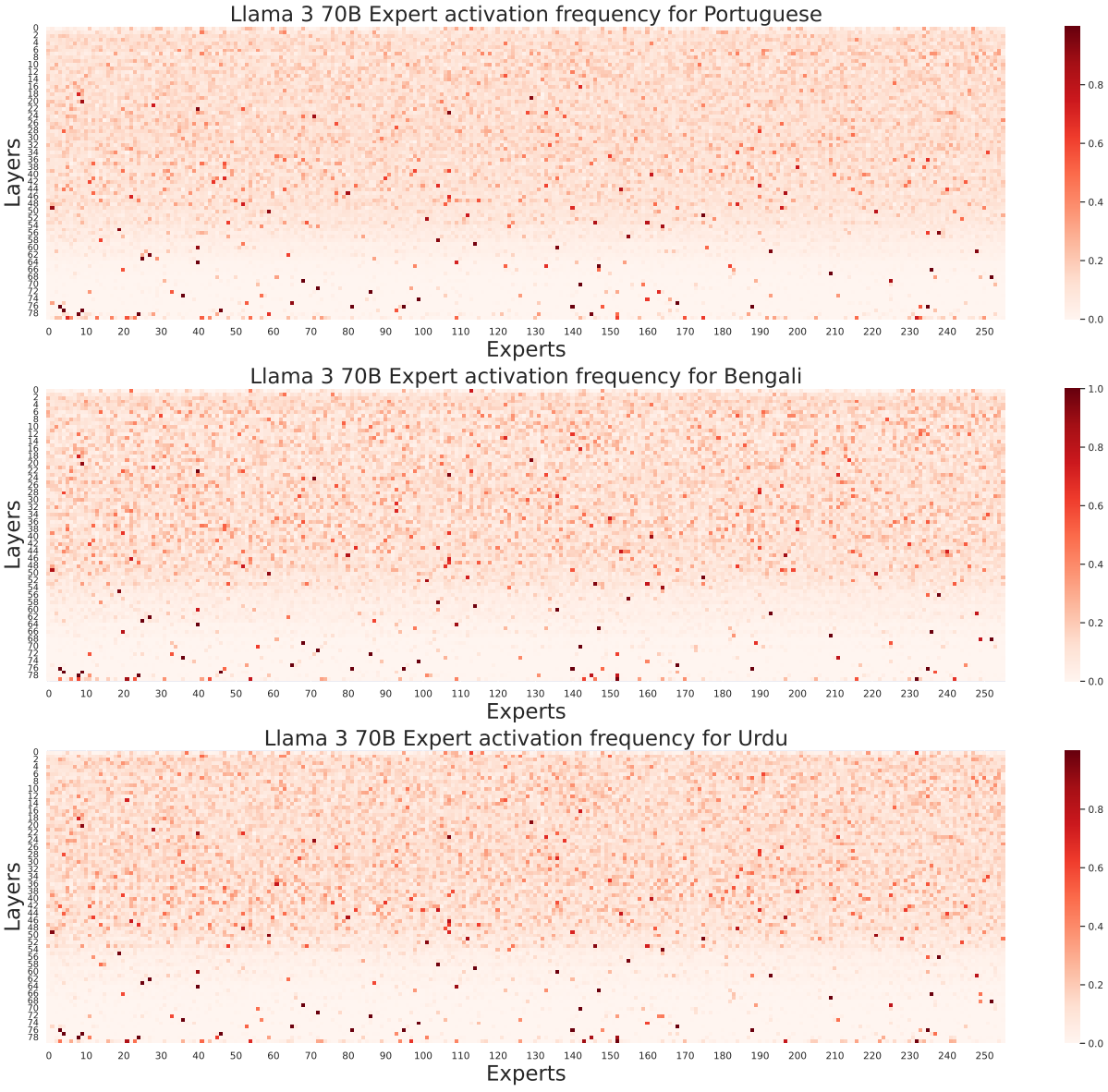


Figure 9: Heatmaps of activation patterns across languages for Llama 3 8B.




Llama 3 70B Expert activation frequency for Portuguese
Layers
Experts
Llama 3 70B Expert activation frequency for Bengali
Layers
Experts
Llama 3 70B Expert activation frequency for Urdu
Layers
Experts

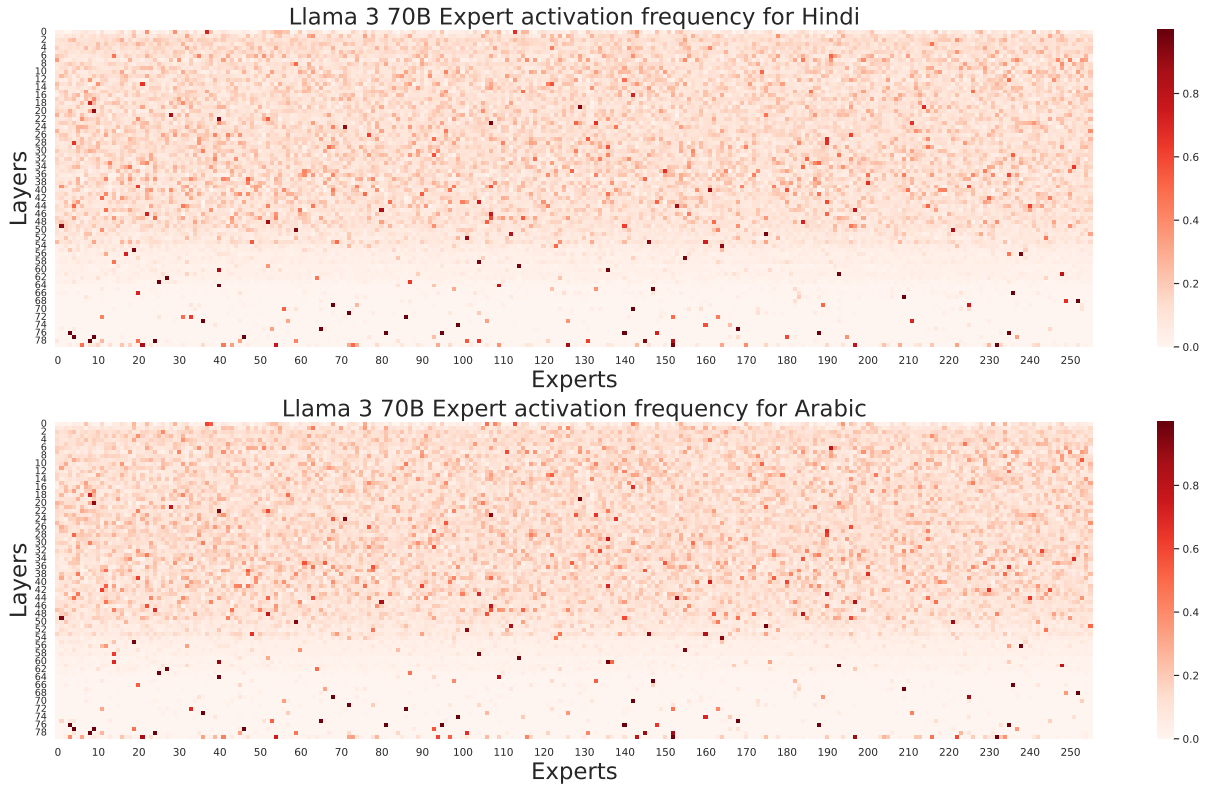


Figure 10: Heatmaps of activation patterns across languages for Llama 3 70B.

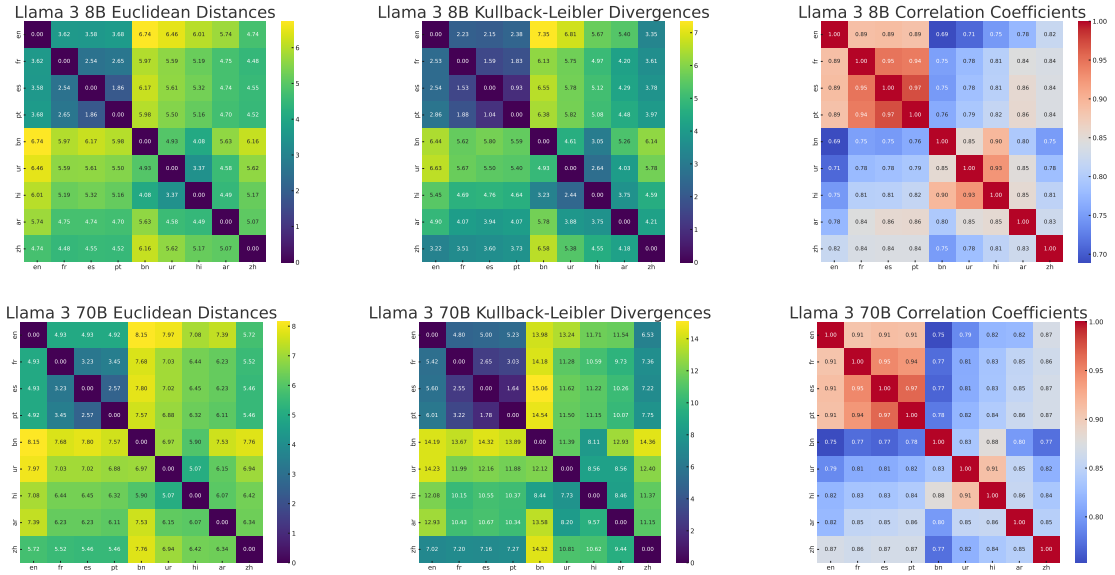


Figure 11: Heatmaps of similarity between activation pattern matrices for different languages in Llama 3 8B and Llama 3 70B. Each value in the Euclidean distance heatmaps represents the square root of the sum of the squares of the differences between corresponding elements of two matrices. Each value in the Kullback-Leibler (KL) divergence heatmaps represents the cumulative sum of the KL divergences computed row-wise between two matrices. Each value in the correlation coefficients heatmaps represents the mean of the Pearson correlation coefficients calculated row-wise between two matrices. The smaller the Euclidean distance/KL divergence, the more similar the two matrices are. The larger the correlation coefficient, the more similar the two matrices are.

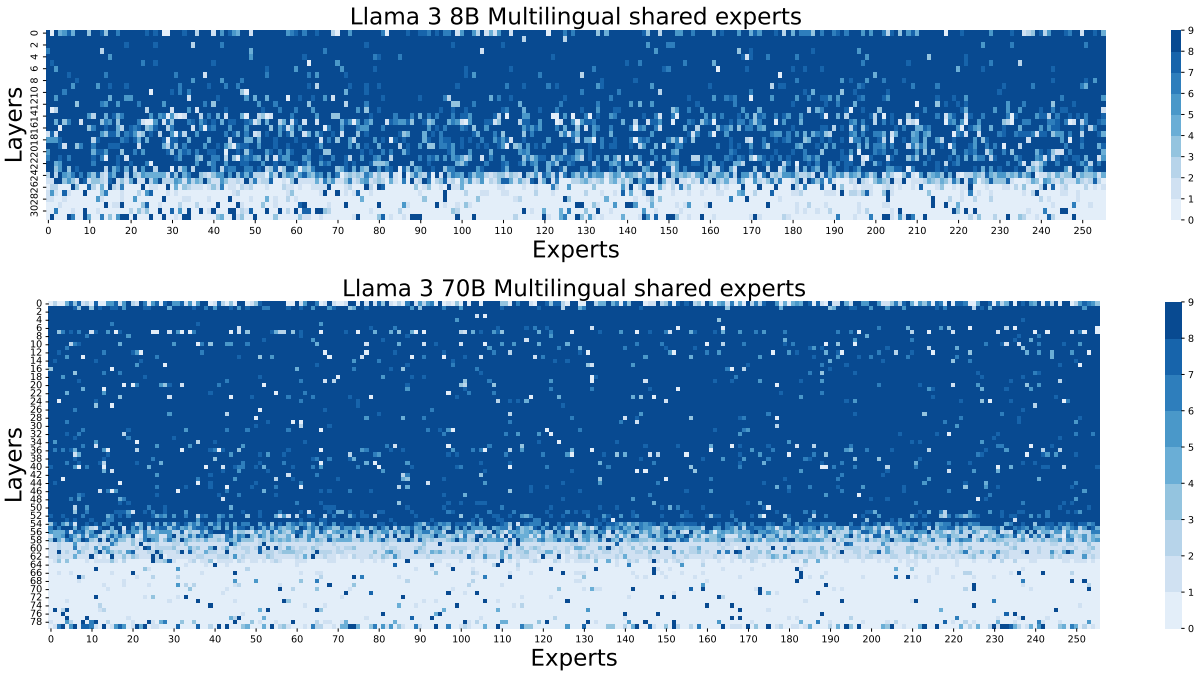
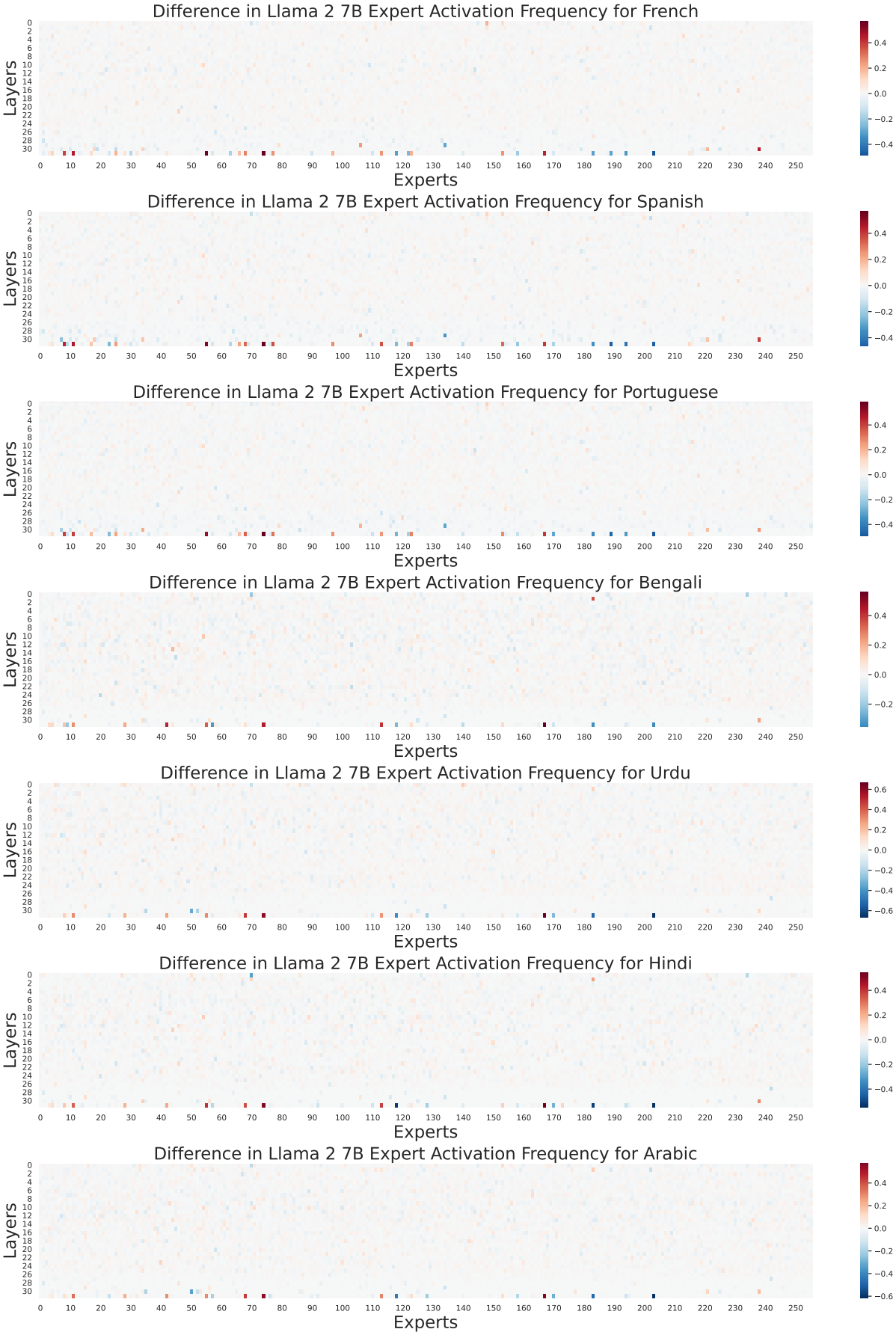


Figure 12: The heatmaps of multilingual shared experts for Llama 3 8B and Llama 3 70B. The color shade of each cell indicates how many languages the expert is a high-frequency expert in.


Difference in Llama 2 7B Expert Activation Frequency for French
Difference in Llama 2 7B Expert Activation Frequency for Spanish
Difference in Llama 2 7B Expert Activation Frequency for Portuguese
Difference in Llama 2 7B Expert Activation Frequency for Bengali
Difference in Llama 2 7B Expert Activation Frequency for Urdu
Difference in Llama 2 7B Expert Activation Frequency for Hindi
Difference in Llama 2 7B Expert Activation Frequency for Arabic
Layers
Experts

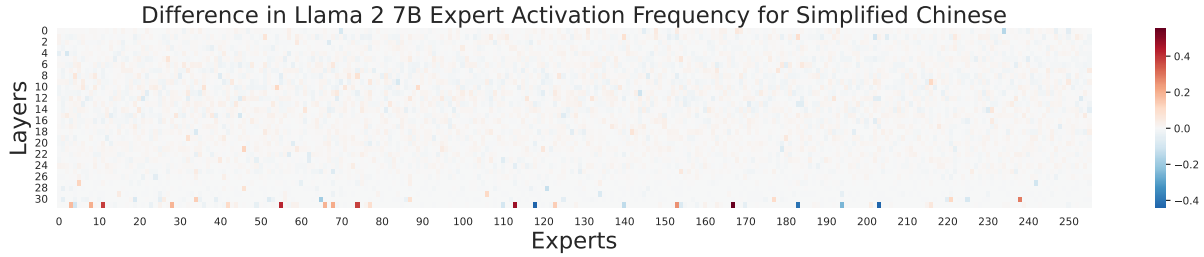
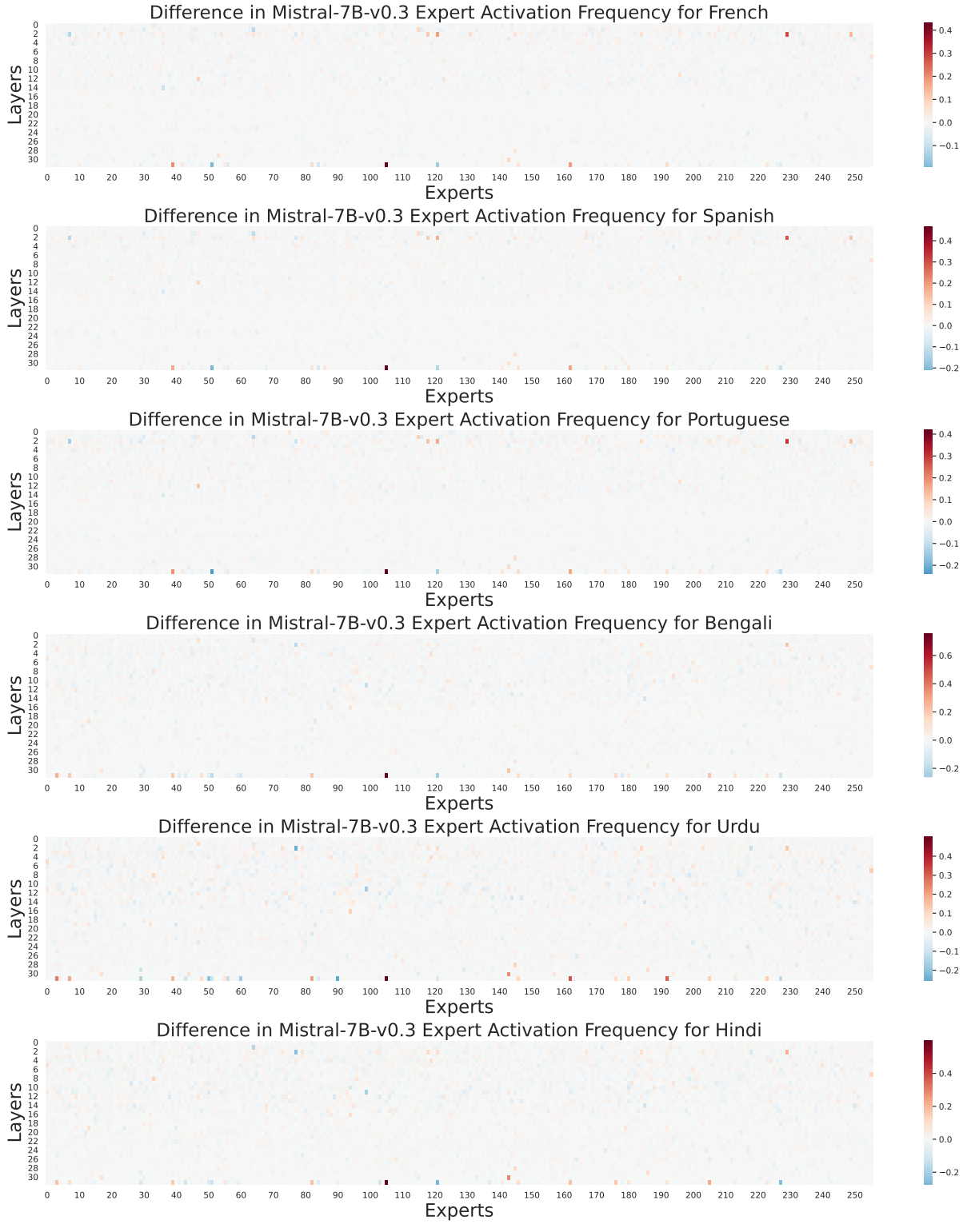


Figure 13: Changes in expert activation frequency of Llama 2 7B instruction tuning variants across different languages compared to the original pre-trained model.



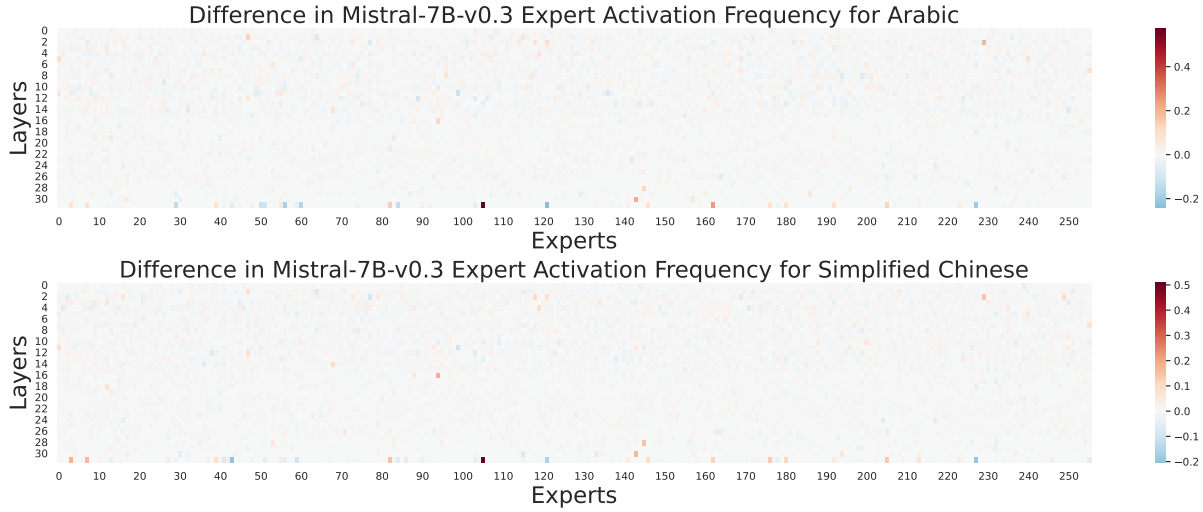
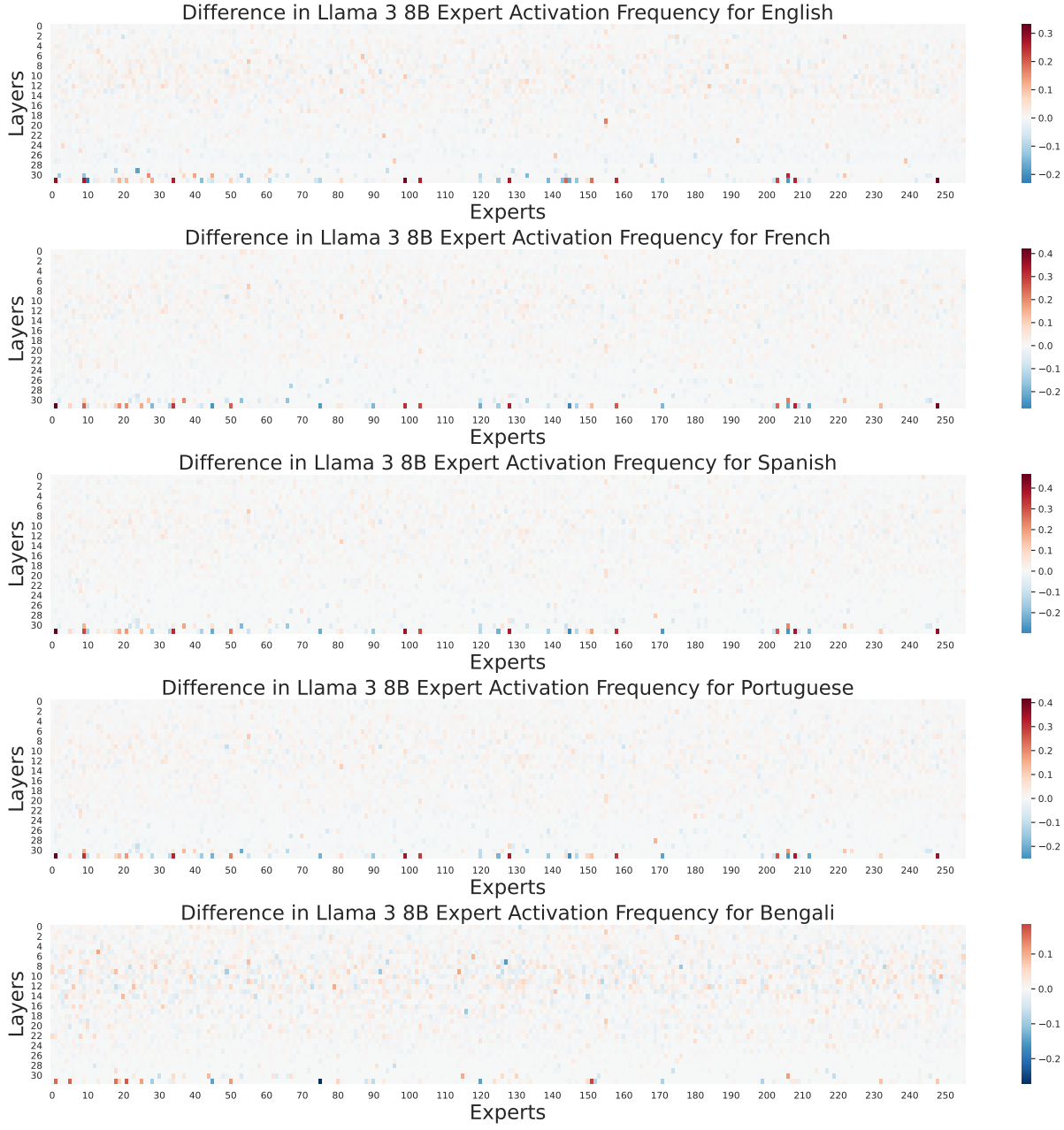


Figure 14: Changes in expert activation frequency of Mistral-7B-v0.3 instruction tuning variants across different languages compared to the original pre-trained model.



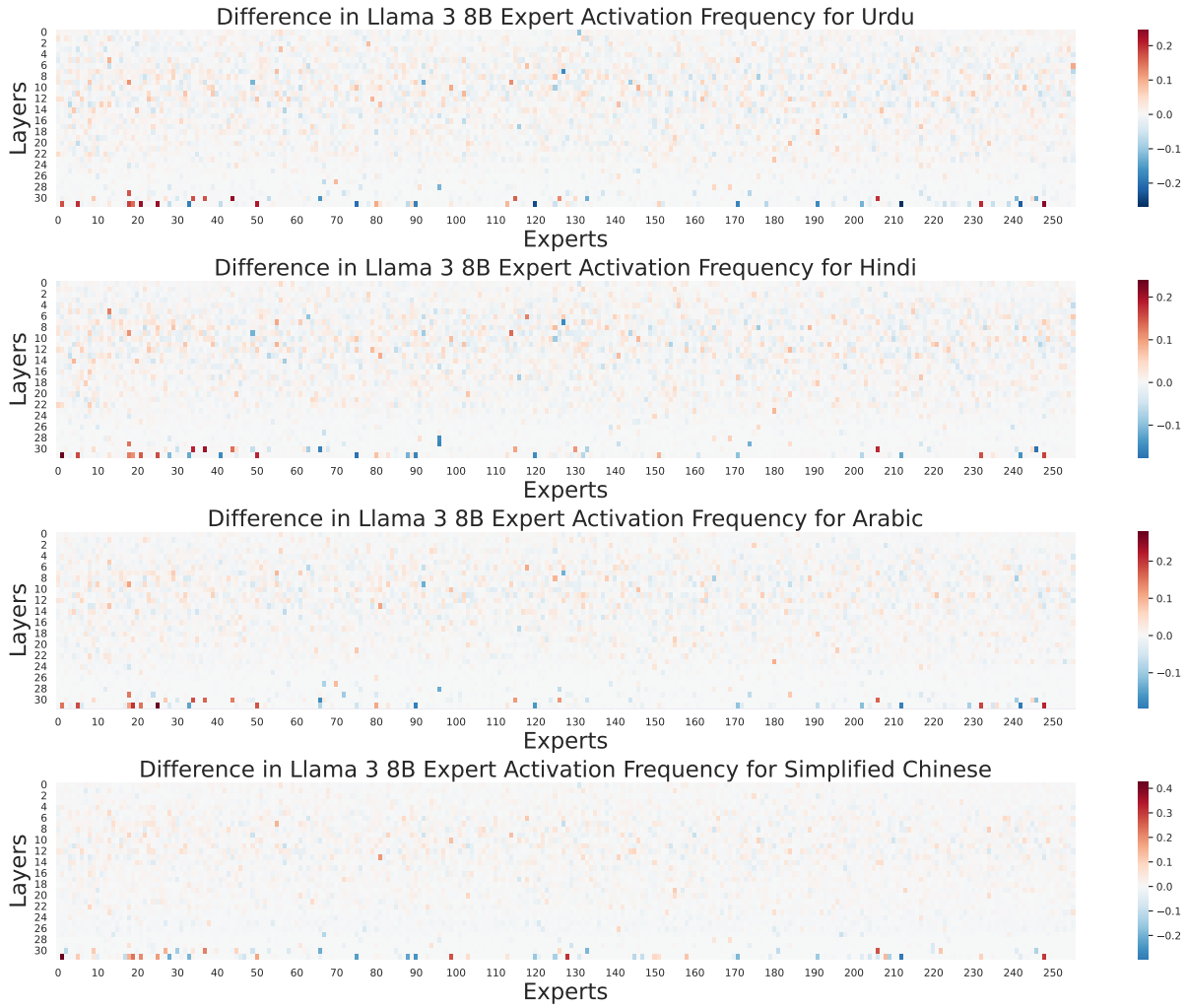
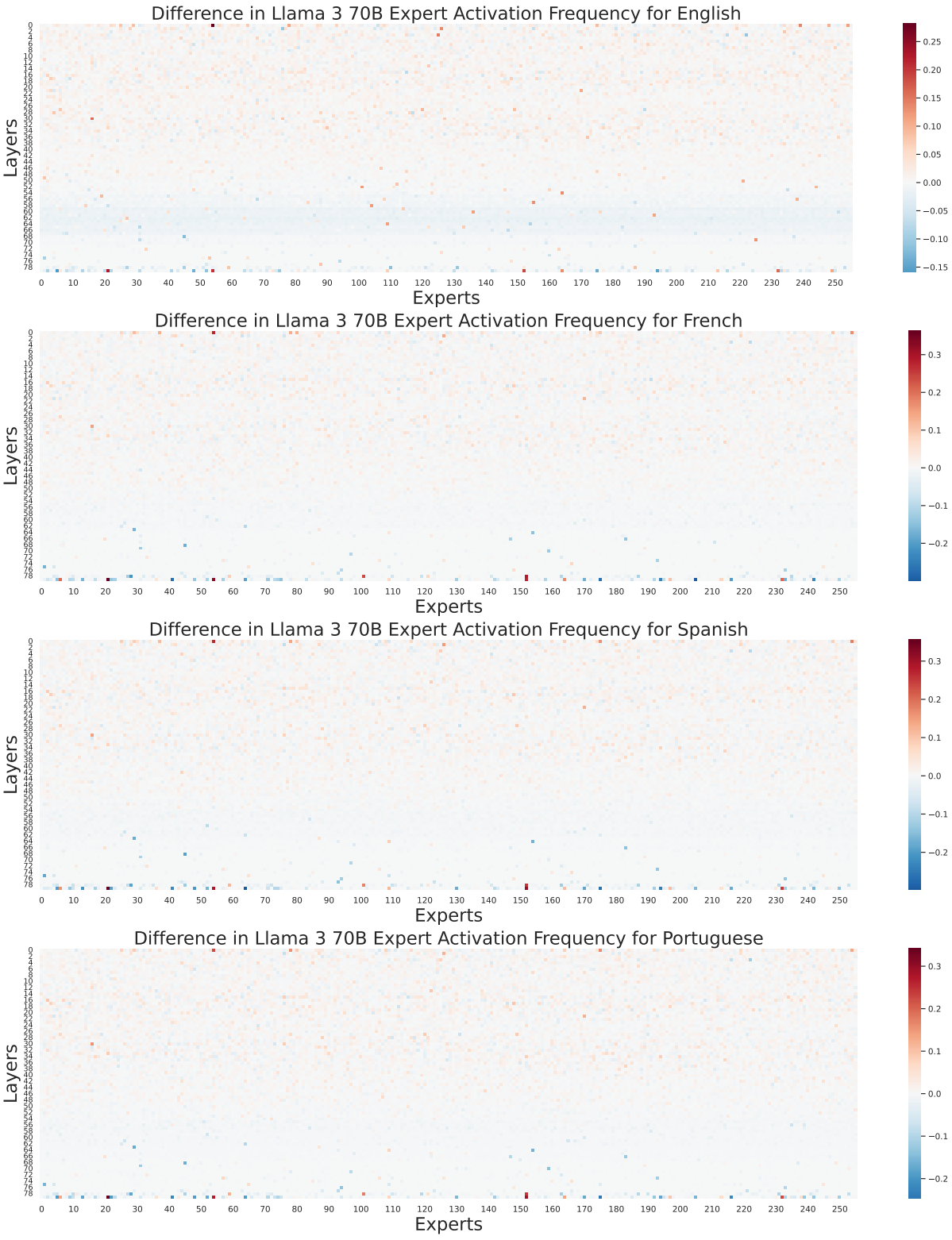
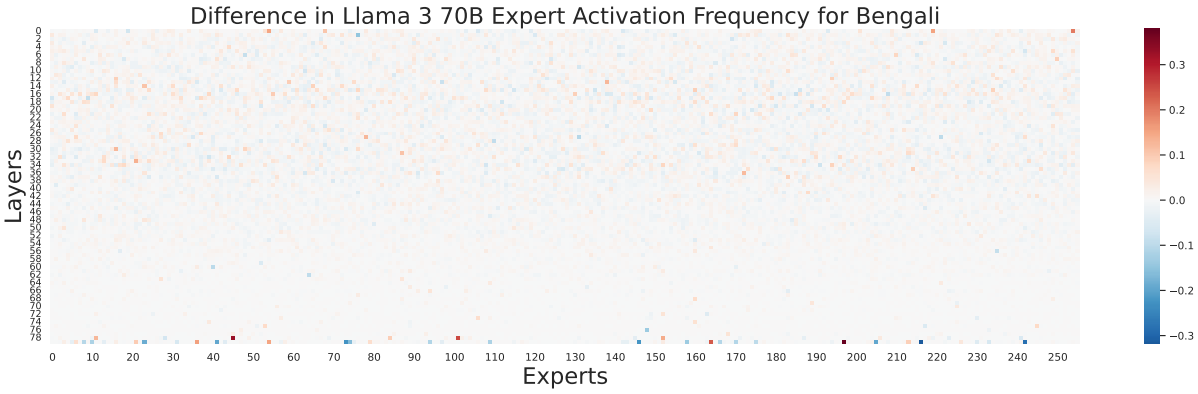
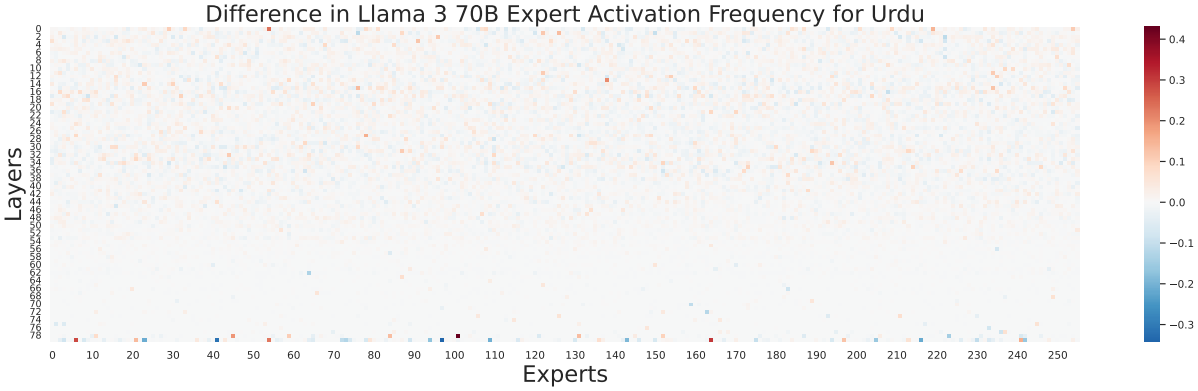
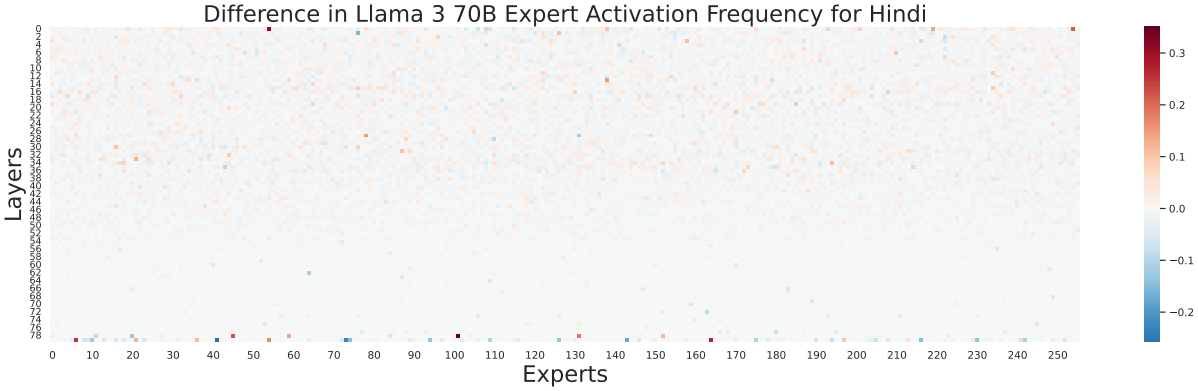


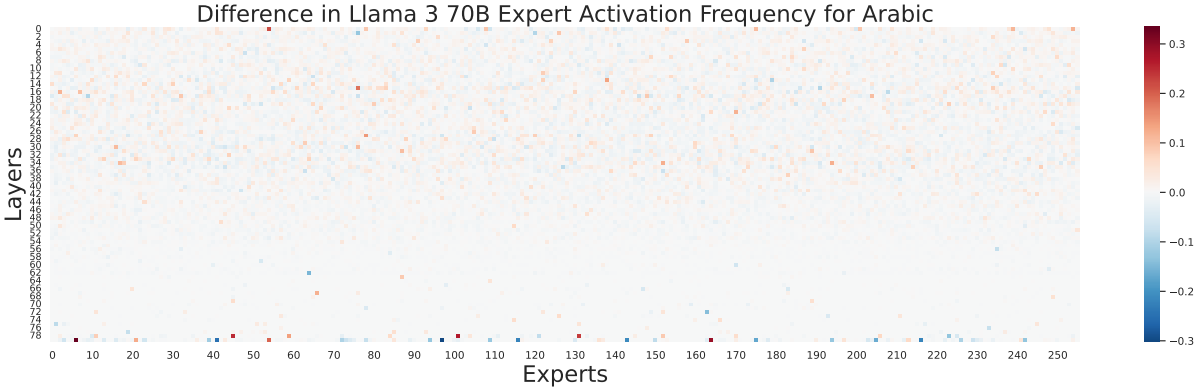
Figure 15: Changes in expert activation frequency of Llama 3 8B instruction tuning variants across different languages compared to the original pre-trained model.


Difference in Llama 3 70B Expert Activation Frequency for English
Layers
Experts
Difference in Llama 3 70B Expert Activation Frequency for French
Layers
Experts
Difference in Llama 3 70B Expert Activation Frequency for Spanish
Layers
Experts
Difference in Llama 3 70B Expert Activation Frequency for Portuguese
Layers
Experts


Difference in Llama 3 70B Expert Activation Frequency for Bengali
Layers
Experts


Difference in Llama 3 70B Expert Activation Frequency for Urdu
Layers
Experts


Difference in Llama 3 70B Expert Activation Frequency for Hindi
Layers
Experts


Difference in Llama 3 70B Expert Activation Frequency for Arabic
Layers
Experts

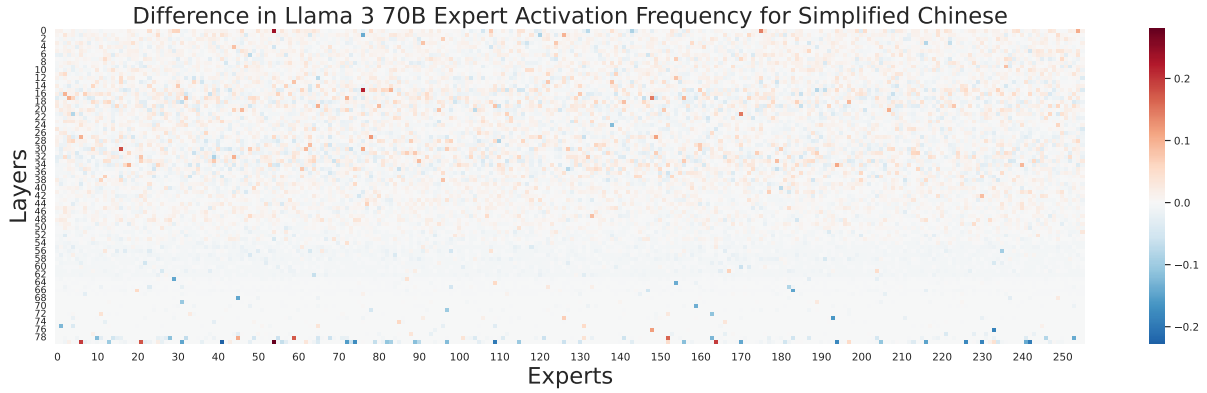


Figure 16: Changes in expert activation frequency of Llama 3 70B instruction tuning variants across different languages compared to the original pre-trained model.

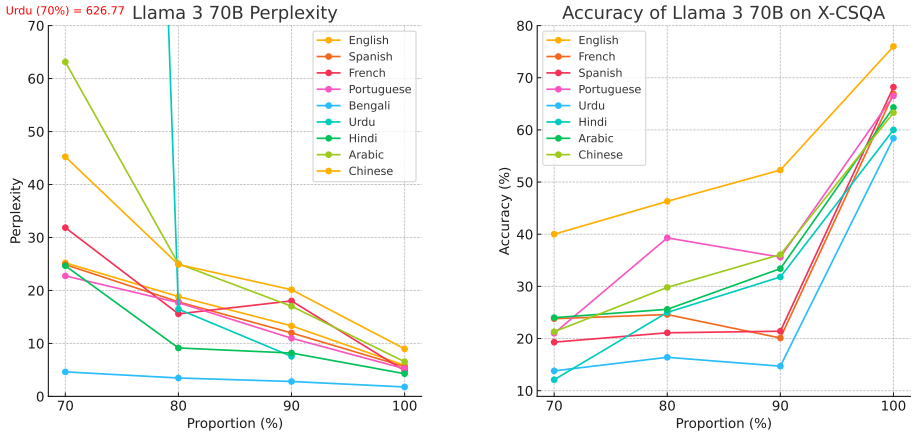


Figure 17: Results of Llama 3 70B pruning based on frequency sorting.

	Expert activation frequency $\geq 0.5\%$			Expert activation frequency $\geq 0.1\%$			
Language	Proportion	Random	Experts	Proportion	Random	Experts	Origin
English	90.1%	4698.40 ± 5822.59	8.55	93.3%	65061.69 ± 91423.24	7.05	5.86
French	87.7%	2995.53 ± 3686.66	7.34	93.2%	3318.31 ± 3260.60	5.56	4.86
Spanish	87.5%	1680.02 ± 2196.96	7.87	93.1%	60.45 ± 21.81	6.41	5.62
Portuguese	87.5%	187.94 ± 137.34	7.61	93.2%	19.27 ± 5.33	6.02	5.23
Bengali	87.9%	21.47 ± 15.56	2.14	91.8%	16.43 ± 11.99	1.93	1.78
Urdu	85.7%	295.45 ± 166.41	5.08	91.7%	18.14 ± 3.96	3.91	3.23
Hindi	86.0%	1433.74 ± 1322.22	6.13	90.5%	94.81 ± 96.82	5.12	4.28
Arabic	84.8%	364.06 ± 106.11	13.28	91.7%	57.79 ± 32.34	8.97	6.54
Chinese	87.0%	1437.81 ± 1704.64	15.64	92.8%	36.01 ± 7.77	11.28	8.98
Average	87.13%	1457.16	8.18	92.37%	7631.43	6.25	5.15

Table 8: The perplexity results of Llama 3 70B. The smaller the value, the better the model performance.

	Expert activation frequency $\geq 5\%$			Expert activation frequency $\geq 1\%$			
Language	Proportion	Random	Experts	Proportion	Random	Experts	Origin
English	77.1%	5641.07 ± 6999.41	20.54	89.6%	19.14 ± 1.41	11.78	9.43
French	76.3%	866.76 ± 100.47	34.77	90.8%	21.78 ± 0.80	14.50	10.09
Spanish	76.3%	89.55 ± 5.88	43.13	90.7%	23.45 ± 0.46	15.78	11.70
Portuguese	75.9%	177709.36 ± 250709.26	39.95	91.0%	23.26 ± 1.72	15.97	10.78
Bengali	67.5%	181486.00 ± 254425.86	37.24	89.5%	21.63 ± 3.99	10.76	5.22
Urdu	68.7%	145160.34 ± 165880.84	27.36	89.7%	16.27 ± 0.67	8.85	7.24
Hindi	70.9%	68728.70 ± 69301.79	11.02	89.8%	10.23 ± 0.21	6.05	4.86
Arabic	72.1%	426638.39 ± 602646.86	20.44	90.0%	14.08 ± 2.30	10.63	7.11
Chinese	76.4%	67081.13 ± 94749.98	34.38	91.7%	21.78 ± 1.68	11.21	9.28
Average	73.5%	119266.81	29.87	90.3%	19.07	11.73	8.41

Table 9: The perplexity results of Llama 2-Chat 7B. The smaller the value, the better the model performance.

	Expert activation frequency $\geq 5\%$			Expert activation frequency $\geq 1\%$			
Language	Proportion	Random	Experts	Proportion	Random	Experts	Origin
English	77.1%	41.7 ± 4.5	50.0	89.6%	49.3 ± 2.3	**57.2**	56.8
French	76.3%	15.3 ± 8.9	33.4	90.8%	41.9 ± 0.6	45.0	45.3
Spanish	76.3%	19.9 ± 14.5	38.7	90.7%	41.5 ± 1.0	42.4	44.5
Portuguese	75.9%	19.0 ± 13.5	31.9	91.0%	35.9 ± 1.6	38.2	39.8
Chinese	76.4%	28.5 ± 5.5	34.2	91.7%	37.6 ± 0.8	38.4	38.4
Average	76.4%	24.9	37.6	90.8%	41.2	44.2	45.0

Table 10: Accuracy (%) of Llama 2-Chat 7B on the X-CSQA dataset.