Unveiling Language-Specific Features in Large Language Models via Sparse Autoencoders

Boyi Deng[1], **Yu Wan**[*], **Yidan Zhang**[2], **Baosong Yang**, **Fuli Feng**[1*]
[1]University of Science and Technology of China,
[2]Sichuan University
dengboyi@mail.ustc.edu.cn

Abstract

The mechanisms behind multilingual capabilities in Large Language Models (LLMs) have been examined using neuron-based or internal-activation-based methods. However, these methods often face challenges such as superposition and layer-wise activation variance, which limit their reliability. Sparse Autoencoders (SAEs) offer a more nuanced analysis by decomposing the activations of LLMs into a sparse linear combination of SAE features. We introduce a novel metric to assess the monolinguality of features obtained from SAEs, discovering that some features are strongly related to specific languages. Additionally, we show that ablating these SAE features only significantly reduces abilities in one language of LLMs, leaving others almost unaffected. Interestingly, we find some languages have multiple synergistic SAE features, and ablating them together yields greater improvement than ablating individually. Moreover, we leverage these SAE-derived language-specific features to enhance steering vectors, achieving control over the language generated by LLMs. The code is publicly available at https://github.com/Aatrox103/multilingual-llm-features.

1 Introduction

Large Language Models (LLMs) (OpenAI et al., 2024; Grattafiori et al., 2024; Qwen et al., 2025) exhibit impressive abilities in various domains such as text generation (OpenAI et al., 2024; Grattafiori et al., 2024; Xu et al., 2025), instruction following (Zhang et al., 2024a; Lou et al., 2024), and reasoning (Huang and Chang, 2023). Recently, considerable efforts have been made to enhance the multilingual capabilities of LLMs to meet the growing demand for their deployment in multilingual environments (Qin et al., 2024; Huang et al., 2025). For instance, Gemini 1.5 incorporates a variety of multilingual data in its training process and emphasizes its multilingual capabilities (Team et al., 2024a). Yang et al. (2024) claim that Qwen2 supports over 30 languages and achieves great performance on multilingual benchmarks. Moreover, multilingual training data comprises approximately 3% of the training data for Llama 3, and there are also high-quality multilingual instruction-tuning data for 8 languages (Grattafiori et al., 2024). As the significance of multilingual capabilities in LLMs continues to grow, it is crucial to delve into the mechanisms of these capabilities to enhance them further.

Works focusing on the mechanisms of multilingual capabilities in LLMs can be broadly divided into neuron-based and internal-activation-based methods. Neuron-based methods aim to identify language-specific neurons and analyze their impact on the corresponding language (Zhang et al., 2024b; Zhao et al., 2024; Tang et al., 2024; Kojima et al., 2024). And activation-based method attempts to obtain token distributions at intermediate layers using the unembedding matrix in the final layer (Zhong et al., 2024; Wendler et al., 2024). However, neuron-based methods are sometimes unreliable, due to “superposition” (Elhage et al., 2022), which suggests that neural networks often consolidate multiple unrelated concepts into a single neuron. Additionally, activation-based method often has significant errors except in the last few layers, due to the varying distribution of activations across different layers. As such, it is important to use a more reliable and interpretable method to analyze multilingual capabilities in LLMs.

To achieve this, we use Sparse Autoencoders (SAEs) (Bricken et al., 2023; Cunningham et al., 2023), which are designed to decompose language model activations in each layer into a sparse linear combination of SAE features. The advantages of SAEs in analyzing multilingual capabilities in LLMs are threefold. First, SAEs can be applied to individual tokens, providing a more monosemous

[*]Corresponding authors

analysis compared to neuron-based methods. Second, SAEs are trained on each layer separately, making them more reliable when analyzing activations from different layers than current activation-based methods. Third, multilingual data is naturally parallel, meaning that ideally, the main difference between multilingual data is the language, so it is easy to identify monolingual features with SAEs.

Given the advantages of SAEs, we use them to analyze multilingual capabilities in LLMs. Concretely, we start with a preliminary experiment in which we find high activation of some features in a certain language. Inspired by this, we propose a metric to measure the monolinguality of a feature based on the activation difference across different languages. The results show that some features possess strong monolingual characteristics. Moreover, we believe that these language-specific features are not only related to language-specific tokens, so we experiment on a "code-switching" (Kuwanto et al., 2024; Winata et al., 2023) dataset and find that language-specific features are also closely associated with the language-specific linguistic context. Furthermore, we use *directional ablation* (Arditi et al., 2024; Ferrando et al., 2024) to "zero out" language-specific features during the forward pass of LLMs, resulting in a loss of capabilities in only certain language. Interestingly, we observe that some languages may exhibit more than one specific feature. And these features have a synergistic relationship, meaning ablating these features together results in a significant improvement compared to ablating them individually.

The language-specific features we find are of great monolinguality, so we further leverage them to improve *steering vectors* (Turner et al., 2024). Concretely, we use language-specific features as gating signals to control steering vectors and achieve better control over the language generated by LLMs, which validates the practical potential of these language-specific features.

In summary, our main contributions are:

- We use SAE, a more human-interpretable method, to analyze multilingual capabilities of LLMs, and propose a metric to measure the monolinguality of SAE features.
- We find some SAE features that are not only related to language-specific tokens but also related to language-specific linguistic context.
- We find that ablating language-specific features only significantly decreases the language-specific capabilities of LLMs.
- We use language-specific features as gating signals to improve *steering vectors*, and achieve better control over the language generated by LLMs.

2 Preliminary

SAEs. SAEs are a specialized form of autoencoders (Hinton and Zemel, 1993) designed to decompose language model activations into a sparse linear combination of learned feature directions. Given a language model activation $\mathbf{x} \in \mathbb{R}^N$ in certain layer[1], the SAE computes a feature activation $\mathbf{f} \in \mathbb{R}^M$, where $M \gg N$, and reconstructs the input as $\hat{\mathbf{x}}$. The typical reconstruction process is described by the equations:

$$\mathbf{f}(\mathbf{x}) := \mathrm{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}), \quad (1)$$

$$\hat{\mathbf{x}}(\mathbf{f}) := \mathbf{W}_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{dec}}. \quad (2)$$

To ensure that $\mathbf{f}$ remains sparse, Bricken et al. (2023); Cunningham et al. (2023) incorporate an L1 penalty on $\mathbf{f}$ into the training loss function. Another approach by Gao et al. (2024) employs Top-K SAEs, which enforce sparsity by selecting only the K most active dimensions of $\mathbf{f}$, setting all the others to zero. Following the notation of Rajamanoharan et al. (2024), we denote the columns of $\mathbf{W}_{\text{dec}}$ as $\mathbf{d}_i$ for $i = 1, \ldots, M$. These columns represent the feature directions into which the SAE decomposes the vector $\mathbf{x}$. For simplicity, we will refer to each column as a "feature" throughout this paper.

Datasets. Flores-200 (Costa-jussà et al., 2022; Goyal et al., 2022) is a parallel corpus that contains translations of English sentences into 200 different languages. Due to the semantic similarity of the translated sentences, this dataset is particularly useful for comparing linguistic features across languages. We extract a subset called Flores-10, which includes 10 languages[2].

Models. To ensure the robustness of our findings, we include a diverse set of LLMs and their corresponding SAEs. We use SAEs from Gemma Scope (Lieberum et al., 2024) for Gemma 2 2B and Gemma 2 9B (Team et al., 2024b), and SAEs from Llama Scope (He et al., 2024) for Llama-3.1-8B.

[1] We use the residual stream at each layer as $\mathbf{x}$ because it is more interpretable (Ferrando et al., 2024; Chanin et al., 2024).

[2] English (en), Spanish (es), French (fr), Japanese (ja), Korean (ko), Portuguese (pt), Thai (th), Vietnamese (vi), Chinese (zh), and Arabic (ar).

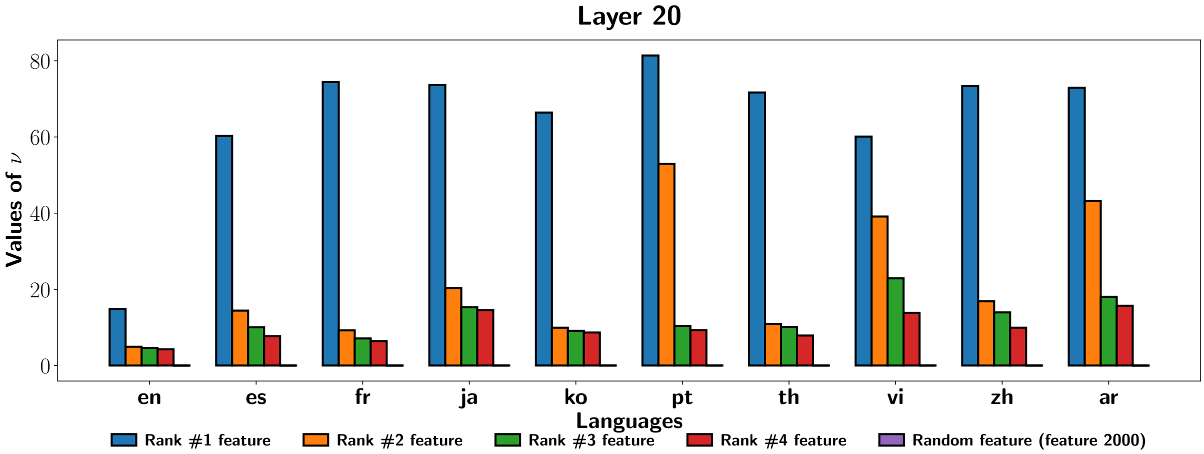


Figure 1: The values of ν, as referenced in Eq. 4, where a larger ν indicates stronger monolingualism, are reported for the top-4 features and a random feature across various languages in layer 20 of Gemma 2 2B. The values of ν for the top-4 features are greater than those of a random feature. In most languages, the top-1 feature possesses a significantly larger ν. Additional results for other layers and LLMs are in Appendix C, exhibiting similar patterns. The value of the random feature (feature 2000) is too small to be visible.

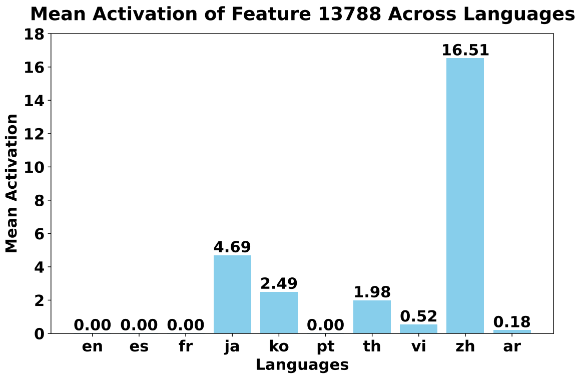


Figure 2: The mean activation of feature 13788 across different languages in layer 10 of Gemma 2 2B. The high mean activation in Chinese suggests that feature 13788 might be related to Chinese.

3 Language-Specific Features

3.1 Finding Language-Specific Features

To find language-specific features, we conduct a preliminary experiment by prompting Flores-10 into the LLMs and analyzing the residual stream using SAEs. We find that the mean activation of some features is particularly high for a certain language, while remaining very low for other languages, as illustrated by the example in Figure 2. Inspired by this, we propose a metric to measure the monolinguality of a feature. Specifically, given a set $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, which contains the residual stream set for a certain layer with K different languages, we calculate the mean activation difference of feature s for a specific language L compared to the other languages as follows:

$$\begin{aligned}\mu_s^L &= \frac{1}{|\mathcal{D}_L|}\sum_{\mathbf{x}\in\mathcal{D}_L}\mathbf{f}_s(\mathbf{x}),\\ \gamma_s^L &= \frac{1}{|\mathcal{D}\setminus\{\mathcal{D}_L\}|}\sum_{\mathcal{D}_I\in\mathcal{D}\setminus\{\mathcal{D}_L\}}\frac{1}{|\mathcal{D}_I|}\sum_{\mathbf{x}\in\mathcal{D}_I}\mathbf{f}_s(\mathbf{x}),\\ \nu_s^L &= \mu_s^L - \gamma_s^L,\end{aligned}\tag{3}$$

where $\mathbf{f}_s(\mathbf{x})$ is the activation of feature s. We calculate ν for all languages and features and rank them from high to low for each language. The top-ranked features are considered language-specific features.

3.2 Monolinguality Analysis

We use the first 100 data points in Flores-10 to calculate ν for each language. The results are shown in Figure 1. From this figure, we make the following observations. (1) The mean activation of the top-4 features is significantly higher than that of a random feature, which remains close to zero. (2) For most languages, the mean activation of the top features decreases rapidly among the first few, and the mean activation of the rank #1 feature is considerably higher than the others. (3) In some languages, the rank #2 feature also shows a substantially large mean activation compared to other features. These results suggest that top-ranked features possess strong monolingual characteristics, and in most scenarios, the top-1 feature suffices in capturing these characteristics.[3]

[3] English is the primary language for most LLMs, and it often exhibits different characteristics compared to other

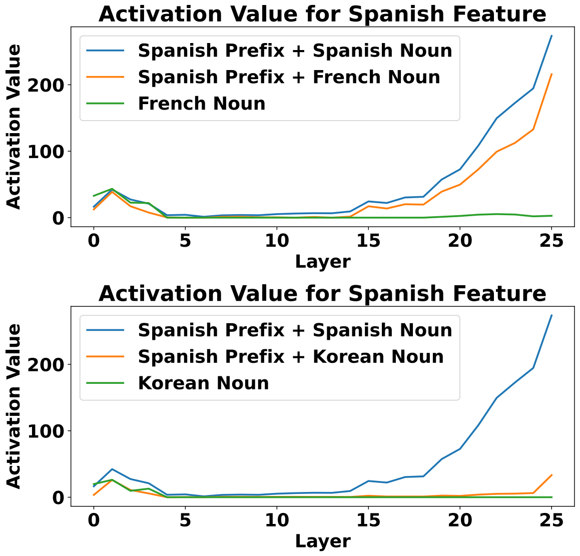


Figure 3: The mean activation values for the Spanish feature with various noun and prefix combinations. Adding a Spanish prefix enhances the Spanish feature activation for non-Spanish nouns, enabling the LLM to process them as if they were "Spanish tokens."

4 Language-Specific Features Extend Beyond Language-Specific Tokens

In earlier sections, we only evaluate language-specific features on monolingual texts. This raises a natural question: are these language-specific features solely related to language-specific tokens? To explore this, we focus on a phenomenon called "code-switching."[4] Our findings indicate that language-specific features are also related to language-specific linguistic context.

4.1 Experimental Settings

Code-Switching Dataset. We use GPT-4o to generate sentences in various languages, each ending with a noun. We then replace the noun with its equivalent in other languages. For each language, we generate 5 simple sentences, and each sentence has 8 variants where the noun is substituted with its equivalent in different languages. Example data are shown in Figure 9. We only report results of Gemma 2 2B for Spanish prefix, additional results with the same patterns are in Appendix E.

Metric To analyze the impact of different language prefixes on ending nouns, we calculate the mean activation of language-specific features for the ending nouns both with and without a prefix.

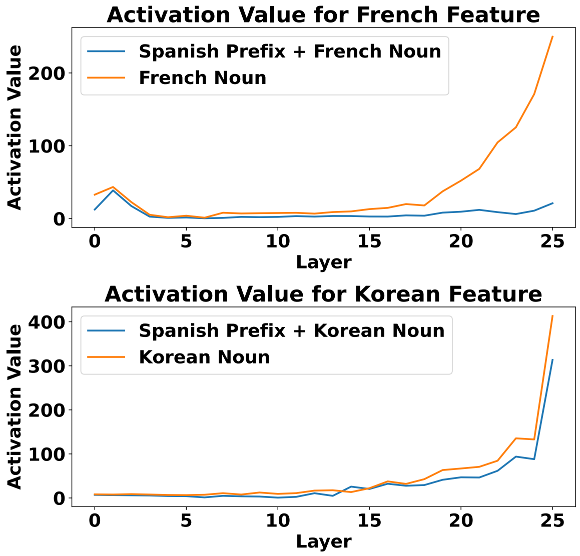


Figure 4: The mean activation values for the French and Korean features with various noun and prefix combinations. Introducing a different language prefix decreases the original language feature activation of nouns.

4.2 Results

Spanish Prefix Enhances Spanish Features in Non-Spanish Nouns. We analyze the mean activation values of the Spanish features for Spanish, French, and Korean nouns, comparing scenarios with and without Spanish prefixes, as illustrated in Figure 3. Our observations are as follows: (1) Introduction of a Spanish prefix to a French or Korean noun results in higher Spanish feature activation values compared to when the French or Korean nouns stand alone. However, the value is still lower than that of the combination of Spanish prefixes and Spanish nouns. (2) The activation value for Spanish features of stand-alone French and Korean nouns remains relatively low across all layers. (3) Both French and Korean nouns with a Spanish prefix show greater increases in Spanish feature activations at deeper layers than at shallower ones. (4) Adding a Spanish prefix results in a larger increase in the Spanish feature for French nouns compared to Korean nouns. These findings suggest that adding a Spanish prefix enhances the Spanish feature activation for non-Spanish nouns, enabling the LLM to process them as if they were "Spanish tokens." Consequently, this allows the LLM to use these non-Spanish tokens within a consistent language context.

Spanish Prefix Decreases Non-Spanish Features in Non-Spanish Nouns. We also analyze the mean activation values of the French and Korean features for corresponding nouns, comparing sce-

languages (Qin et al., 2024), so we only focus on non-English results in the subsequent sections.

[4] Code-switching refers to the practice of alternating between two or more languages within a single text (Kuwanto et al., 2024; Winata et al., 2023).

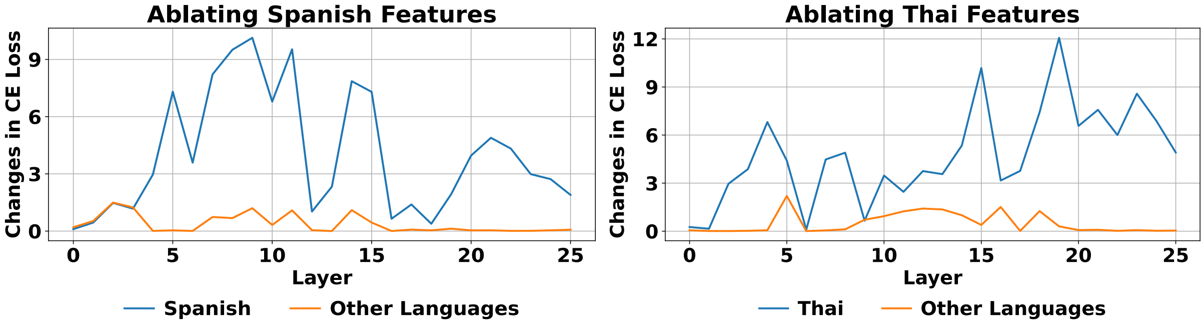


Figure 5: The changes in CE loss on texts in the target language and texts in other languages after ablating language-specific features. Ablating language-specific features has a much larger impact on the CE loss of texts in the target language compared to texts in other languages. We provide results for Gemma 2 2B here, additional results can be found in Appendix D.

narios with and without Spanish prefixes, as presented in the provided Figure 4. Our observations are as follows: (1) For French and Korean nouns, the original language feature activation is significantly higher when the nouns are standalone than when preceded by a Spanish prefix. (2) Both French and Korean nouns show greater decreases in their original language feature activations at deeper layers than at shallower ones. (3) Adding a Spanish prefix results in a larger decrease in the corresponding feature for French nouns compared to Korean nouns. These findings reveal that introducing a different language prefix decreases the original language feature activation of nouns, making them less like nouns from their original language.

Language-Specific Features Extend Beyond Language-Specific Tokens. The results in Figures 3 and 4 suggest that language-specific features are not solely tied to specific language tokens but are also closely associated with the language-specific linguistic context. This suggests that the linguistic characteristics recognized by the model extend beyond individual words to encompass the contextual environment in which these words appear. Notably, the influence of a Spanish prefix is more pronounced on French nouns than on Korean nouns, potentially due to the linguistic similarities between Spanish and French. This highlights the model's ability to dynamically adapt its feature activations based on the surrounding linguistic context, effectively reinterpreting non-Spanish tokens within a Spanish framework while diminishing their original language attributes.

5 Ablating Language-Specific Features Leads to Language-Specific Changes

In the previous section, we identified language-specific features that are closely related to monolingual texts. In this section, we examine how these language-specific features affect the language-specific capabilities of LLMs. Specifically, inspired by Arditi et al. (2024); Ferrando et al. (2024), we use *directional ablation* to "zero out" language-specific features and observe the changes in the cross-entropy (CE) loss of texts in different languages within LLMs.

5.1 Model Interventions

Directional Ablation. To analyze the impact of a feature $\mathbf{d} \in \mathbb{R}^N$ on the inference process of LLMs, Arditi et al. (2024); Ferrando et al. (2024) introduce *directional ablation* to "zero out" a feature in the residual stream activation $\mathbf{x} \in \mathbb{R}^N$. This is done by subtracting the projection of $\mathbf{x}$ onto $\mathbf{d}$ from $\mathbf{x}$:

$$\mathbf{x}' \leftarrow \mathbf{x} - \hat{\mathbf{d}}\hat{\mathbf{d}}^{\intercal}\mathbf{x}, \tag{4}$$

where $\hat{\mathbf{d}}$ is the unit vector of $\mathbf{d}$. After obtaining the ablated residual stream, replace $\mathbf{x}$ with $\mathbf{x}'$ and continue the forward pass of the LLMs.

5.2 Ablation of Language-Specific Features

For each target language, layer, and LLM, we intervene in the inference process of LLMs using Eq. 4 on the top-2 language-specific features of the target language. We then measure the changes in CE loss for both texts in the target language and texts in other languages after ablating language-specific features. The results are shown in Figure 5. We observe that: (1) Ablating language-specific features has a much larger impact on the CE loss of texts in the target language compared to texts in other languages. (2) For different layers, the changes in CE loss of target language texts vary significantly. These findings suggest that language-specific features play a crucial role in controlling

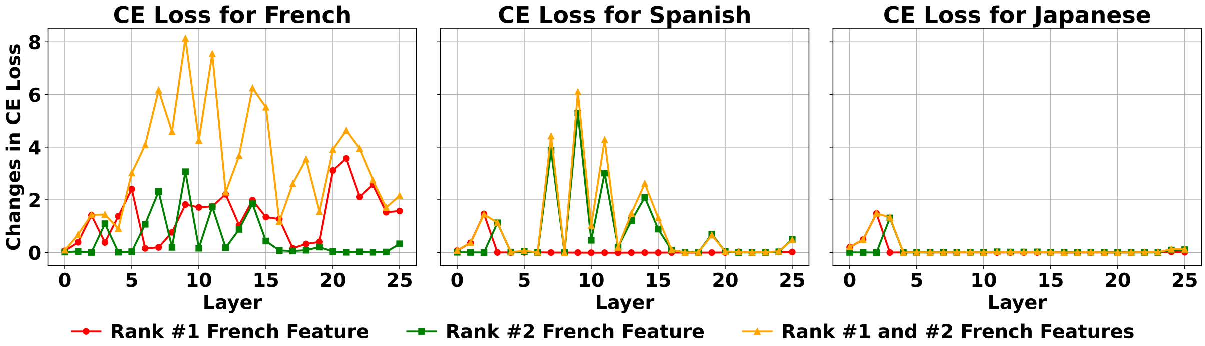


Figure 6: The change in CE loss for three languages after ablating French Features. Simultaneously ablating multiple French features exhibits a synergistic effect in French, while showing no synergistic effect on other languages. We provide results for Gemma 2 2B here, and additional results can be found in Appendix F.

the generation process for the target language. Ablating these features from the generation process of LLMs can lead to a loss of only specific language capabilities.

5.3 Synergistic Language Features

We compare the CE loss for French, Spanish, and Japanese when using different numbers of French features for directional ablation. The results are shown in Figure 6. From this, we make the following observations: (1) In some layers, simultaneously ablating the top 2 French features for French significantly impacts the CE loss more than ablating these features individually. (2) In all layers, simultaneously ablating the top 2 French features for Spanish and Japanese results in a CE loss impact approximately equal to the sum of the effects when these features are ablated individually. (3) The changes in CE loss for French are larger than those for Spanish and Japanese. The changes for Spanish are large in some layers, while for Japanese, they are nearly zero across all layers.

Based on these observations, we can conclude that for any target language, there exists a synergistic relationship among its features. Ablating multiple features simultaneously impacts significantly more than the sum of the effects when each feature is ablated individually. This synergistic effect is observed only when ablating language-specific features within its language. Interestingly, in layers 7, 9, 10, 11, 14, and 15, the rank #2 French feature is also among the top-2 Spanish features, explaining the significant changes in Spanish in some layers.

6 Enhancing Steering Vectors Using Language-Specific Features

Having studied the basic characteristics of language-specific features, we now explore how to leverage these features in practice. Concretely, we use language-specific features as signals to guide steering vector (Turner et al., 2024; Rimsky et al., 2024; Mayne et al., 2024), in order to control the language in the model.

6.1 Experimental Settings

Tasks for Evaluation. We propose two tasks for evaluation. In the first task, *Adversarial Language Identification*, given a text in language A, we prompt the model to identify its language. Our goal is to make the model identify the text as language B instead. We use the CE loss for predicting language B as the metric. In the second task, *Cross-Lingual Continuation*, given a text in language A, our goal is to make the model continue the text in language B. We use a language identification model from Burchell et al. (2023) to verify if the continuation is in language B. The success rate is used as the metric. Additionally, to measure the impact of the method on other language capabilities of LLMs, we also calculate the CE loss on Flores-10 without the original language when using the method.

6.2 Methods

Steering Vectors. Steering vectors are vectors in the space of model activations that can guide a model's behavior when added to its internal activations (Turner et al., 2024; Rimsky et al., 2024; Mayne et al., 2024). To extract steering vectors for previously mentioned tasks, we use a subset in language A from Flores-10 as positive prompt set, and another subset in language B as negative prompt set, then we calculate the difference between the mean activations for positive and negative prompts at all token positions in layer L. This yields a steering vector $\mathbf{v}$, defined as:

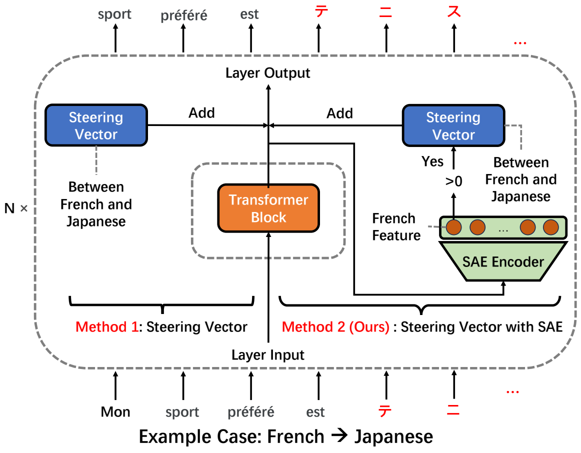


Figure 7: Comparison of steering vector and improved steering vector with SAE (ours): An example case from French to Japanese.

$$\mathbf{v} = \frac{1}{|\mathcal{X}_+|} \sum_{\mathbf{x} \in \mathcal{X}_+} a_L(\mathbf{x}) - \frac{1}{|\mathcal{X}_-|} \sum_{\mathbf{x} \in \mathcal{X}_-} a_L(\mathbf{x}), \quad (5)$$

where $\mathcal{X}_+$ and $\mathcal{X}_-$ are positive and negative prompt sets, and $a_\mathbf{L}(\mathbf{x})$ represent the mean activations in layer L for prompt $\mathbf{x}$.

During inference, these steering vectors are directly added to the corresponding layer's activations across all tokens, replacing the original activations to continue the forward pass. By modifying the model's activations with the steering vector, the internal activation of the prompt can be steered from language A towards language B, potentially improving performance on the language switching task.

Improved Steering Vectors with Language-Specific Features. There are two main drawbacks of steering vectors: (1) Adding a steering vector across all tokens, including non-target language tokens, leads to an increase in CE loss in non-target language texts. (2) Using a steering vector in multiple layers simultaneously does not lead to better performance since frequent adjustments may drastically change the normal distribution of activation values, as demonstrated in Figure 8. As a result, we propose using the activation of language-specific features as a signal to determine whether to use a steering vector for a token. Concretely, for a steering vector from language B to language A in layer L and a model activation x in layer L, if the activation of the top-2 language B features of $\mathbf{x}$ are non-zero, we add the steering vector to it and continue the forward pass. The rationale for employing language-specific features is twofold: First, it ensures that steering vectors affect only target language tokens, preventing increased CE loss in non-target language texts. Second, by applying steering vectors selectively based on specific language feature activations, it minimizes excessive adjustments across layers, maintaining activation value distribution and enhancing model stability.[5]

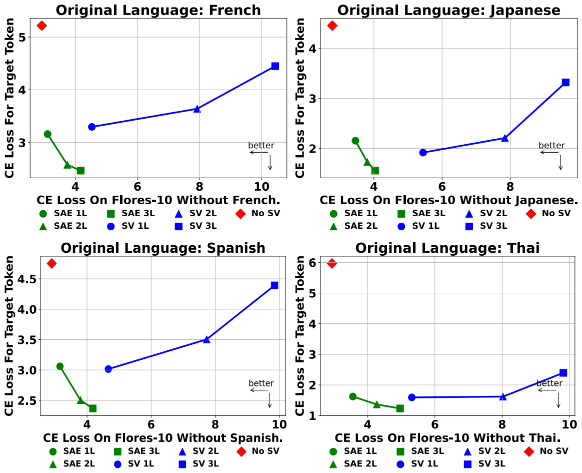


Figure 8: The y-axis shows the CE loss of LLMs identifying text in the original language as if it were in the target language, with lower values indicating better performance. The x-axis shows the impact on non-original language texts, with lower values indicating less impact. "SV" is the "steering vector" method, while "SAE" is our enhanced method. The suffix kL indicates results from modifying k consecutive layers. Our method (green) provides a better balance between the metrics.

6.3 Results

Better Performance on Adversarial Language Identification. In Figure 8, we present the results for *Adversarial Language Identification*. The experiments cover the steering vector both with and without SAE across one, two, and three consecutive layers, denoted as $1L$, $2L$, and $3L$. From the figure, we make the following observations: (1) As the number of modified layers increases, our SAE method achieves better performance in *Adversarial Language Identification*, whereas the performance of the SV method decreases. This indicates that SAE is more effective at *Adversarial Language Identification* and more robust when applied to multiple layers. (2) As the number of modified layers increases, the CE loss on other languages also increases. However, the rate of increase with

[5] In cases where there is no ambiguity, we abbreviate the "steering vector" as "SV" and our improved "steering vector with SAE" as "SAE." And we add a suffix kL to indicate modification of k consecutive layers.

Model	Method	Success rate ↑/ CE loss on other language ↓								
		Es	Fr	Pt	Ja	Ko	Th	Vi	Zh	Ar
Gemma 2 2B	SV L1	92.1 / 4.7	92.6 / 4.5	84.2 / 4.7	86.1 / 5.4	95.2 / 5.3	85.7 / 5.3	91.1 / 4.6	**84.7** / 5.2	**88.3** / 4.6
	SAE L3	**95.8 / 4.2**	**96.7 / 4.2**	**84.4 / 4.4**	**89.2 / 4.0**	**95.4 / 4.4**	**90.7 / 5.0**	**91.3 / 3.4**	71.9 / **4.3**	81.3 / **3.9**
Gemma 2 9B	SV L1	82.2 / 4.1	85.3 / 4.0	76.4 / 4.1	83.4 / 4.5	93.0 / 4.6	**88.7 / 4.6**	83.6 / 4.1	**79.5** / 4.4	**84.2** / 4.1
	SAE L3	**96.2 / 3.4**	**94.6 / 2.9**	**86.1 / 3.2**	**86.3 / 3.0**	**93.6 / 4.3**	85.6 / 4.9	**95.3 / 2.8**	77.0 / **3.2**	78.3 / **4.0**
Llama-3.1-8B	SV L1	85.7 / 3.7	86.8 / 3.5	79.7 / 3.6	79.1 / 4.4	90.0 / 4.3	**88.4** / 4.4	85.0 / **3.6**	77.0 / 4.1	85.2 / 3.8
	SAE L3	**97.0 / 3.0**	**96.1 / 2.7**	**80.0 / 2.7**	**86.6 / 3.6**	**95.2 / 3.2**	78.0 / **3.3**	**94.8** / 4.7	**91.2 / 3.7**	**88.4 / 2.7**

Table 1: The results of *Cross-Lingual Continuation* task. Our SAE method can surpass the SV method across both metrics in most cases and achieve a much better balance between the metrics.

our SAE method is much smaller than with the SV method. In most cases, the CE loss on other languages for SAE applied to three consecutive layers is even lower than that for SV applied to a single layer. These results suggest that our SAE method achieves a better balance between the two metrics on *Adversarial Language Identification*.

Better Performance on Cross-Lingual Continuation. As illustrated in Figure 8, the performance of the SV method declines rapidly; hence, we only report the results of SV 1L for *Cross-Lingual Continuation*. The results are shown in Table 1, where we observe the following: (1) For different languages and models, SAE 3L outperforms SV 1L in both success rate and CE loss in most cases. (2) In some cases, SAE 3L achieves a better CE loss but with a lower success rate compared to SV 1L. Since these two metrics are generally a trade-off, this does not imply that the SAE method is inferior to the SV method. These results suggest that our SAE method can surpass the SV method across both metrics in most cases.

7 Related Works

Multilingual Mechanism of LLMs Multilingual mechanisms of LLMs are mainly studied through neuron-based and "logit lens" (nostalgebraist, 2020) methods. Neuron-based aim to identify language-specific neurons within LLMs and modify these neurons to assess their impact on the corresponding language (Zhang et al., 2024b; Zhao et al., 2024; Tang et al., 2024; Kojima et al., 2024). For example, Zhang et al. (2024b) discover that removing certain neurons in LLMs leads to a significant performance decrease in some languages. Zhao et al. (2024) introduce PLND to identify activated neurons for inputs in different languages, and hypothesize that in the intermediate layers, LLMs employ English for thinking. Moreover, Tang et al. (2024); Kojima et al. (2024) explore methods to identify language-specific neurons within LLMs. However, these methods can be complex and unreliable due to "superposition," (Elhage et al., 2022) where multiple concepts can be encoded in a single neuron. "Logit lens" methods derive token distributions from intermediate layers using the output layer's unembedding matrix. Wendler et al. (2024) find that Llama2 (Touvron et al., 2023) might use English as an internal language in intermediate layers, and Zhong et al. (2024) extend the conclusion, showing that LLMs with continued pre-training in Japanese employ both Japanese and English in intermediate layers. Due to the varying distribution of residual streams across different layers, the "logit lens" method often has significant errors except in the last few layers, making the analysis sometimes unreliable.

SAEs SAEs are a specialized form of autoencoders designed to decompose language model activations into a linear combination of SAE feature directions (Bricken et al., 2023; Cunningham et al., 2023). Typically, the activations of neurons in deep neural networks do not have a straightforward, human-understandable interpretation. However, SAEs can transform these activations into a higher-dimensional latent space, which is potentially more interpretable. For instance, Cunningham et al. (2023) identify features associated with apostrophes. Meanwhile, Ferrando et al. (2025) discover features that indicate whether LLMs recognize a particular entity. Furthermore, Paulo et al. (2024) develop an open-source automated pipeline to generate and evaluate natural language explanations for SAE features using LLMs. Their work confirms that SAEs are indeed significantly more interpretable than individual neurons.

8 Conlusion

In this study, we explored the underlying mechanisms of multilingual capabilities in LLMs using SAEs to achieve a more refined analysis. By introducing a novel metric for monolinguality, we found that certain features were strongly tied to specific languages. And directional ablation confirmed the significant role these features play in enhancing language-specific capabilities, with combined feature ablation yielding greater improvements than individual ablation. Additionally, we improved steering vectors using these SAE-derived features, achieving better performance and robustness. Building upon the insights gained from this work, an exciting avenue for future research is to utilize these language-specific features to guide the training process of multilingual language models.

Limitations

This study has several limitations that we plan to address in the future. First, although our method performs well across 10 different languages, it does not yet cover certain low-resource languages. Investigating these underrepresented languages will enhance our analysis. Second, while our SAE-based steering vectors outperform the original steering vectors in most cases (see Table 1), there are instances where our method falls short. Therefore, exploring a more refined approach to improve performance is worthwhile. Third, the SAEs used in our experiments were not trained on curated multilingual data. It would be advantageous to train SAEs using high-quality multilingual datasets.

Acknowledgements

This research was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 865–879. Association for Computational Linguistics.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. 2024. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. *Preprint*, arXiv:2411.14257.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. *Preprint*, arXiv:2411.14257.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Geoffrey E. Hinton and Richard S. Zemel. 1993. Autoencoders, minimum description length and helmholtz free energy. In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 3–10. Morgan Kaufmann.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6919–6971. Association for Computational Linguistics.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Preprint*, arXiv:2303.10475.

Harry Mayne, Yushi Yang, and Adam Mahdi. 2024. Can sparse autoencoders be used to decompose and interpret steering vectors? *Preprint*, arXiv:2411.08790.

nostalgebraist. 2020. Interpreting gpt: the logit lens. *LessWrong*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. 2024. Automatically interpreting millions of features in large language models. *Preprint*, arXiv:2410.13928.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *Preprint*, arXiv:2407.14435.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15504–15522. 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5701–5715. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15366–15394. Association for Computational Linguistics.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. Personalized generation in large model era: A survey. *arXiv preprint arXiv:2503.02614*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024a. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6228–6247. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *Preprint*, arXiv:2408.10811.

A Implementation Details

For each experiment, we don't perform any sampling during generation to avoid randomness. For SAEs from Gemma Scope (Lieberum et al., 2024), we choose the one with the second smallest L_0 value for each layer. For SAEs from Llama Scope (He et al., 2024), we use the model available at `https://huggingface.co/fnlp/Llama3_1-8B-Base-LXR-8x/tree/main`.

B Flores-10

We extract a subset called Flores-10 from Flores-200, which includes 10 languages: English (en), Spanish (es), French (fr), Japanese (ja), Korean (ko), Portuguese (pt), Thai (th), Vietnamese (vi), Chinese (zh), and Arabic (ar). In Section 3, we use the first 100 data points in the dev set for each language to identify language-specific features. In Section 5, we use the first 100 data points in the dev set for each language to generate the steering vector, and perform experiments using 500 data points in the dev set for each language that do not overlap with the first 100 data points.

C Additional Results of ν

Additional results of ν for 3 different LLMs are demonstrated in Figure 10-12. We report the results at four different levels: "first layers", "$\frac{1}{3}$ of the total layers", "$\frac{2}{3}$ of the total layers", and the "final layer". Similarly to the results in Figure 1, the top-ranked features possess strong monolingual characteristics, and in most scenarios, the top-1 feature suffices in capturing these characteristics.

D Additional Results for Directional Ablation

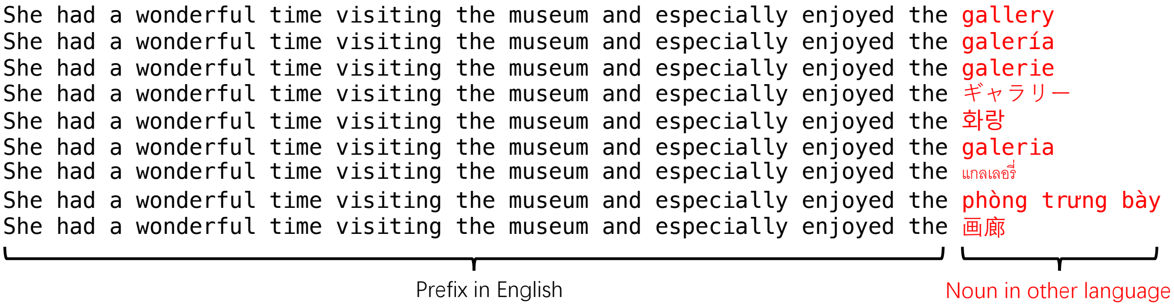
Additional results for directional ablation for 3 different LLMs are demonstrated in Figure 13-15. The results are similar to those in Figure 5, where ablating language-specific features has a much larger impact on the CE loss of texts in the target language compared to texts in other languages.

E Additional Results for Code-Switching

Additional results for code-switching are demonstrated in Figure 16-33. The results are similar to those in Figure 3 and 4.

F Additional Results for Multiple Features

Additional results for multiple features are demonstrated in Figure 34-42. The results are similar to those in Figure 6.



```
She had a wonderful time visiting the museum and especially enjoyed the gallery
She had a wonderful time visiting the museum and especially enjoyed the galería
She had a wonderful time visiting the museum and especially enjoyed the galerie
She had a wonderful time visiting the museum and especially enjoyed the ギャラリー
She had a wonderful time visiting the museum and especially enjoyed the 화랑
She had a wonderful time visiting the museum and especially enjoyed the galeria
She had a wonderful time visiting the museum and especially enjoyed the แกลเลอรี่
She had a wonderful time visiting the museum and especially enjoyed the phòng trưng bày
She had a wonderful time visiting the museum and especially enjoyed the 画廊
```

Figure 9: Example data of our code-switching dataset.

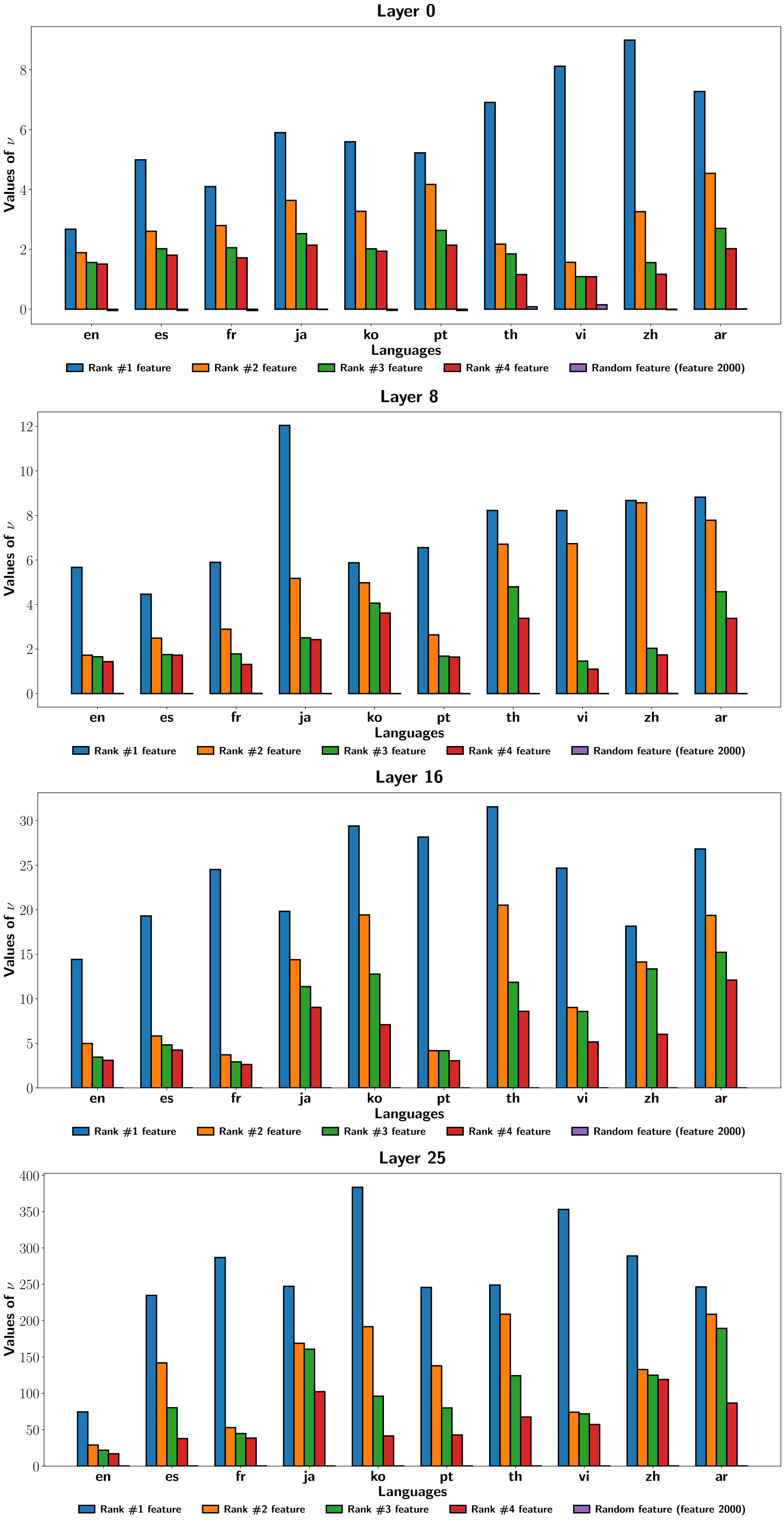


Figure 10: The values of ν of Gemma 2 2B.

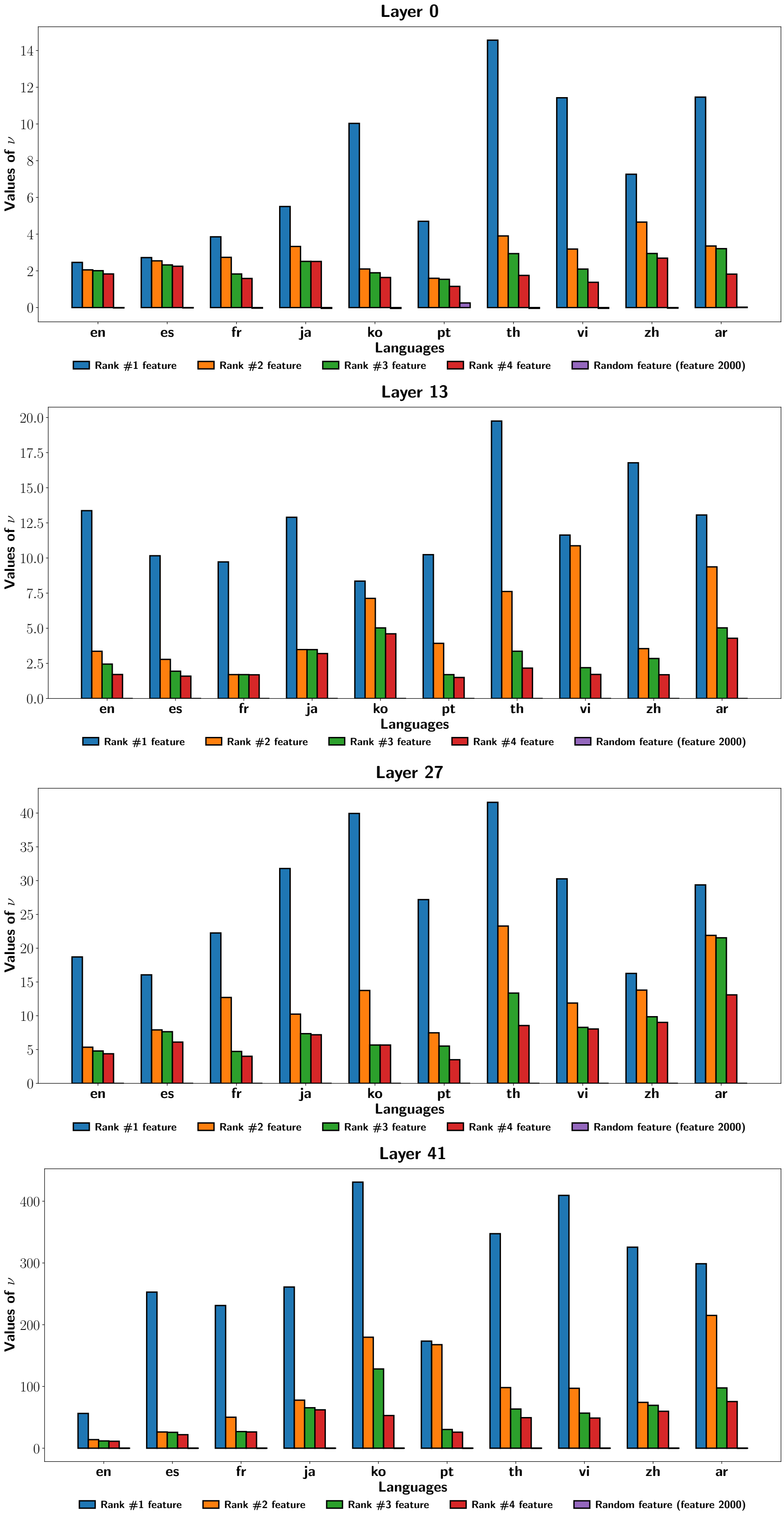
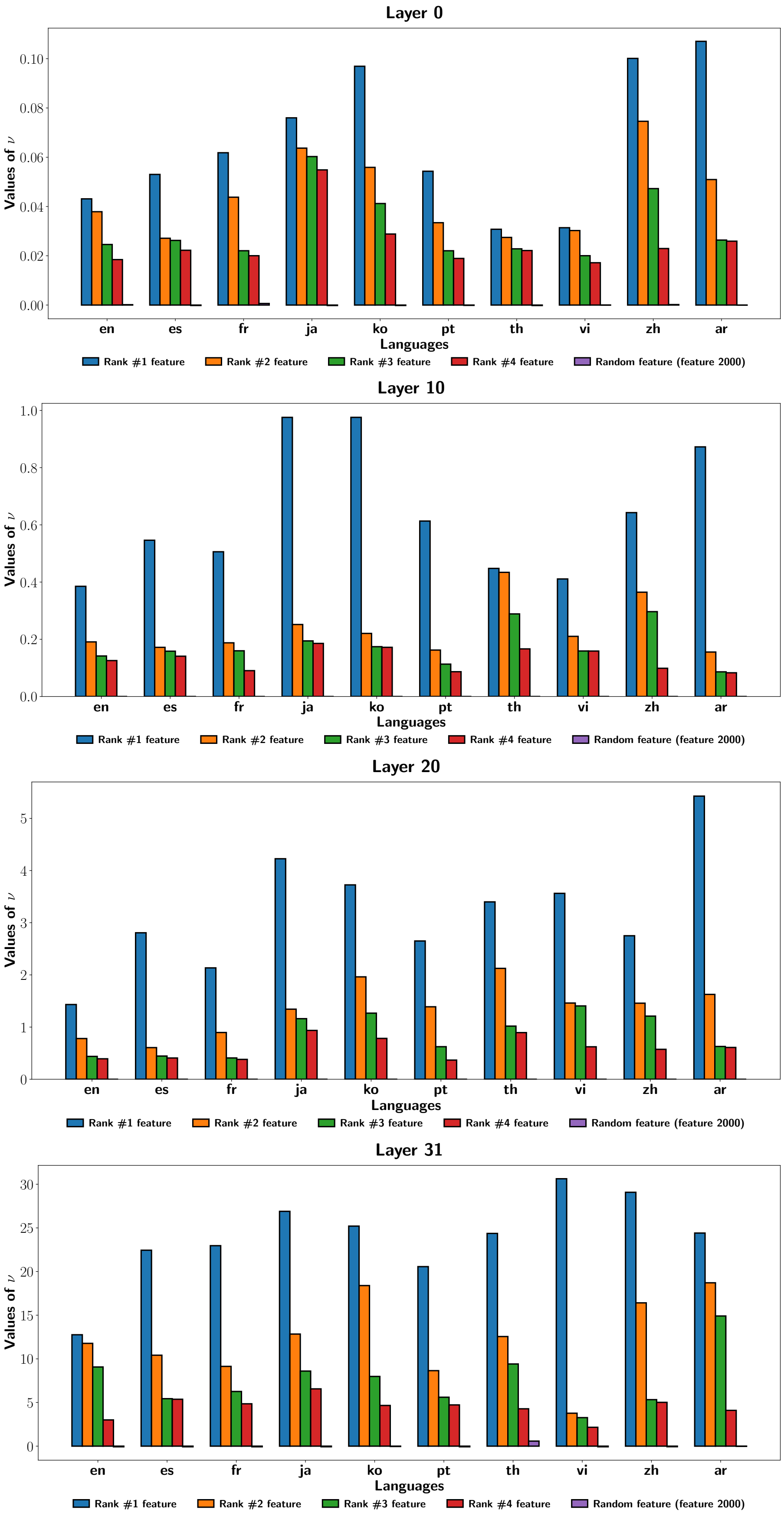


Figure 11: The values of ν of Gemma 2 9B.



Layer 0

Layer 10

Layer 20

Layer 31

Figure 12: The values of ν of Llama-3.1-8B.

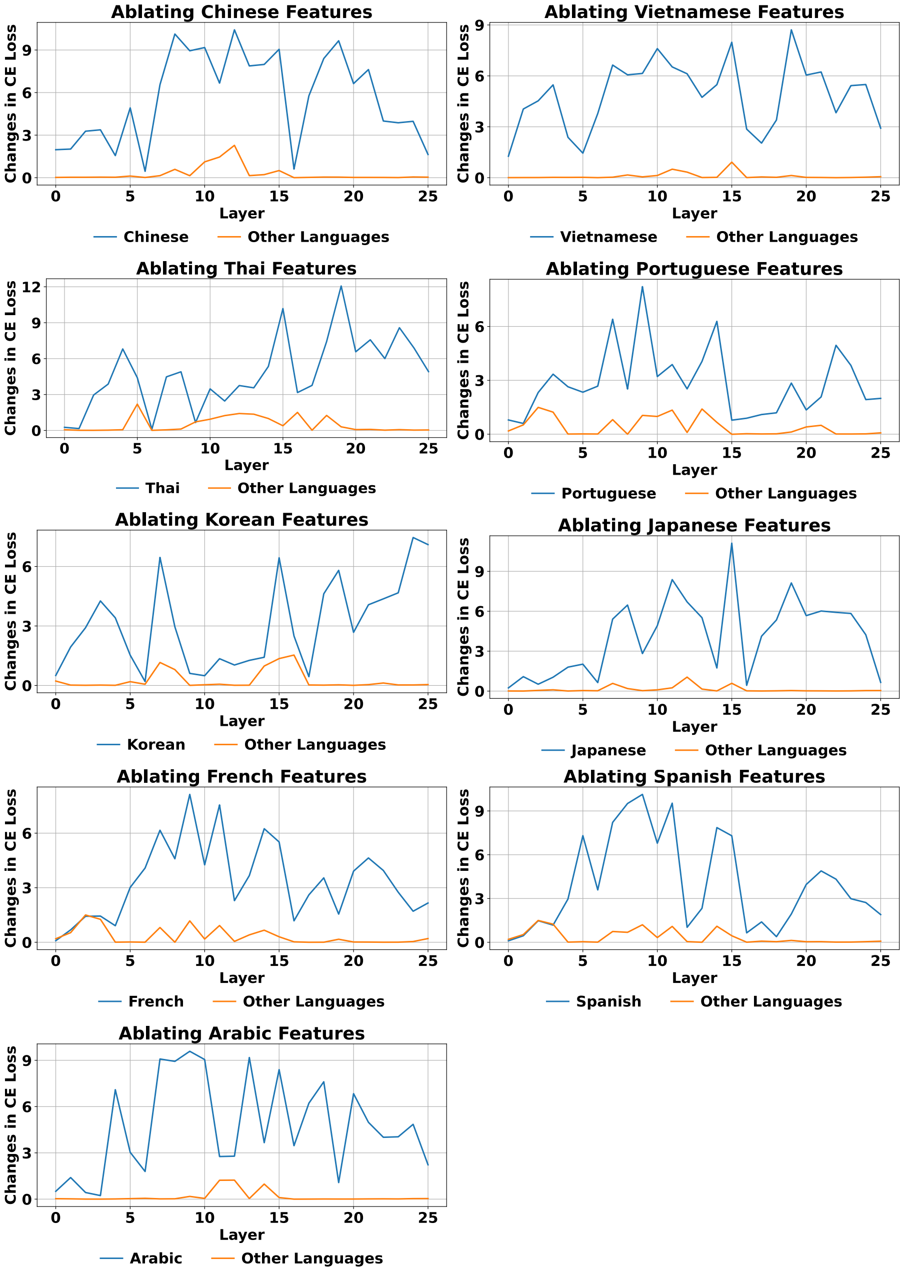


Figure 13: The changes in CE loss on texts in the target language and texts in other languages after ablating language-specific features for Gemma 2 2B.

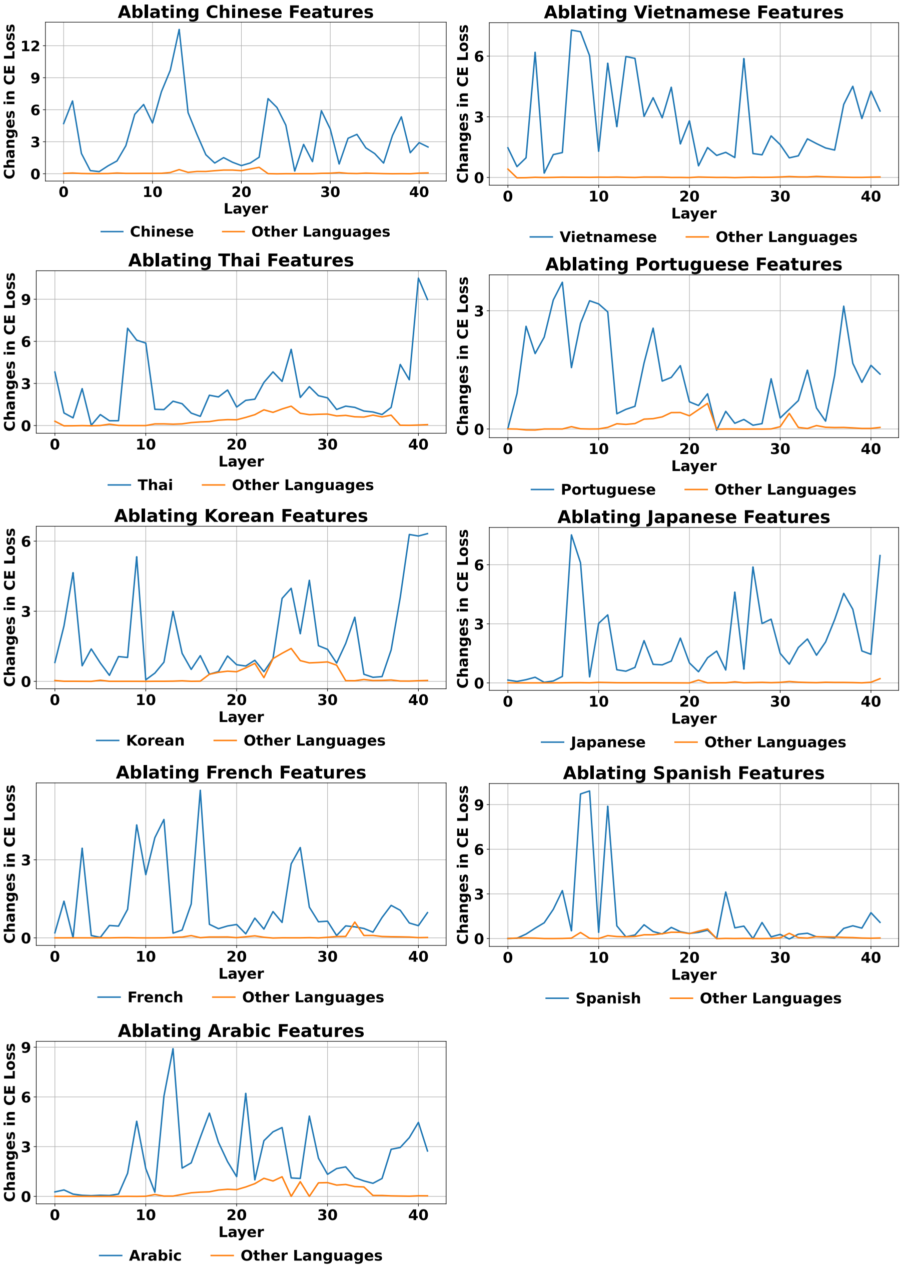


Figure 14: The changes in CE loss on texts in the target language and texts in other languages after ablating language-specific features for Gemma 2 9B.

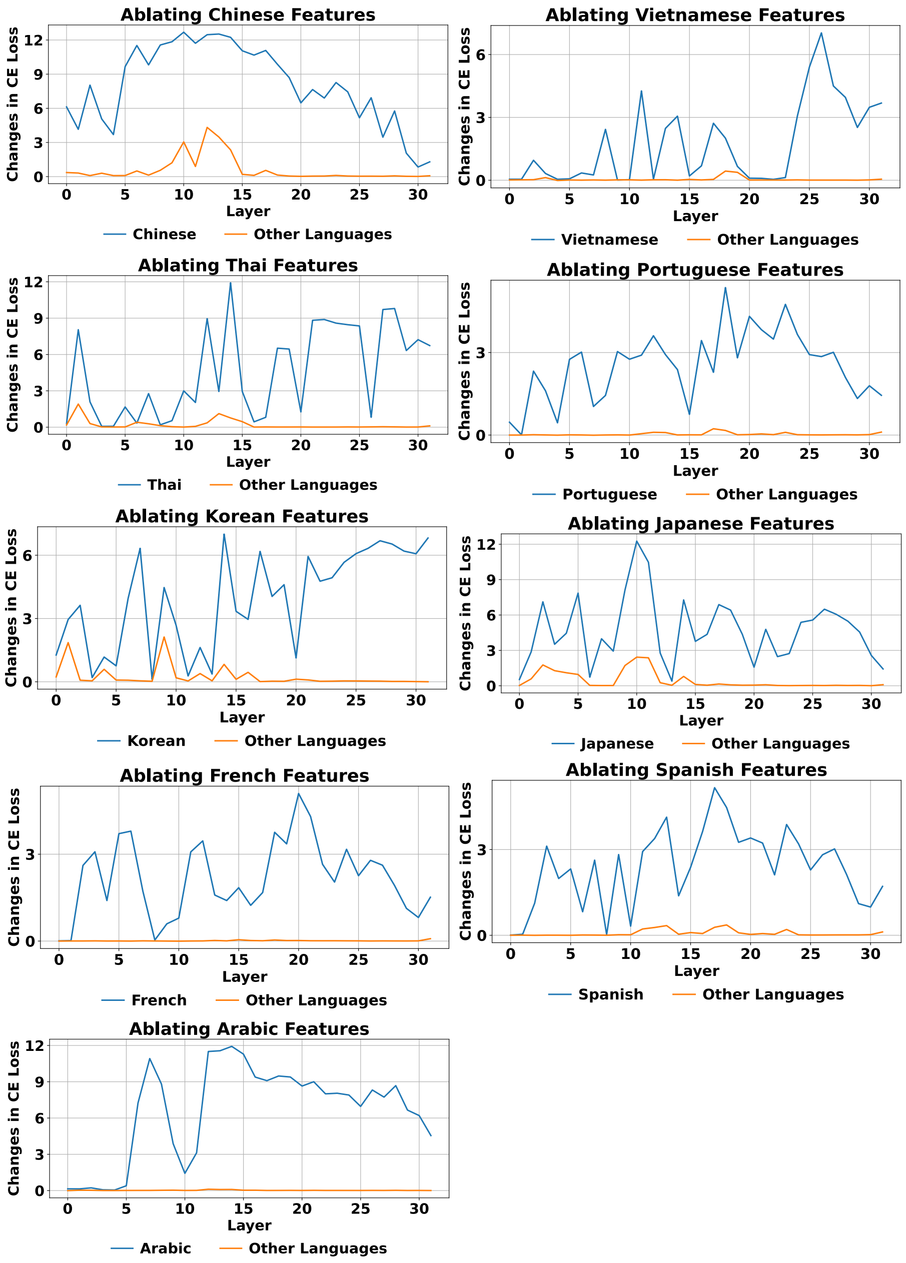


Figure 15: The changes in CE loss on texts in the target language and texts in other languages after ablating language-specific features for Llama-3.1-8B.

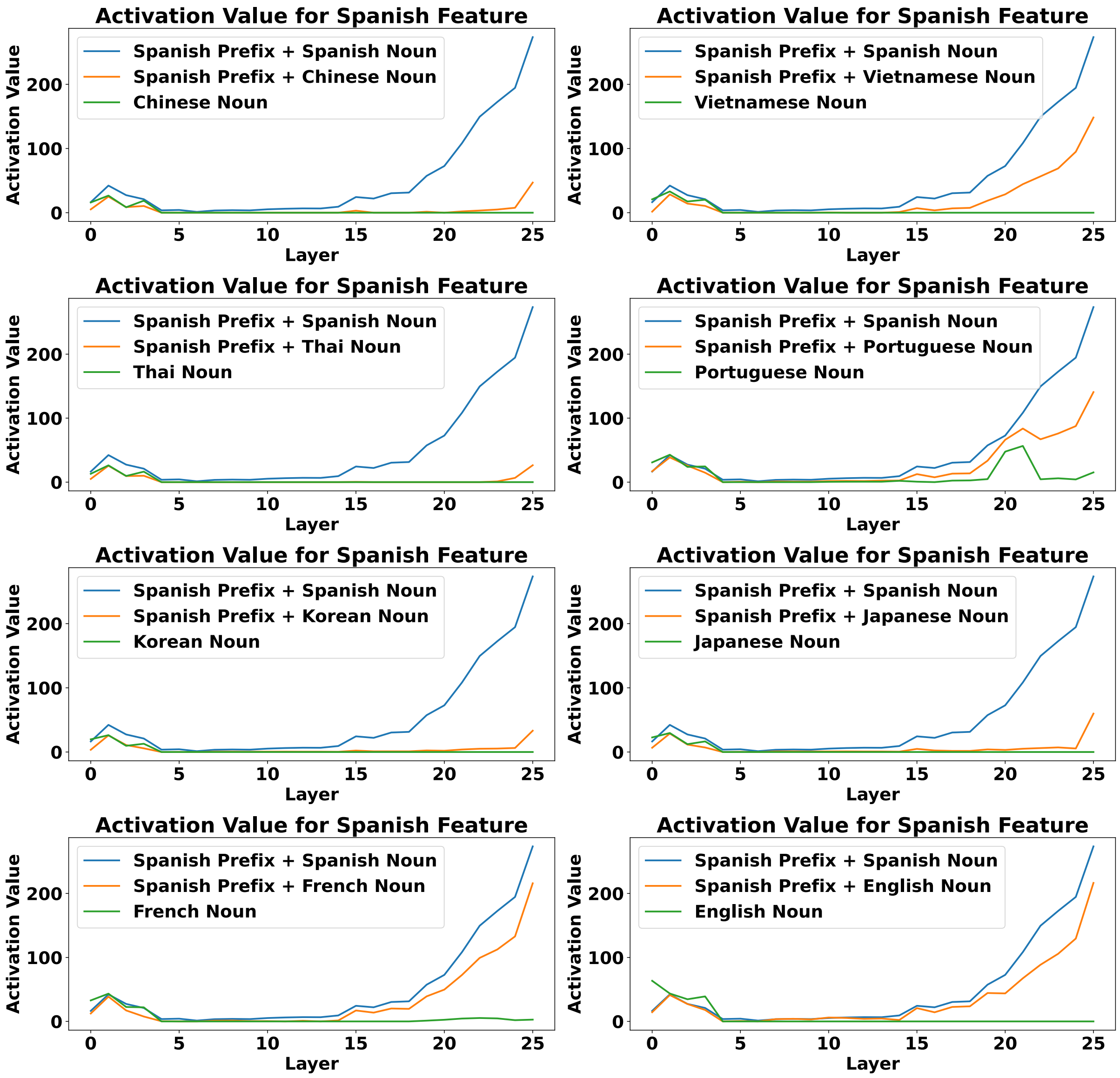


Figure 16: The mean activation values for the Spanish feature with various noun and prefix combinations for Gemma 2 2B.

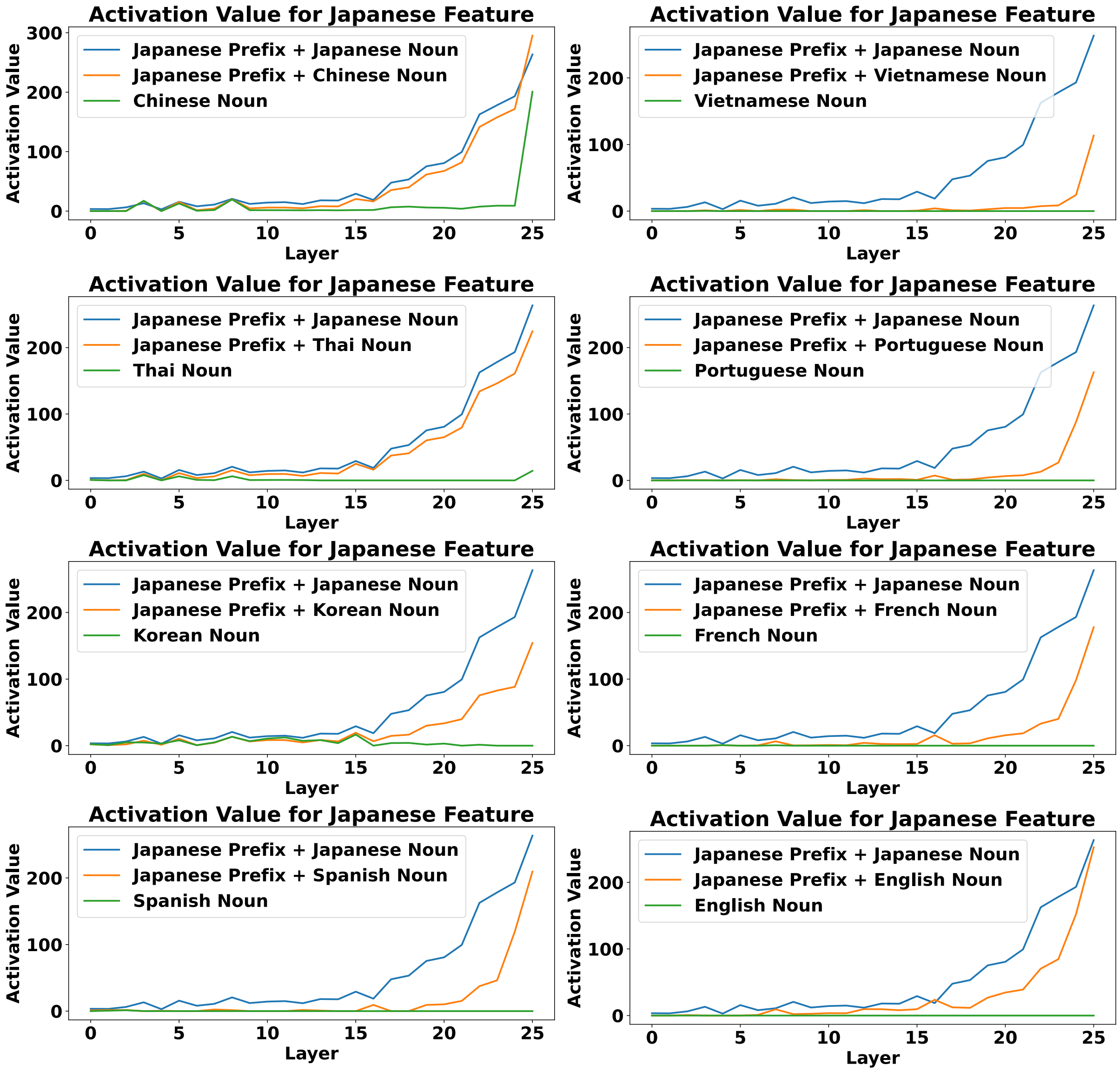


Figure 17: The mean activation values for the Japanese feature with various noun and prefix combinations for Gemma 2 2B.

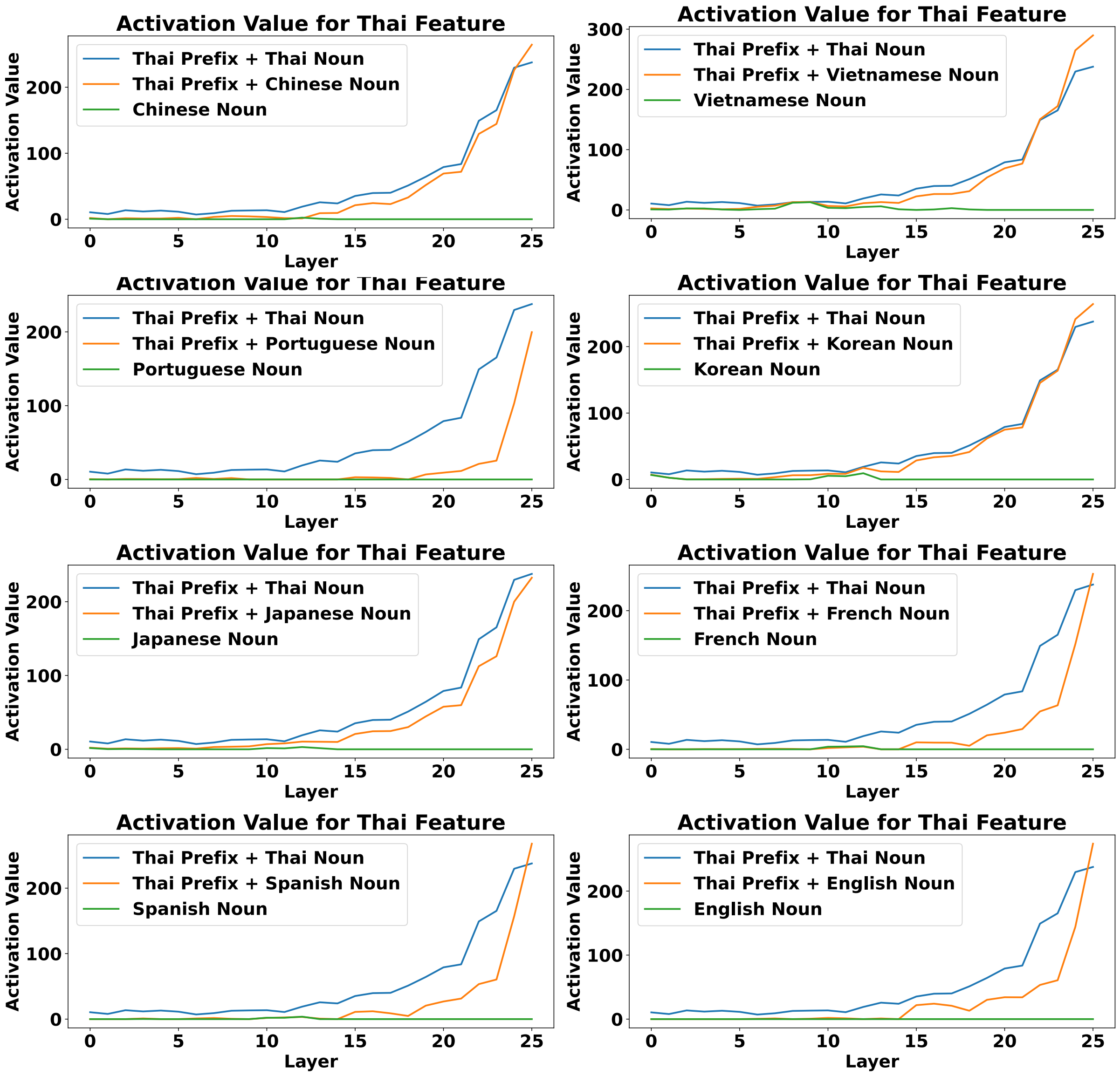


Figure 18: The mean activation values for the Thai feature with various noun and prefix combinations for Gemma 2 2B.

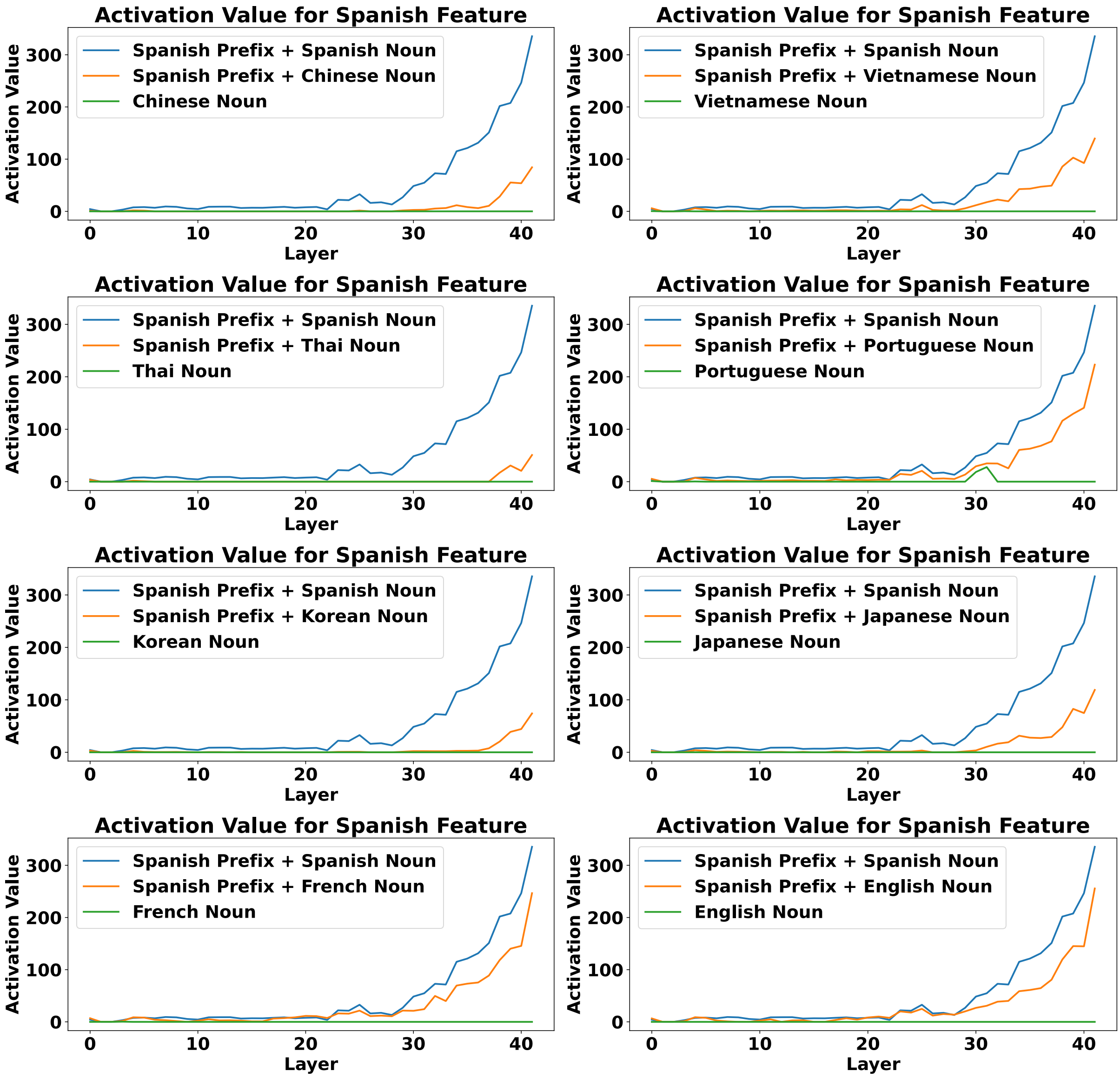


Figure 19: The mean activation values for the Spanish feature with various noun and prefix combinations for Gemma 2 9B.

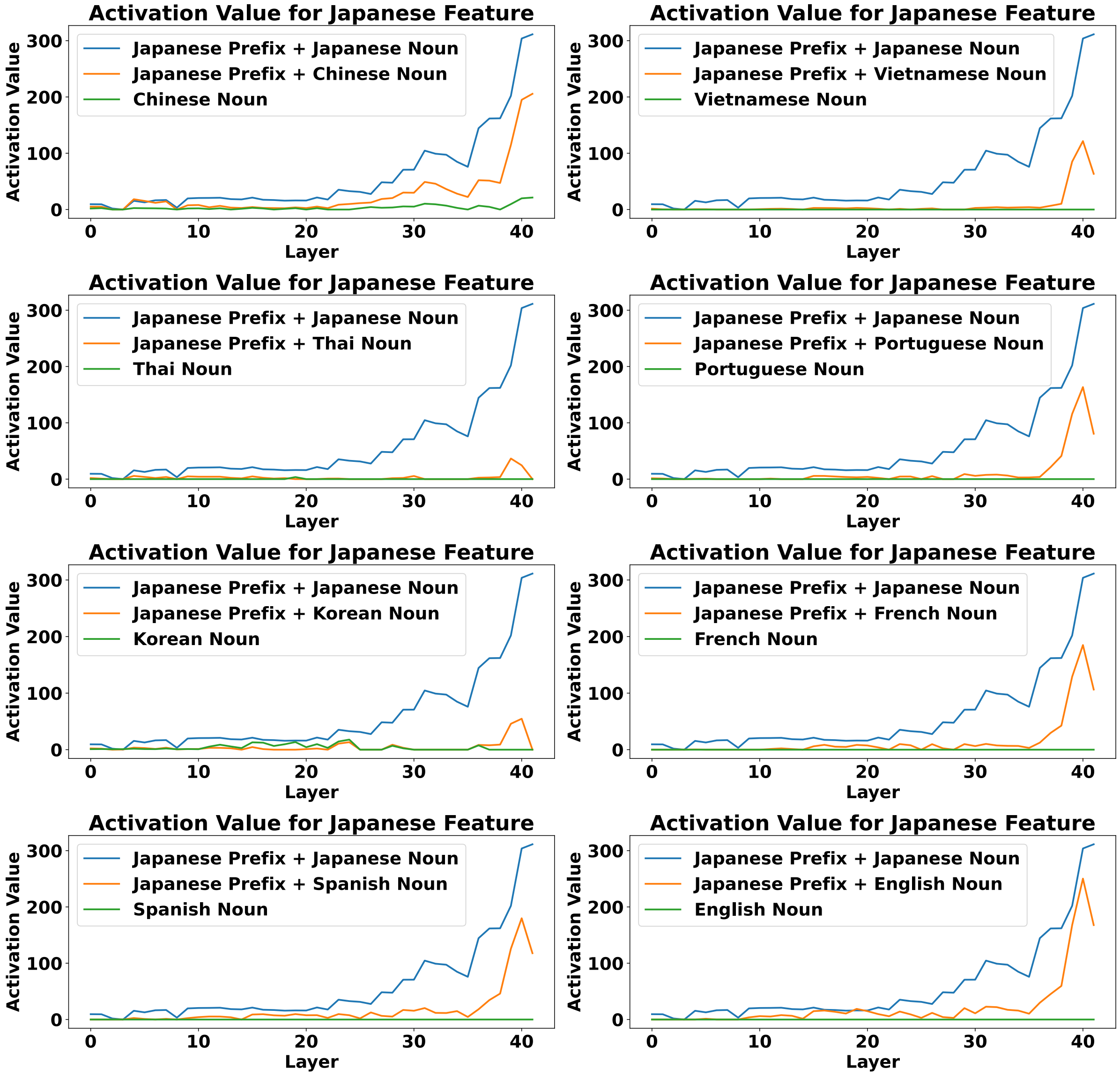


Figure 20: The mean activation values for the Japanese feature with various noun and prefix combinations for Gemma 2 9B.

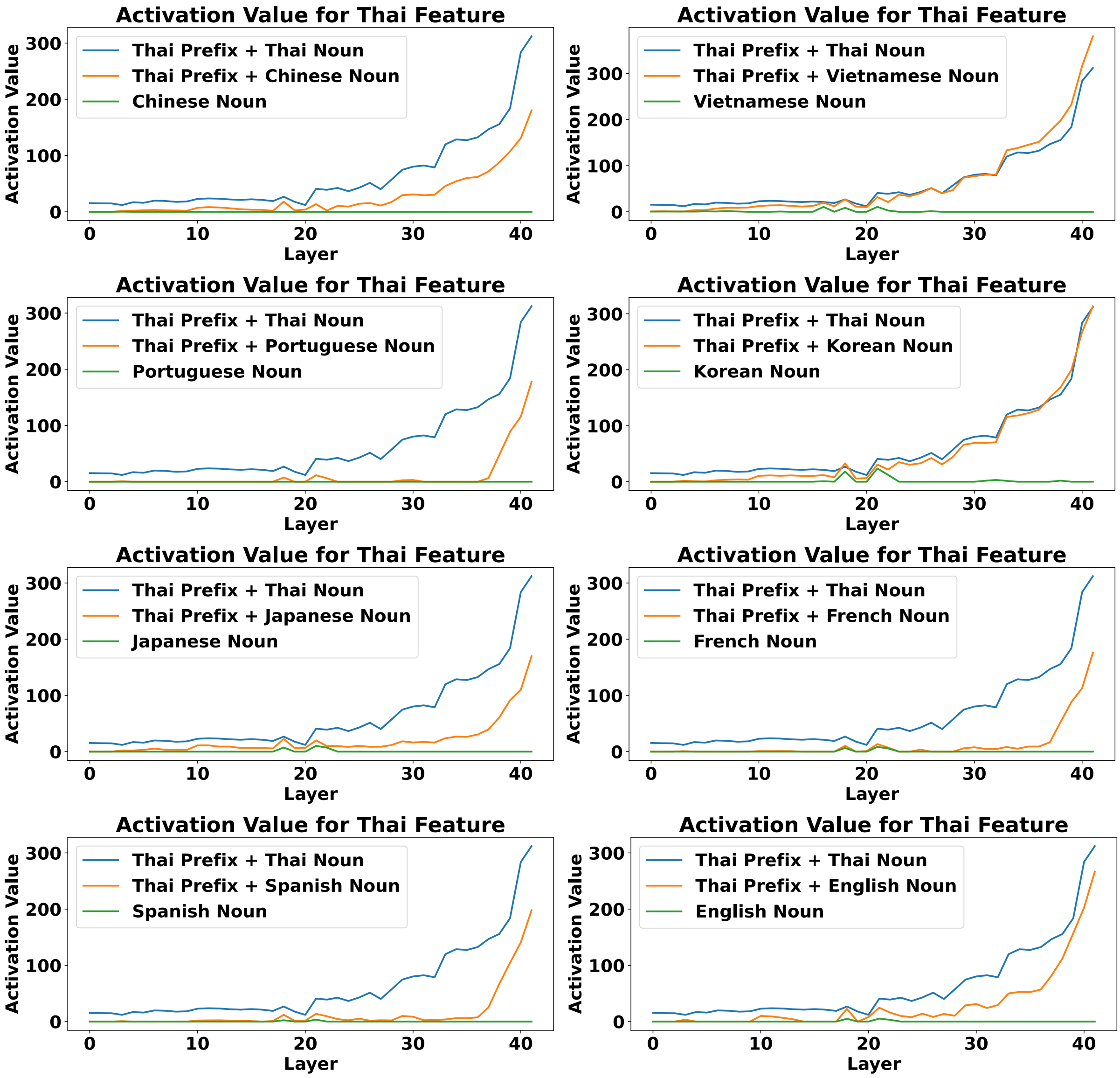


Figure 21: The mean activation values for the Thai feature with various noun and prefix combinations for Gemma 2 9B.

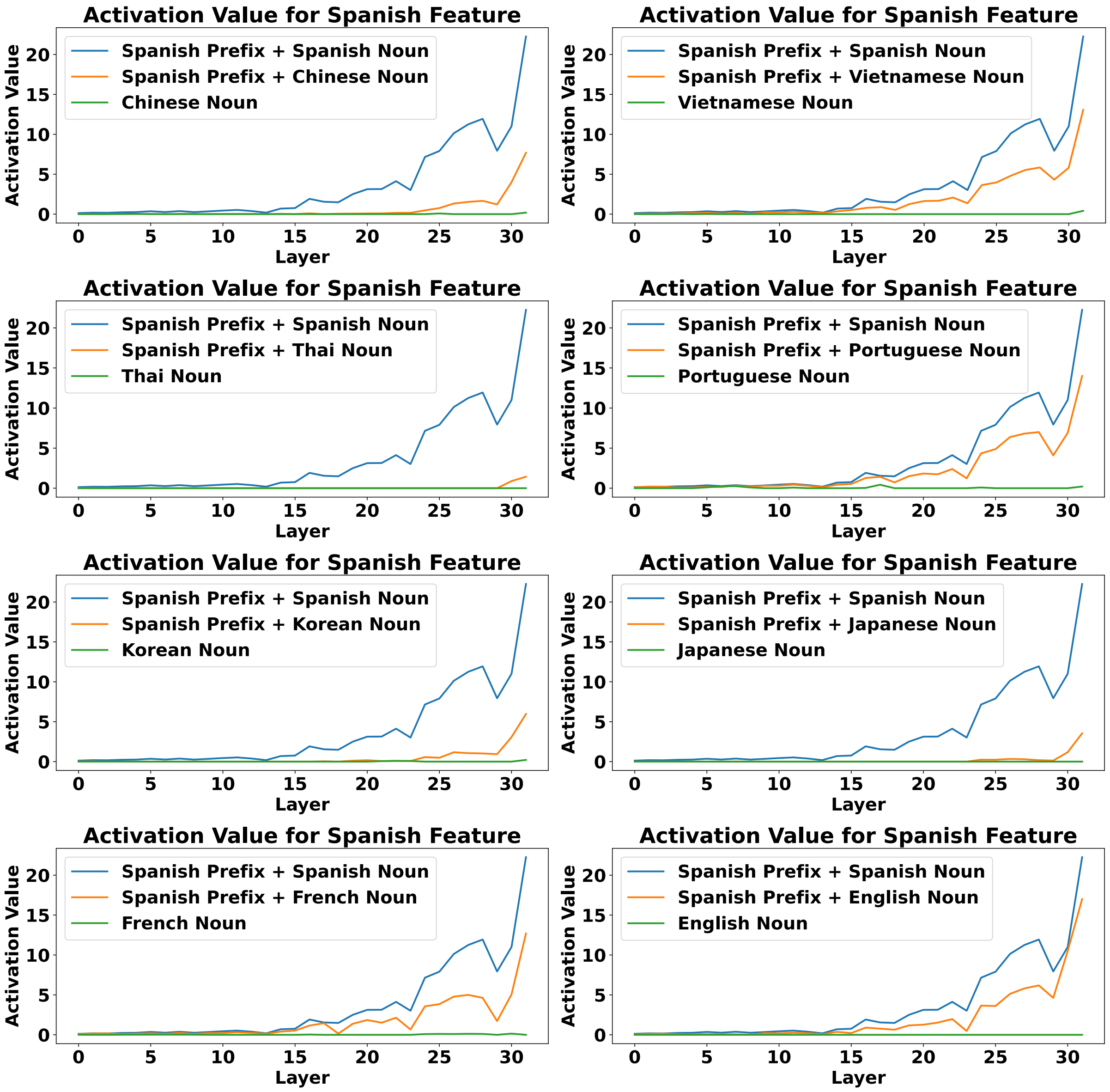


Figure 22: The mean activation values for the Spanish feature with various noun and prefix combinations for Llama-3.1-8B.

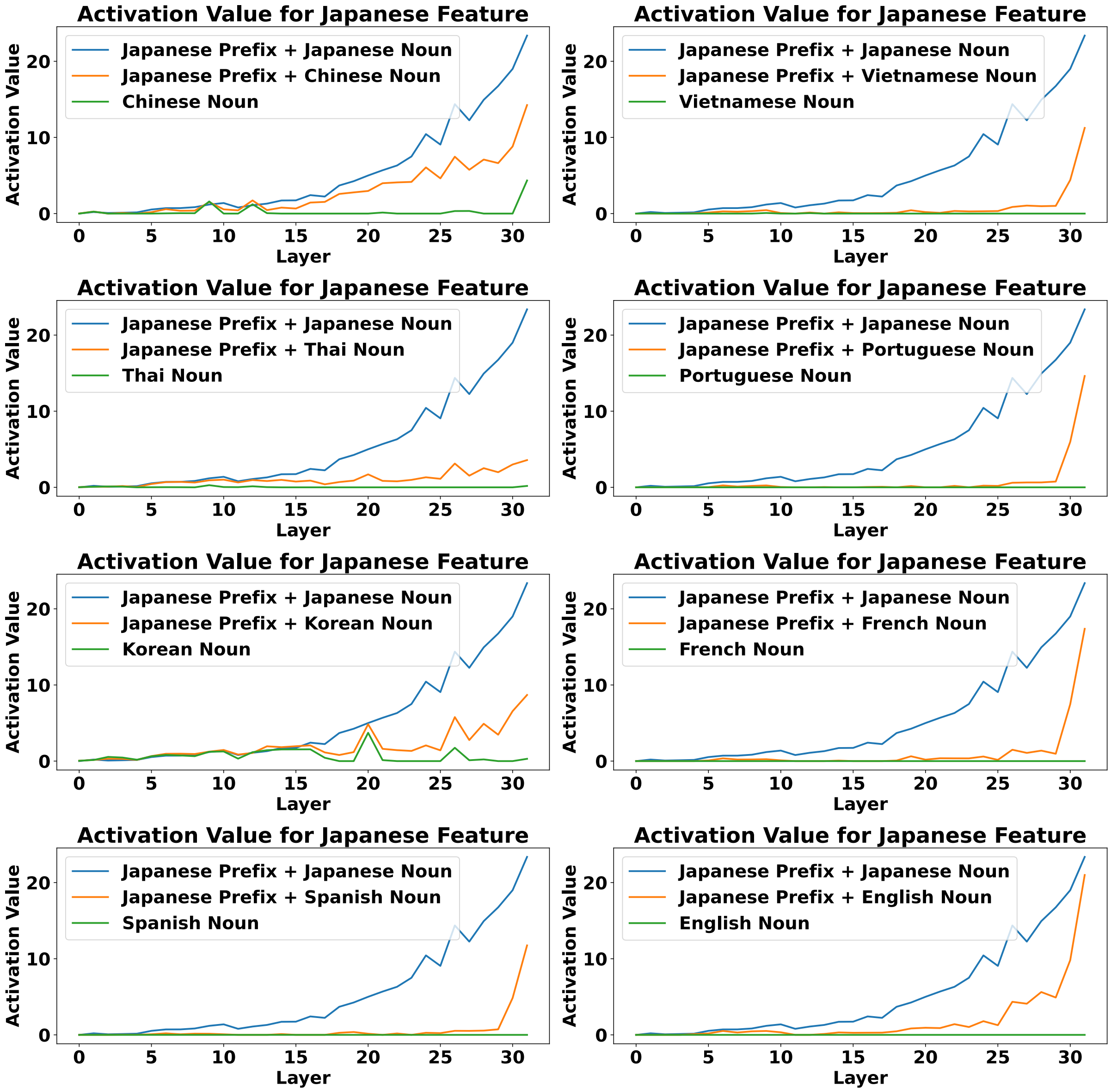


Figure 23: The mean activation values for the Japanese feature with various noun and prefix combinations for Llama-3.1-8B.

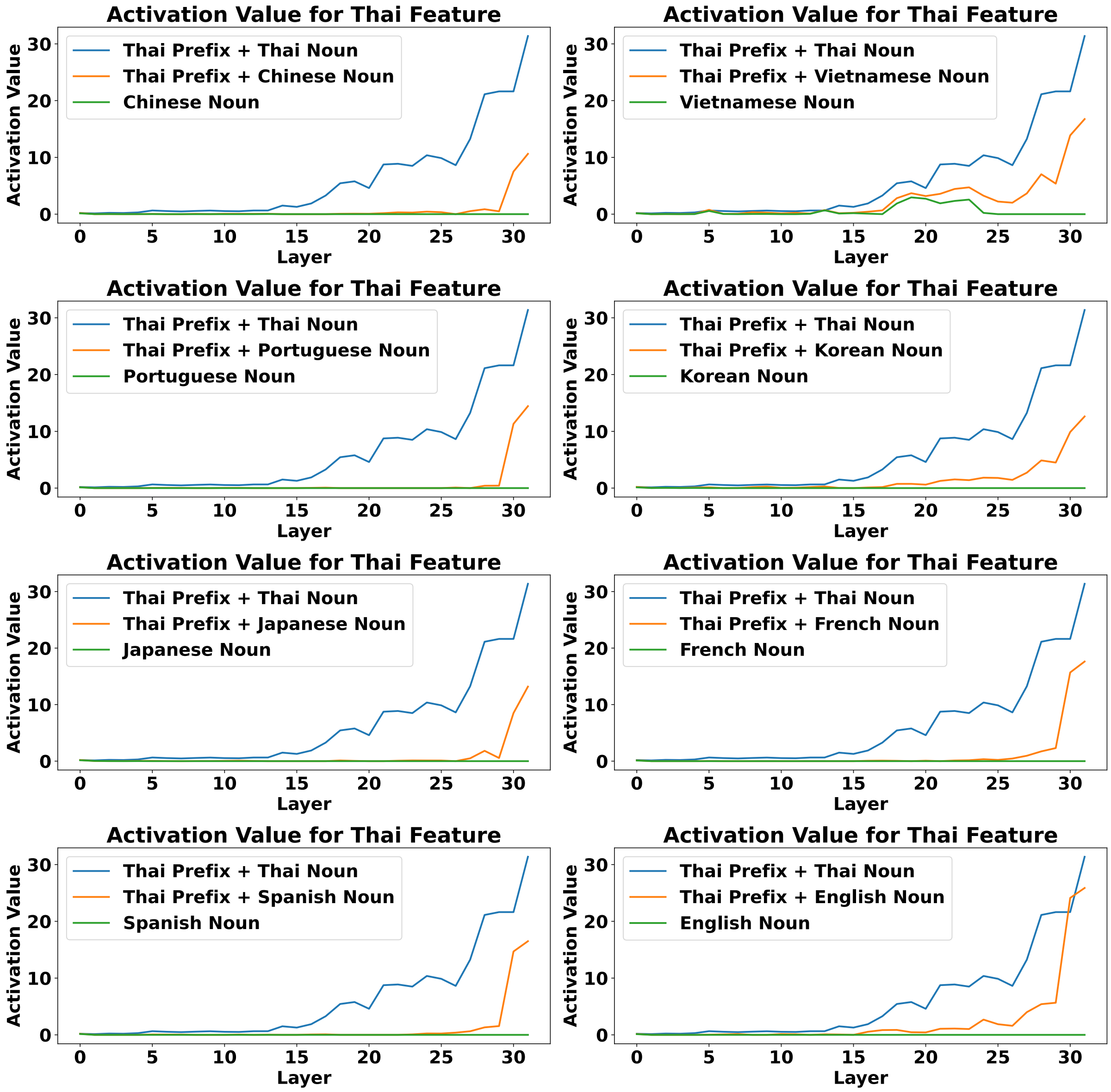


Figure 24: The mean activation values for the Thai feature with various noun and prefix combinations for Llama-3.1-8B.

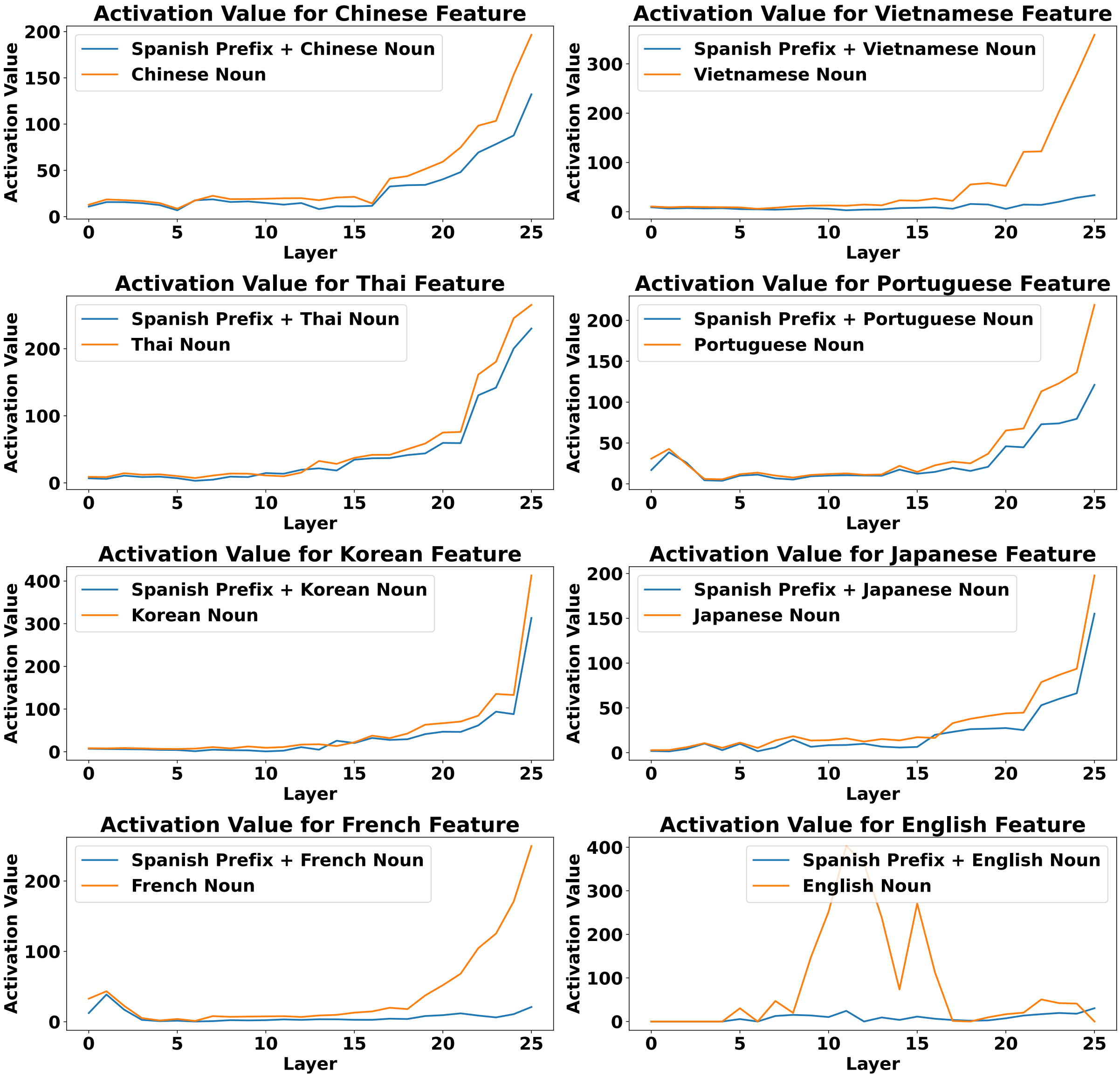


Figure 25: The mean activation values for various features with Spanish prefix for Gemma 2 2B.

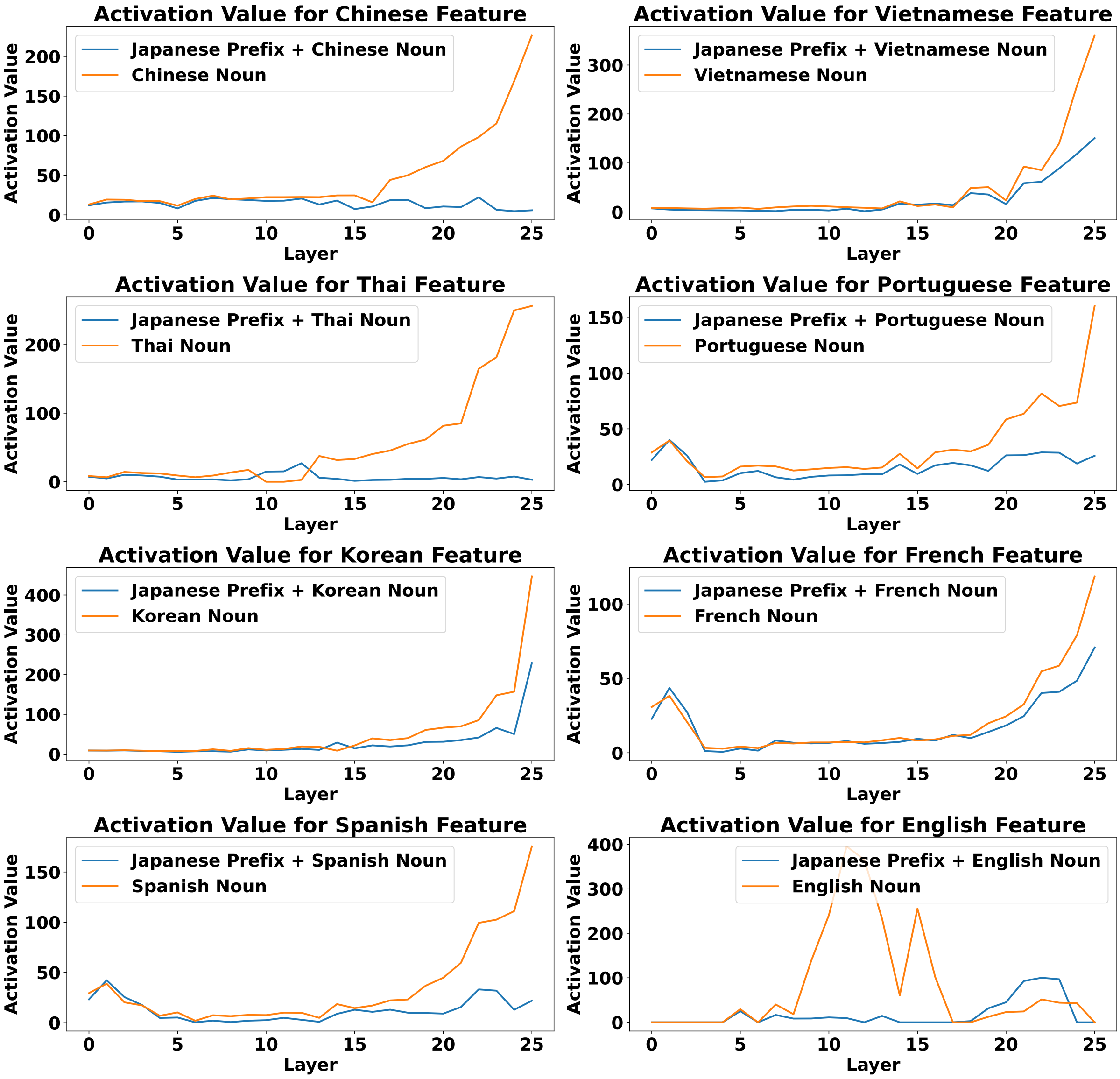


Figure 26: The mean activation values for various features with Japanese prefix for Gemma 2 2B.

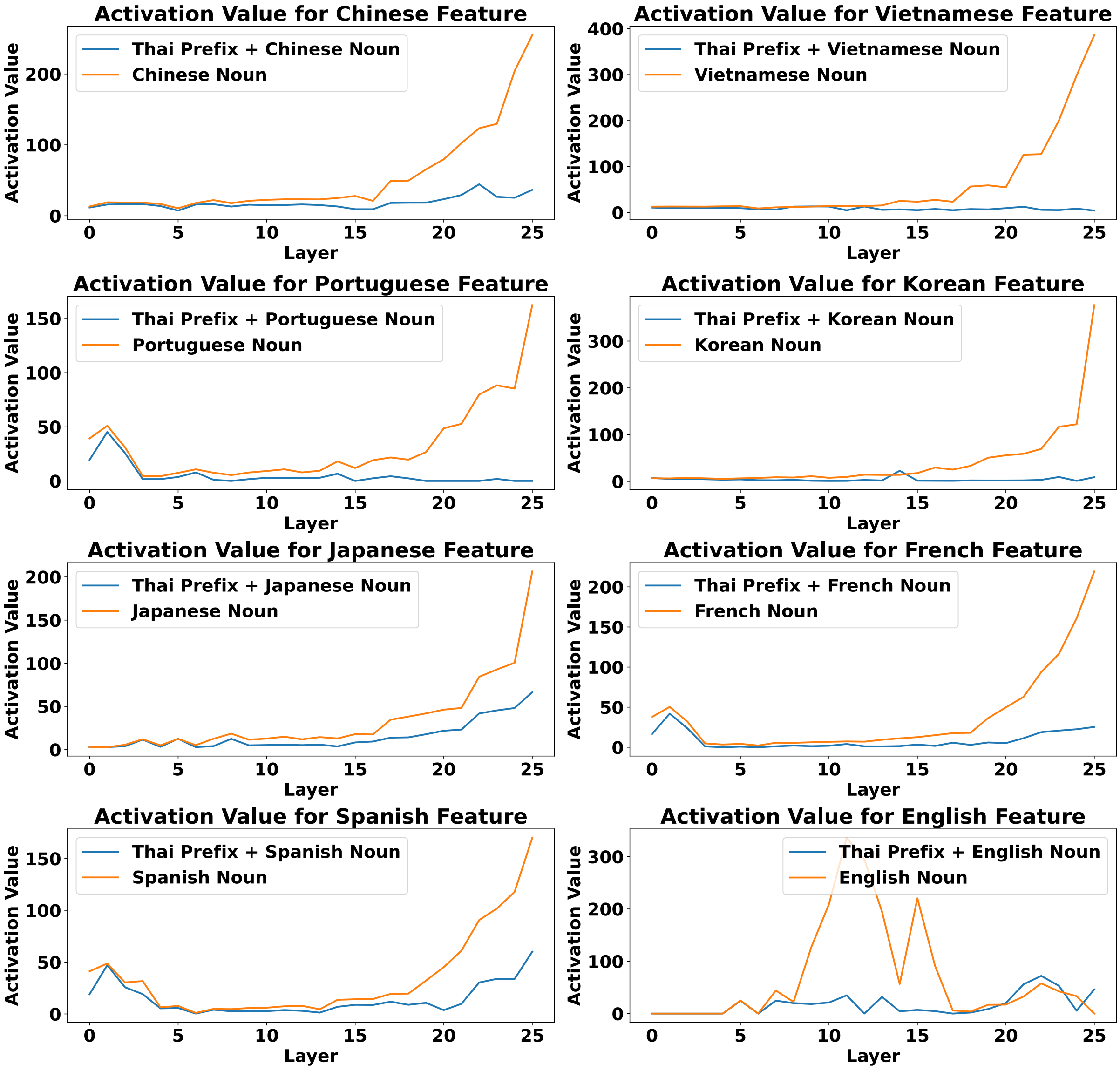


Figure 27: The mean activation values for various features with Thai prefix for Gemma 2 2B.

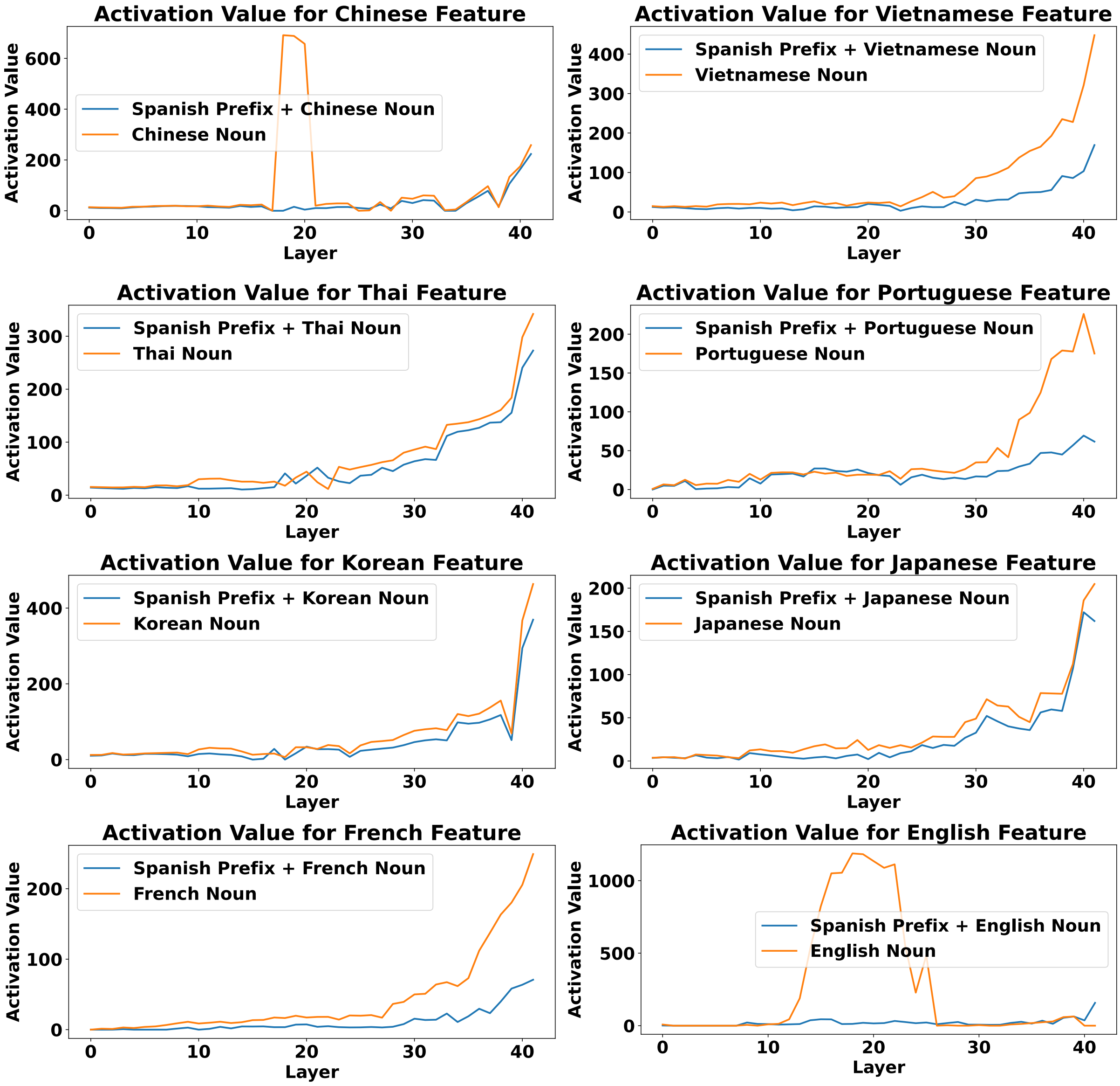


Figure 28: The mean activation values for various features with Spanish prefix for Gemma 2 9B.

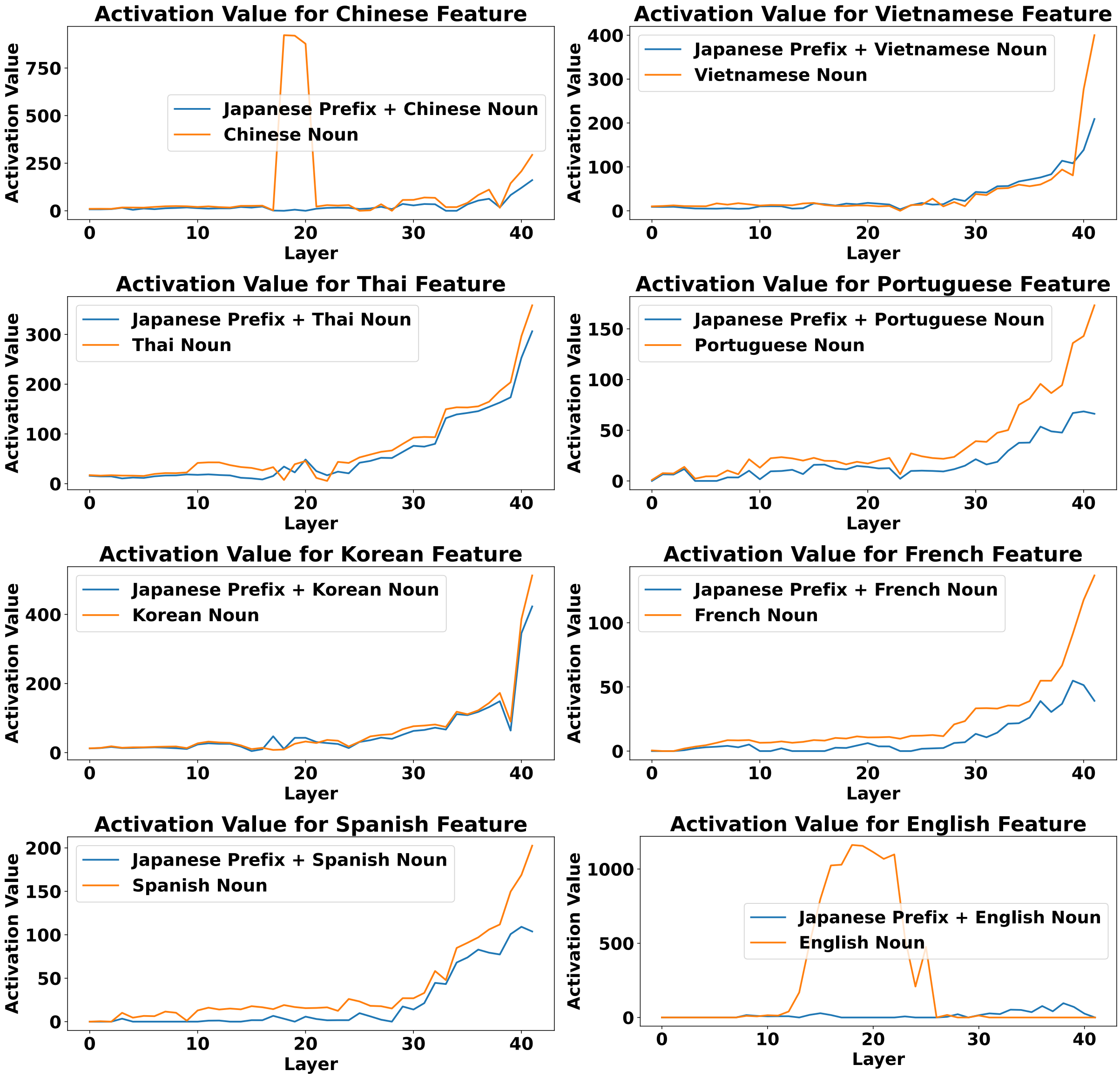


Figure 29: The mean activation values for various features with Japanese prefix for Gemma 2 9B.

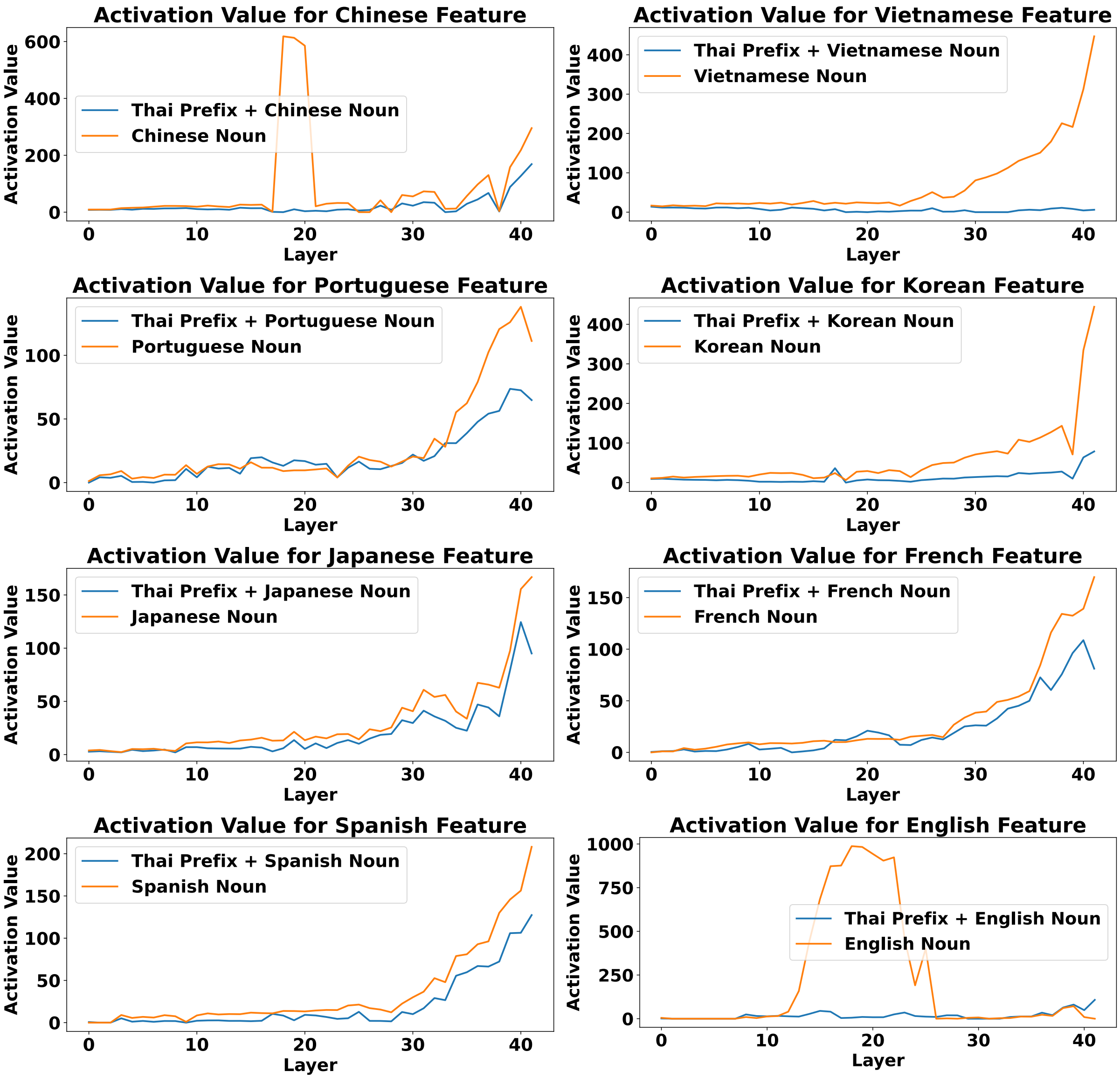


Figure 30: The mean activation values for various features with Thai prefix for Gemma 2 9B.

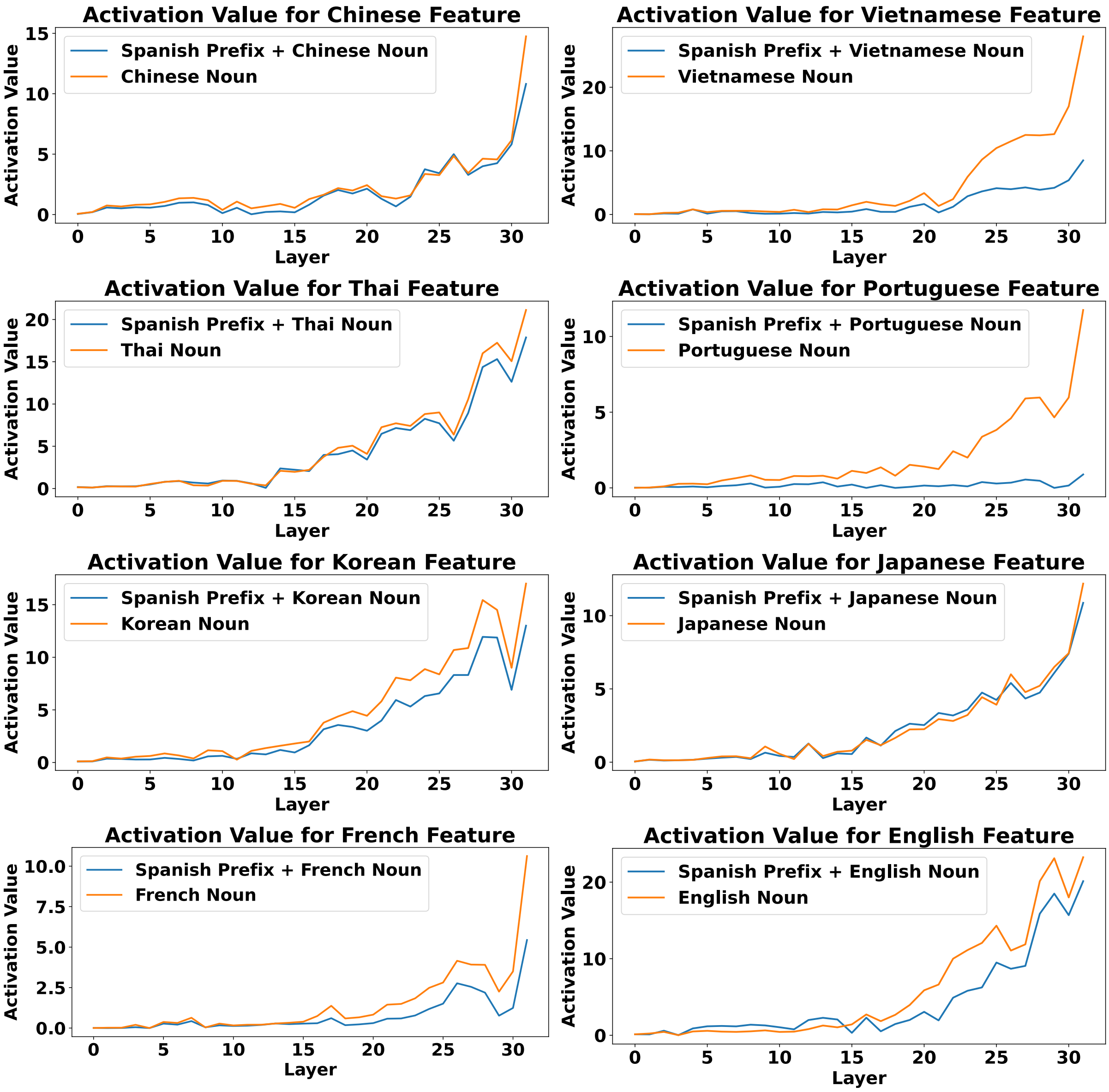


Figure 31: The mean activation values for various features with Spanish prefix for Llama-3.1-8B.

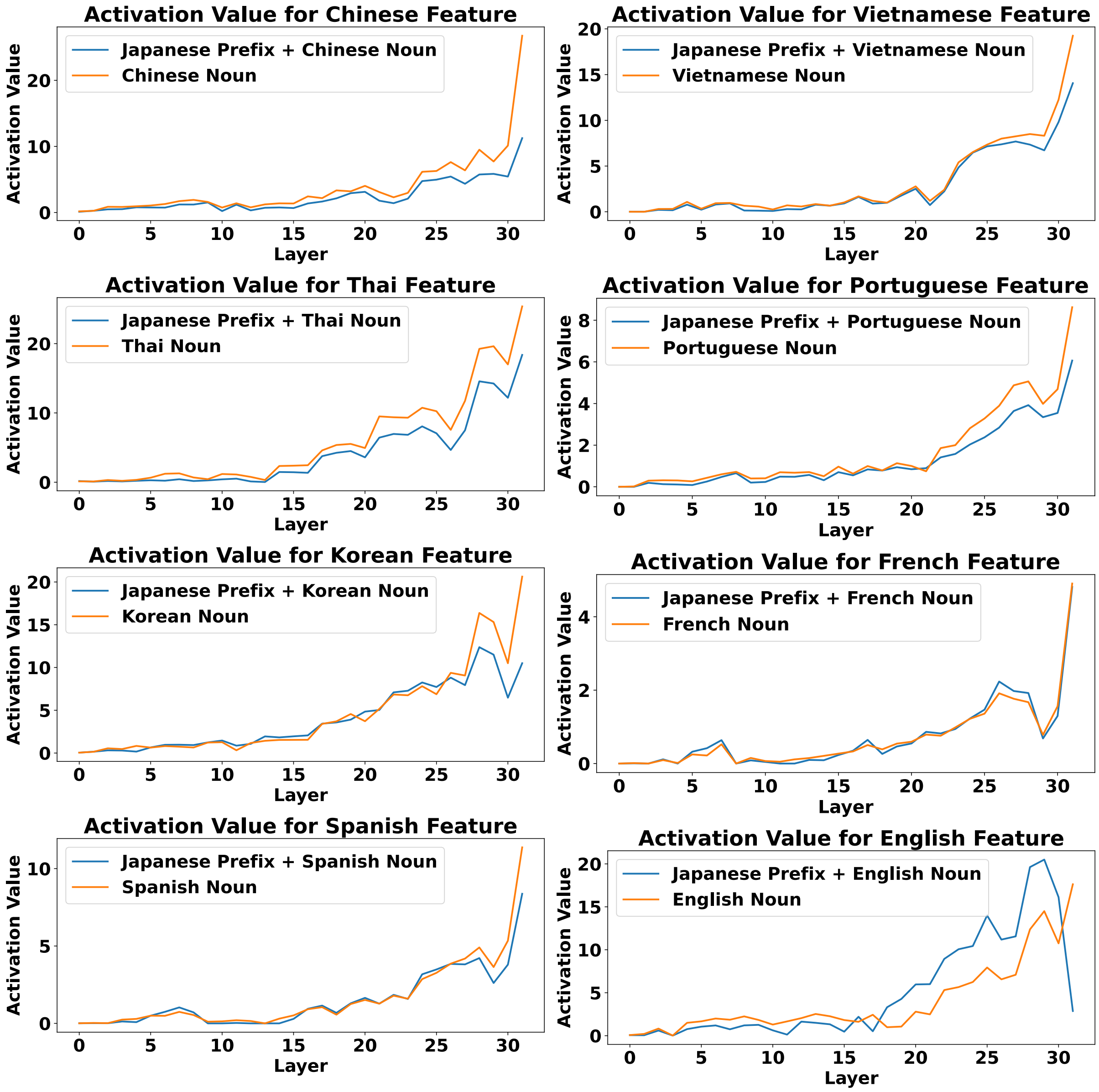


Figure 32: The mean activation values for various features with Japanese prefix for Llama-3.1-8B.

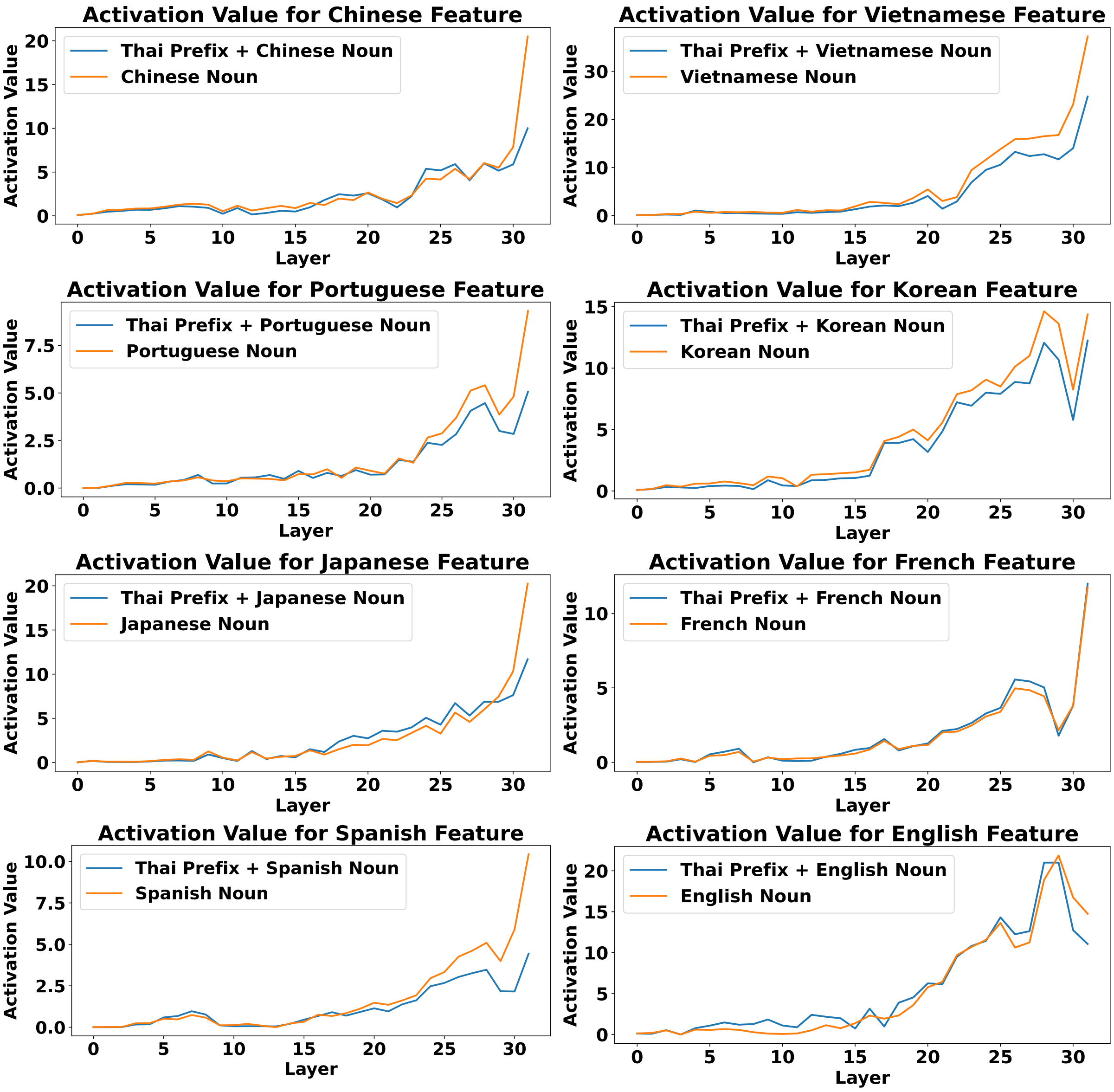


Figure 33: The mean activation values for various features with Thai prefix for Llama-3.1-8B.

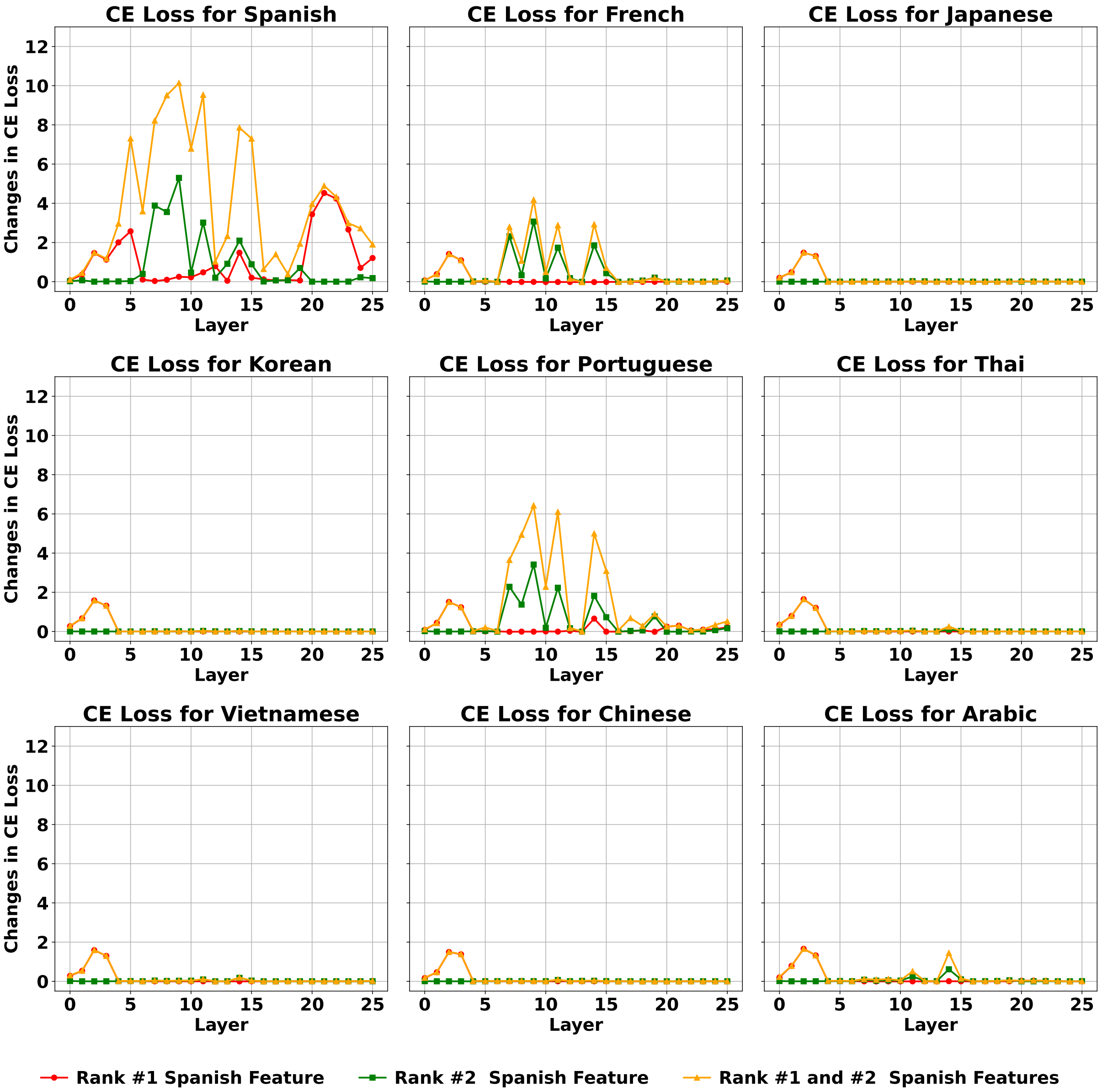


Figure 34: The change in CE loss for various languages after ablating Spanish features for Gemma 2 2B.

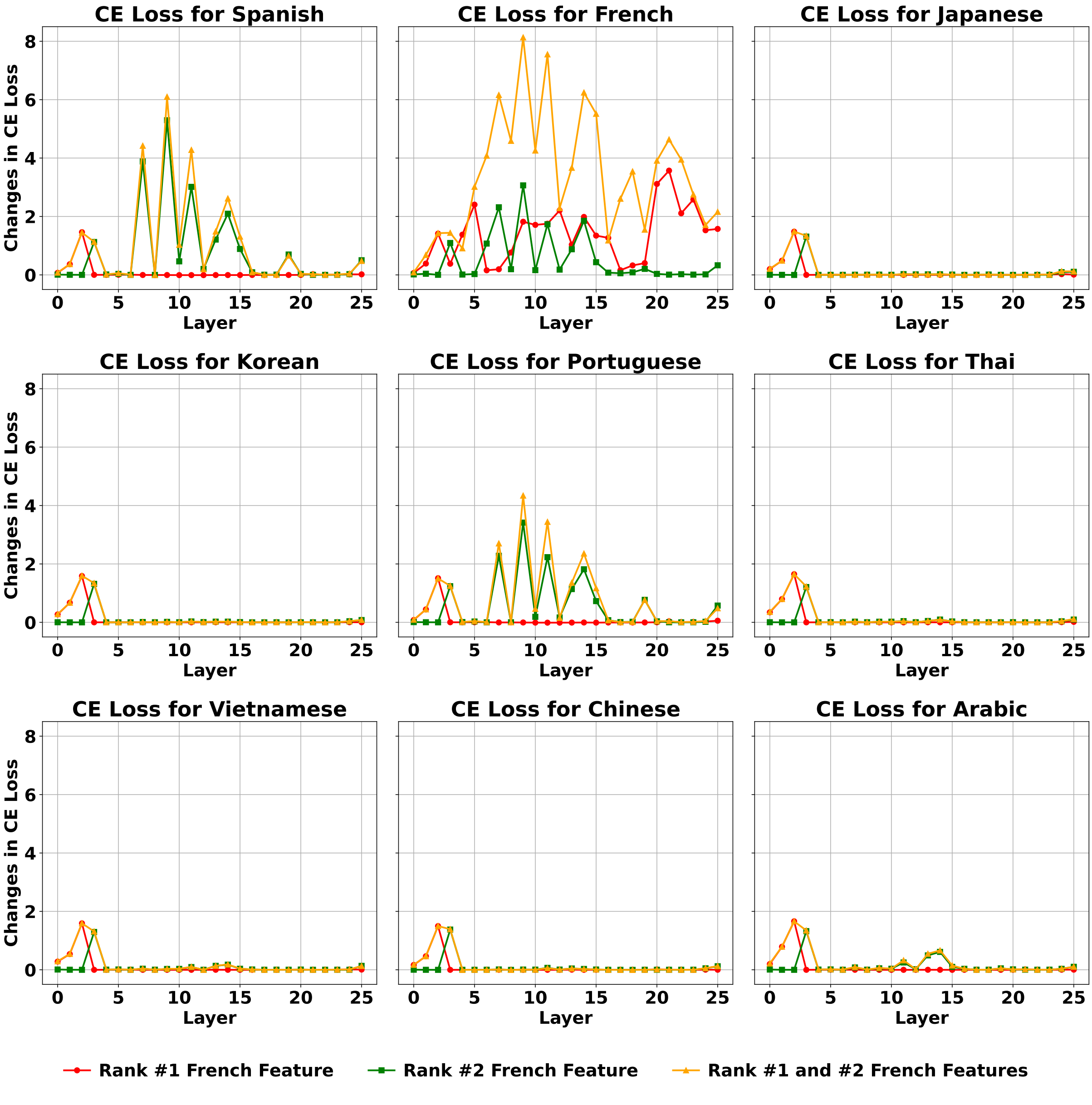


Figure 35: The change in CE loss for various languages after ablating French features for Gemma 2 2B.

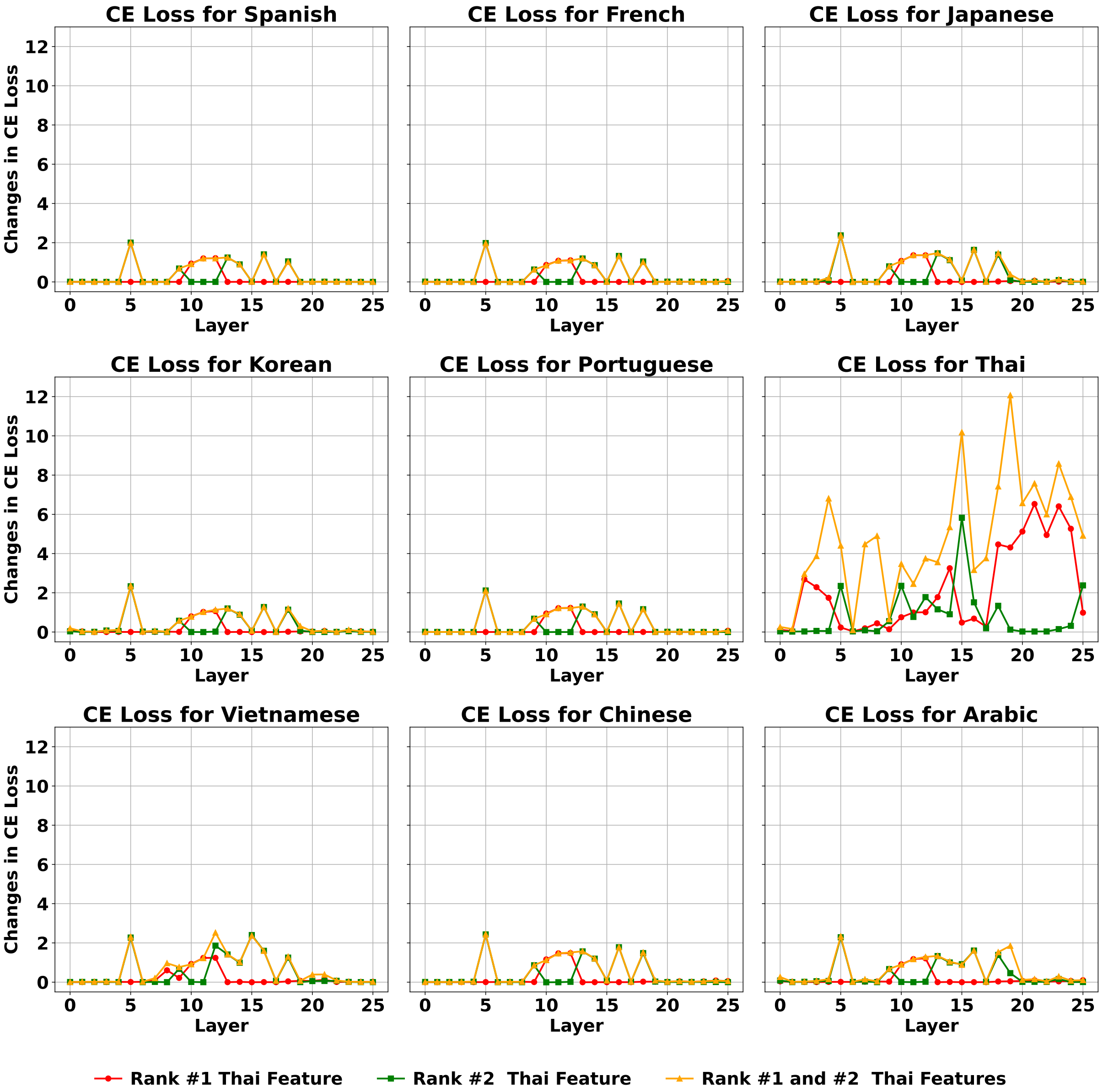


Figure 36: The change in CE loss for various languages after ablating Thai features for Gemma 2 2B.

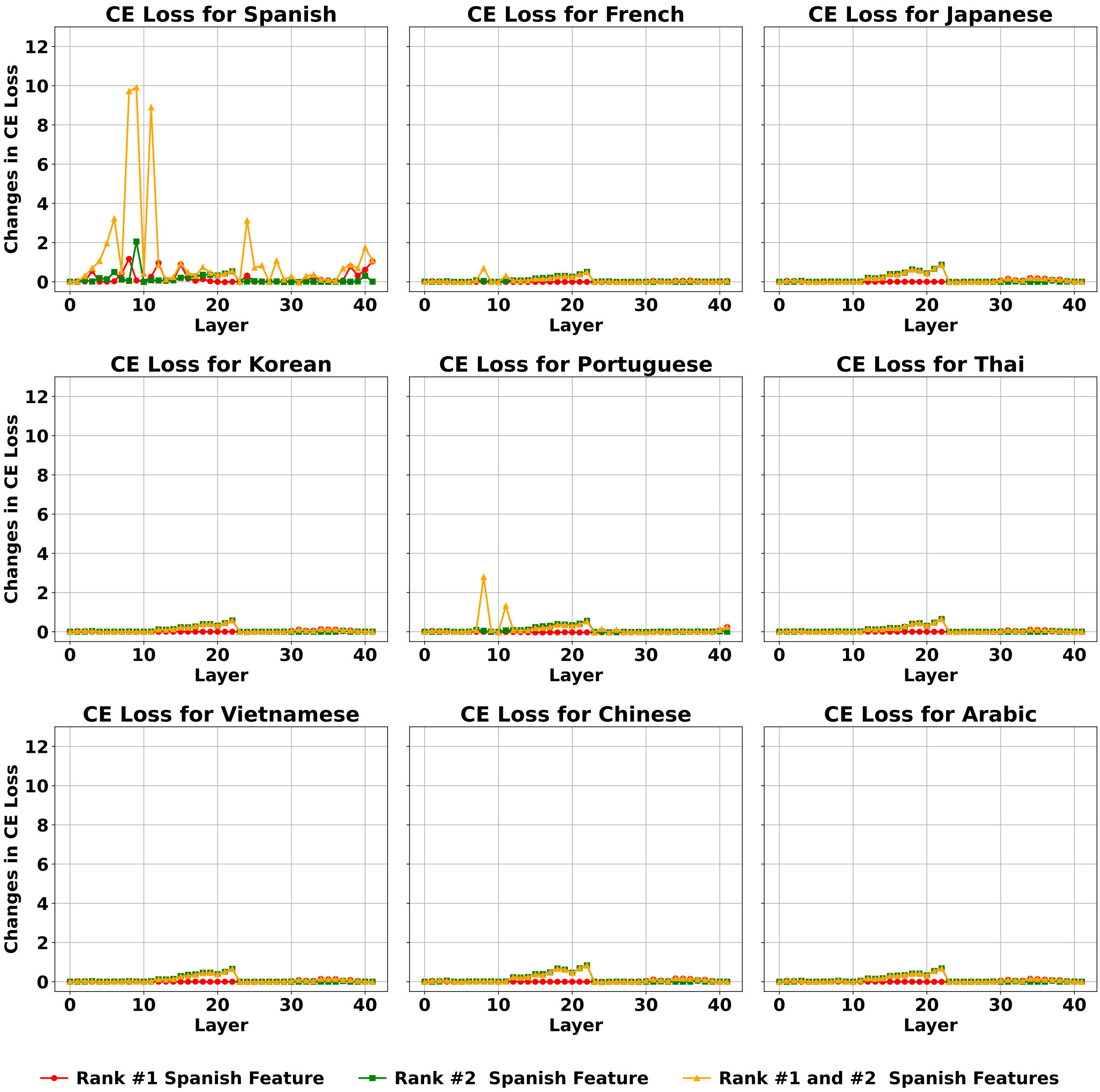


Figure 37: The change in CE loss for various languages after ablating Spanish features for Gemma 2 9B.

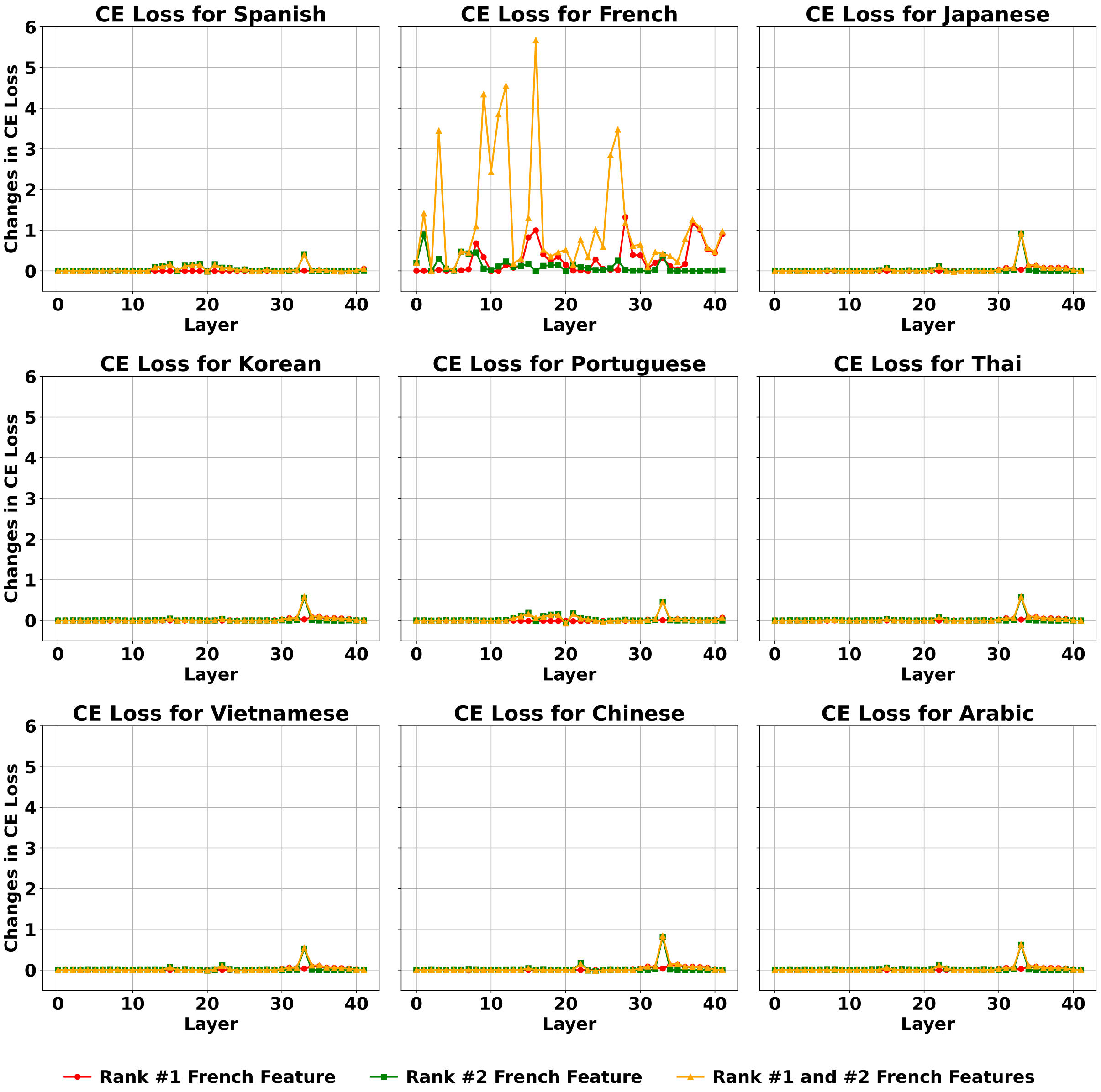


Figure 38: The change in CE loss for various languages after ablating French features for Gemma 2 9B.

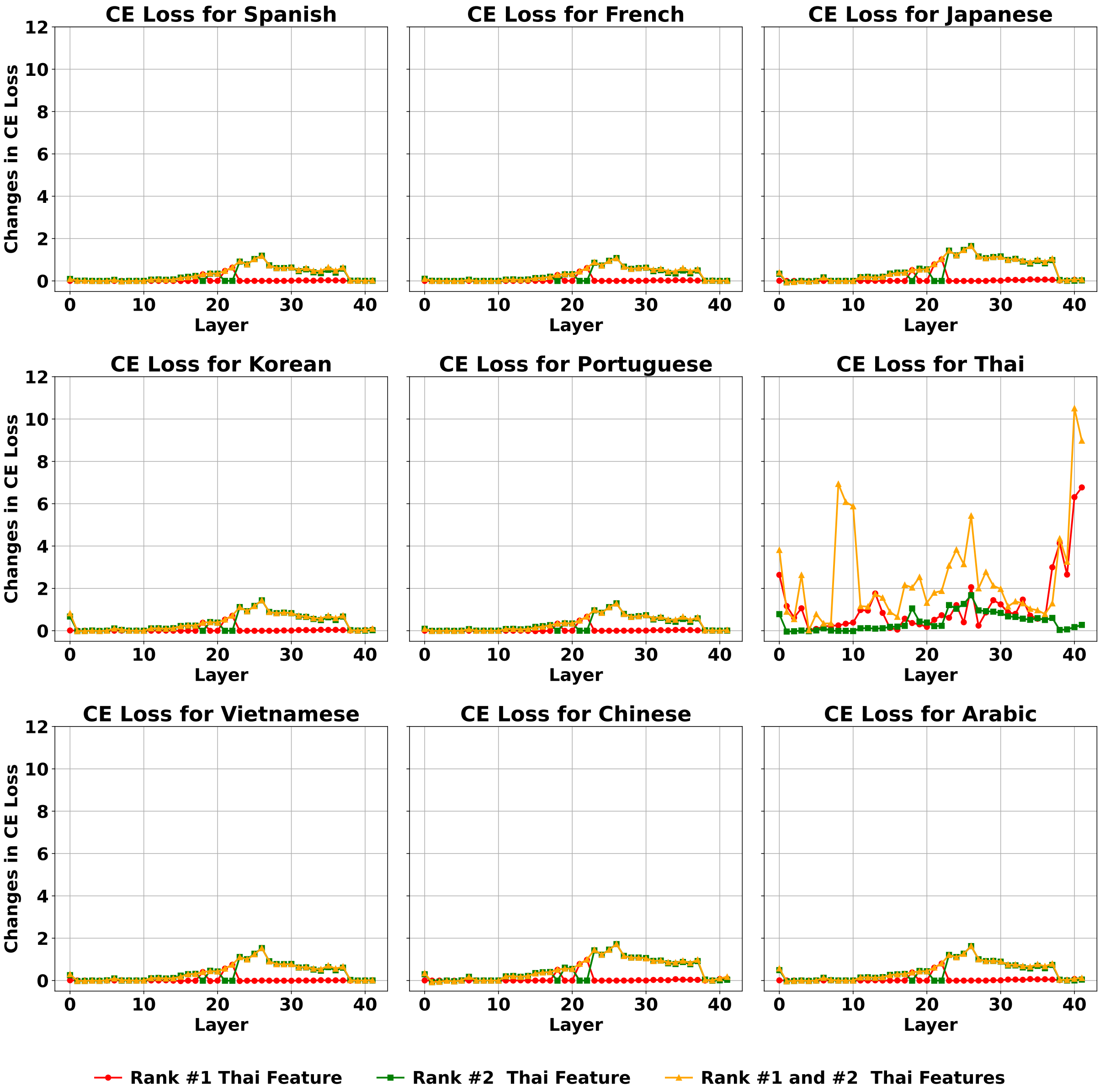


Figure 39: The change in CE loss for various languages after ablating Thai features for Gemma 2 9B.

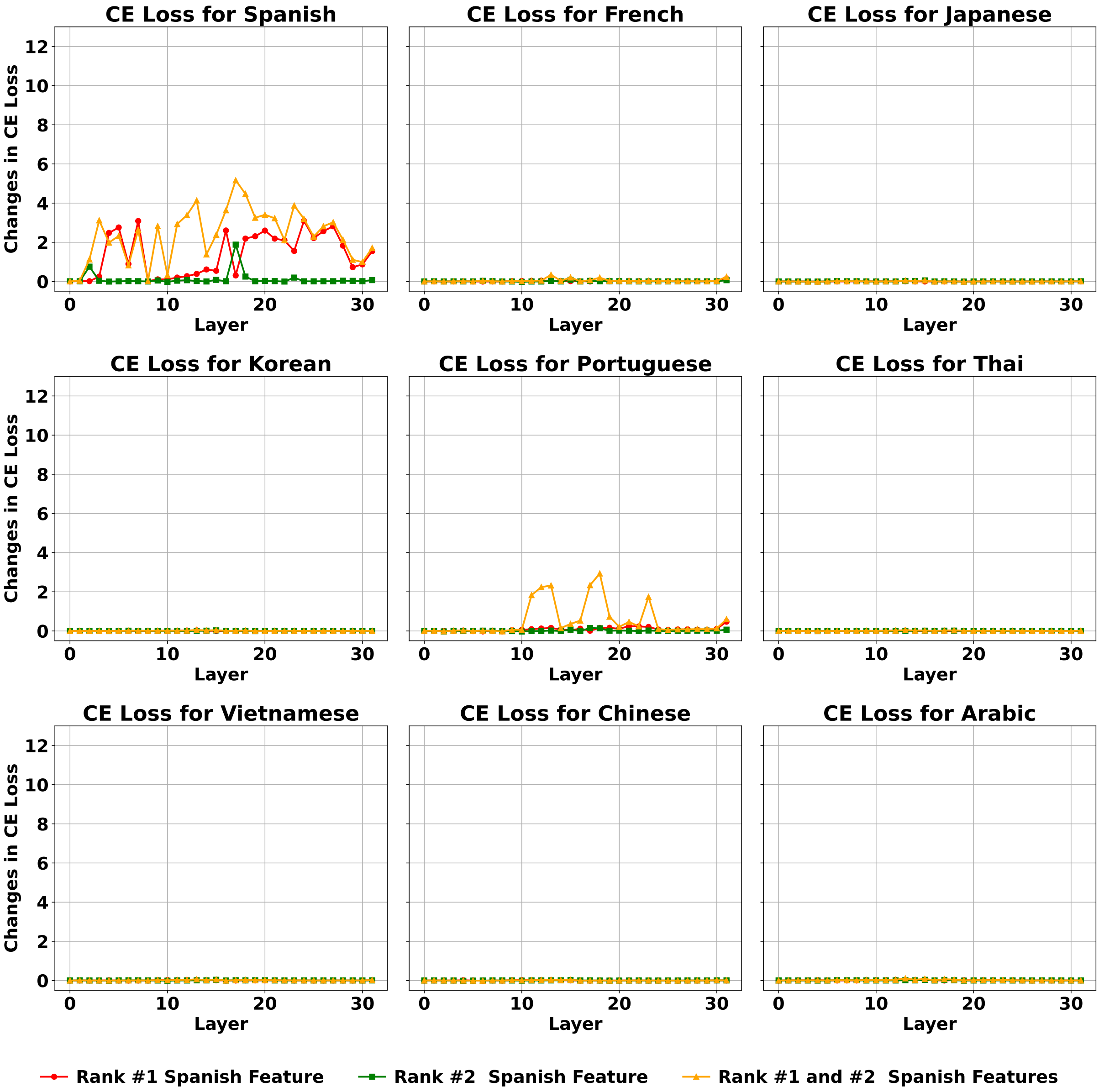


Figure 40: The change in CE loss for various languages after ablating Spanish features for Llama-3.1-8B.

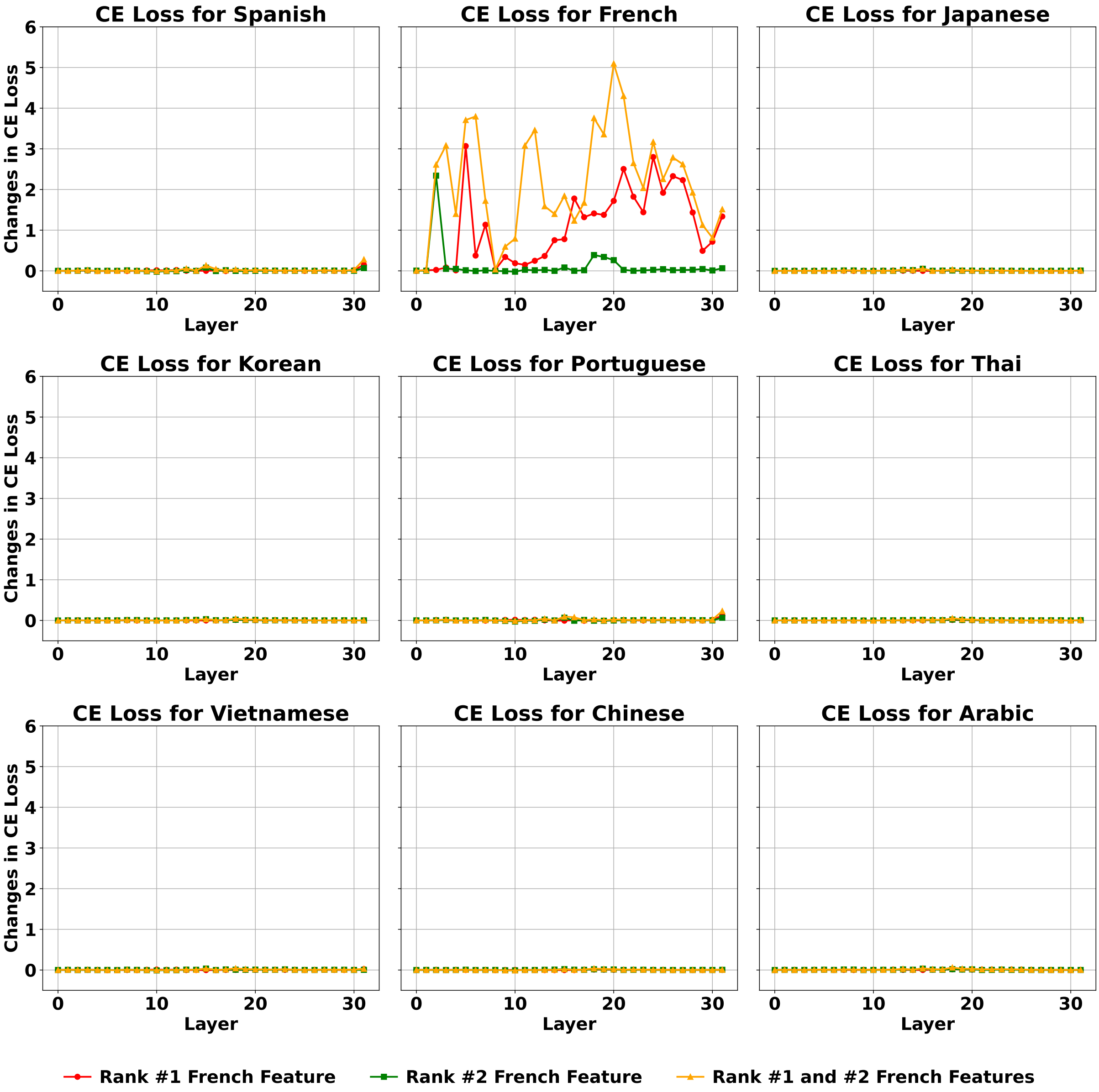


Figure 41: The change in CE loss for various languages after ablating French features for Llama-3.1-8B.

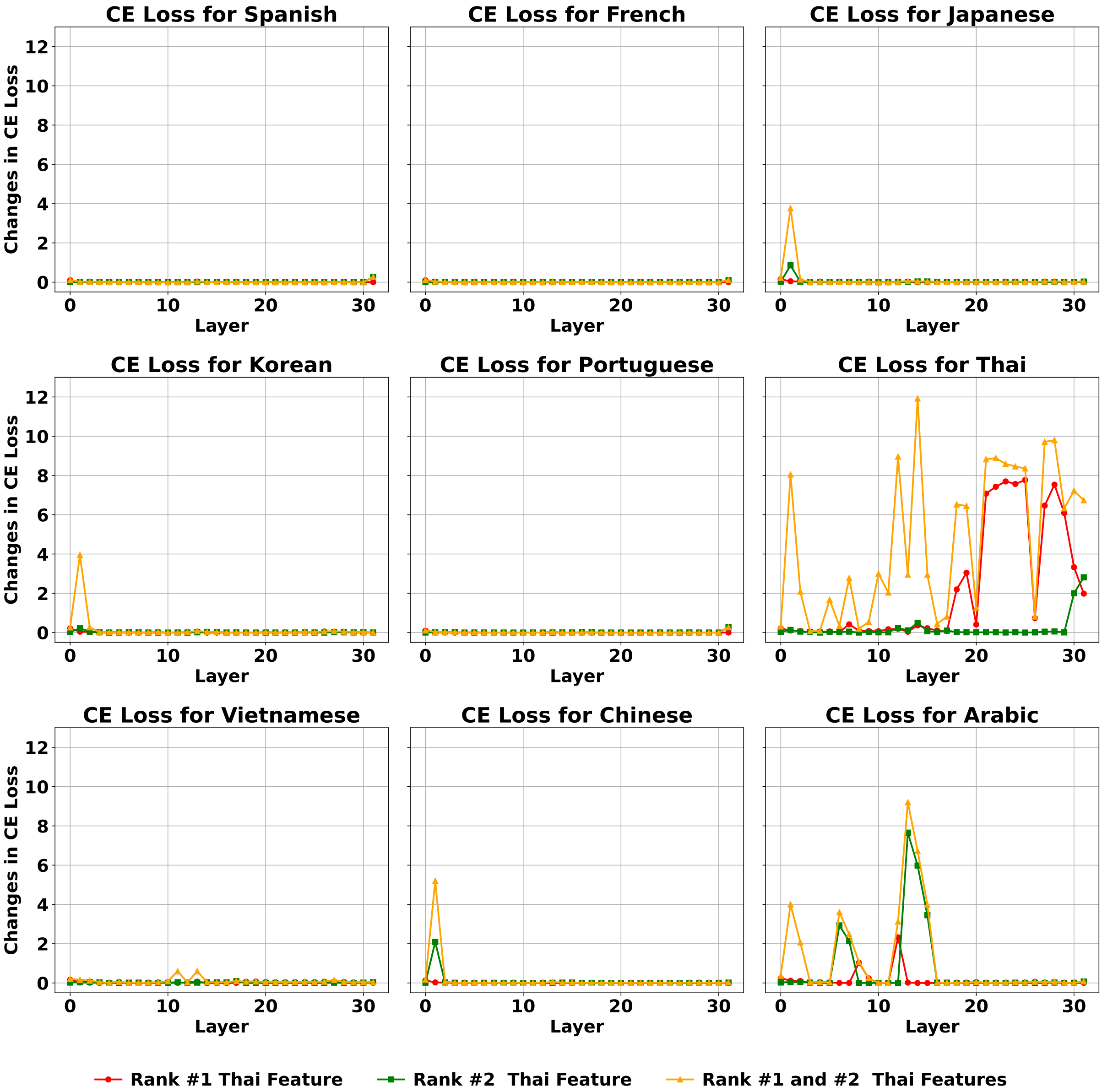


Figure 42: The change in CE loss for various languages after ablating Thai features for Llama-3.1-8B.