

Unveiling the Influence of Amplifying Language-Specific Neurons

Inaya Rahmanisa¹, Lyzander Marciano Andrylie¹, Mahardika Krisna Ihsani²,
Alfan Farizki Wicaksono¹, Haryo Akbarianto Wibowo², Alham Fikri Aji²

¹Faculty of Computer Science, Universitas Indonesia

²Department of Natural Language Processing, MBZUAI

{inaya.rahmanisa, lyzander.marciano}@ui.ac.id, alfan@cs.ui.ac.id

{mahardika.ihsani, haryo.wibowo, alham.fikri}@mbzuai.ac.ae

Abstract

Language-specific neurons, units in LLMs that strongly correlate with individual languages have been shown to influence model behavior by deactivating them. However, their role in amplification remains underexplored. This work investigates the effect of amplifying language-specific neurons through interventions across 18 languages, including low-resource ones, using three models primarily trained in different languages. We compare amplification factors by their effectiveness in steering to the target language using a proposed Language Steering Shift (LSS) evaluation score, then evaluate it on downstream tasks: commonsense reasoning (XCOPA, XWinograd), knowledge (Include), and translation (FLORES). The optimal amplification steering factors effectively steer output toward nearly all tested languages. Intervention using this factor on downstream tasks improves self-language performance in some cases but generally degrades cross-language results. These findings highlight the effect of language-specific neurons in multilingual behavior, where amplification can be beneficial especially for low-resource languages, but provides limited advantage for cross-lingual transfer.¹

1 Introduction

Multilingual large language models (LLMs) have shown remarkable capabilities. However, they still continue to exhibit inconsistencies in downstream tasks (i.e., often producing biased or inaccurate outputs across different linguistic contexts), especially involving low-resource languages (Dang et al., 2024). Understanding the root cause of these inconsistencies remains challenging due to limited transparency in how LLMs represent language-specific features and perform cross-lingual transfer.

¹The code and dataset are made available at <https://github.com/tauimbz/lang-task-neuron>.

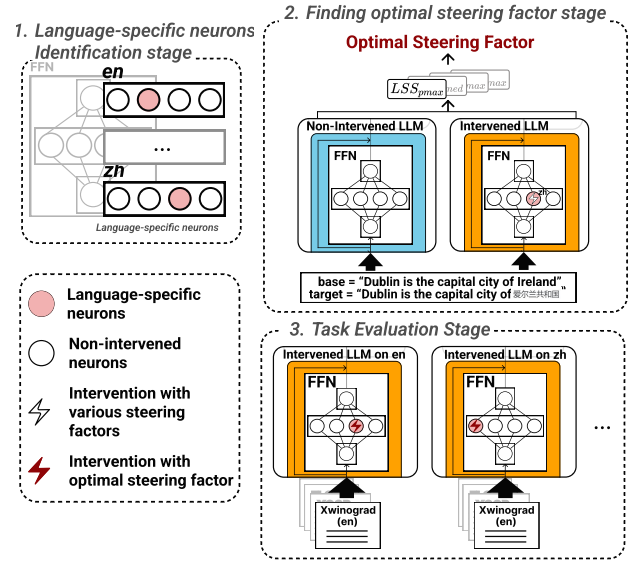


Figure 1: Illustration of our methodology. There are three sequential stages involving: 1). Identification of language-specific neurons, 2). Finding the optimal amplifying steering factor, and 3). Evaluation of the optimal amplifying steering factor on downstream tasks to understand its impact on the models' behavior.

In elucidating this issue, a few recent studies have begun to interpret model inner workings in multilingual contexts, revealing neuron-level correlations with language features and suggesting the existence of language-specific neurons (Kojima et al., 2024; Tang et al., 2024; Zhao et al., 2024). These neurons have been shown to affect multilingual capability following deactivation and signs of language-steering capabilities in generation. Interventions are typically done to *steer* outputs by replacing the neuron values with fixed activation values, called *steering factor*, to control behavior. Most prior works use zero as the steering factor to deactivate targeted neurons. However, we argue that amplifying these neurons can offer deeper insight into their functional capabilities, particularly regarding performance improvements or task and language transfer.

Building on the success of cross-lingual transfer

methods such as training and fine tuning (Conneau et al., 2020; Kim et al., 2019), we aim to investigate whether neuron overlap reflects shared linguistic representations. To this end, we extend existing studies that have been largely focusing on self-language interventions (Mondal et al., 2025; Wang et al., 2024) which showed limited gains, with cross-language interventions. We examine cross-lingual neuron behavior in closely related language pairs, distant language pairs, and low resource languages, to further explore whether neuron-level interventions can similarly provide insights on cross-lingual transfer, particularly revealing where and why such interventions succeed or fail.

This work is conducted by identifying and analyzing two types of neurons: language-specific neurons and language-activated (less-specific) neurons. We then compare various steering factors derived from different settings to evaluate their effect on the steerability of these neurons and analyze their impact on downstream task performance using neurons specific to the target language (self-intervention) and to every other languages (cross-intervention). We evaluate 18 languages including low-resource ones like Quechua, Haitian Creole, and Swahili, across models primarily trained in English and Chinese (Qwen), English (Gemma), and Southeast Asian languages (SeaLLMs) using XCOPA, XWinograd, and Include-lite for reasoning and knowledge tasks, and FLORES-200 for translation.

Several significant findings from this study are as follows:

- Comparison between different steering factors using the proposed Language Steering Shift (LSS) reveals that patched steering factors are the most effective and statistically outperform test-time intervention factors. Amplifying language-specific neurons with these factors consistently directs outputs toward all target languages, achieving over 90% average success in language shifting.
- Optimal amplification of language-specific neurons in self-intervention can reduce perplexity and improve task performance for certain languages, with gains reaching up to 1% in some models, showing improved results for low-resource languages compared to Mondal et al. (2025).
- Cross-lingual amplification can slightly reduce perplexity and enhance performance between closely related languages, but it gener-

ally leads to performance degradation in others, indicating limited effectiveness in cross-lingual transfer.

2 Related Work

Recent studies have explored neuron activations in language (Kojima et al., 2024; Tang et al., 2024; Zhao et al., 2024), task (Leng and Xiong, 2025; Song et al., 2024), and mixed contexts (Wang et al., 2024), often by inhibiting neurons to assess their roles. Kojima et al. (2024) showed that intervening on language-specific neurons can steer output language. Common intervention methods include activation patching (Zhang and Nanda, 2024), model editing (Meng et al., 2023), and suppression via zeroing or percentile thresholds (Mondal et al., 2025). While Mondal et al. (2025) has explored self-language intervention on downstream tasks, to the best of our knowledge, no study has yet focused on comparing amplification factors and provided improving insights into self- and cross-lingual interventions in downstream tasks.

3 Methodology

This section describes our method for identifying language-specific and language-activated neurons, introduces the Language Steering Shift (LSS) score to quantify language steering, and details our approach for determining the steering factors.

3.1 Neuron Identification

As illustrated in Figure 1, initially, we define neurons as the outputs of the activation function within the FFN modules in LLMs as they have been found to store knowledge or features (Dai et al., 2022; Tang et al., 2024).

Identifying Language-Activated Neurons. As a baseline method, we adapt a method from Wang et al. (2024) to define language-activated neurons, i.e., neurons whose activation value exceeds 0 throughout every sentence s in the dataset of language k , denoted S_k , allowing a single neuron to be language-activated for multiple languages. The set of language-activated neurons in layer i for language k , denoted by $A_k(i)$, is described by:

$$A_k(i) = \{j \mid \text{act}(i, j; s) > 0, \forall s \in S_k\}.$$

For neuron j , $\text{act}(i, j; s)$ is the activation value of the j -th neuron in the i -th layer when processing sentence s . These neurons are not further filtered,

resulting in relatively large number of neurons and overlapping activations across languages. We refer to these as *Baseline neurons*.

Language-Specific Neurons. *language activation probability entropy* (LAPE) method (Tang et al., 2024) is adopted to identify language-specific neurons. The probability of activation for the j -th neuron within the i -th layer when processing language k , denoted by $p_{i,j}^k$, is estimated according to:

$$\hat{p}_{i,j}^k = \mathbb{E} [\mathbb{I}(\text{act}(i, j) > 0) \mid \text{language } k \text{ tokens}] .$$

The expectation is taken over all tokens from sentences in language k , where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the condition is true and 0 otherwise; and $\text{act}(i, j)$ denotes the activation value of the j -th neuron in the i -th layer. Let $\mathbf{p}_{i,j} = [p_{i,j}^1, \dots, p_{i,j}^k, \dots, p_{i,j}^{|L|}]$ represent a list of activation probabilities for a set of languages L . These probabilities are then normalized to yield the L1-norm version, denoted by $p'_{i,j}$. Using this normalized version, the LAPE score of the j -th neuron in the i -th layer is formally defined as:

$$\text{LAPE}_{i,j} = - \sum_{k=1}^{|L|} p'_{i,j}^k \log(p'_{i,j}^k) .$$

We filter neurons whose language activation probability exceeds the m -th percentile. Finally, we select neurons that fall within the bottom $n\%$ of LAPE scores.

3.2 Language Steering Shift (LSS)

We propose **Language Steering Shift (LSS)** score to quantify the extent to which same-meaning answers are steered to the target language. Let L be the set of languages used. For every language $k \in L$, we evaluate the dataset for language k by performing: 1) a no intervention scenario, and 2) an intervention scenario where neurons corresponding to language l are intervened where $l \in L \setminus \{k\}$. The LSS score for evaluating language k when intervened upon by neurons specific to language l , denoted by $\text{LSS}(k, l)$, is calculated as follows:

$$\begin{aligned} \text{LSS}(k, l) &= \mathbb{E} [\mathbb{I}(\delta(k, l)|_{\text{int}} - \delta(k, l)|_{\text{non}} > 0)] , \\ \delta(k, l) &= \log p(\text{ans}_l) - \log p(\text{ans}_k) , \end{aligned}$$

where $\log p(\text{ans}_x)$ denotes the log-probability assigned by the model to the answer in language x ;

the symbols $|_{\text{non}}$ and $|_{\text{int}}$ represent the no intervention and intervention scenario, respectively; $\mathbb{E}[\cdot]$ is the expectation value across all sentences in the dataset. Intuitively, this metric measures the tendency of a model to produce the correct answer in target language k given an incomplete sentence written in source language l after intervening neurons corresponding to language k . For example, given the English sentence “Dublin is the capital city of Ireland”, the LSS score measures how much it shifts toward the Chinese translation, “Dublin is the capital city of 爱尔兰共和国”, when intervened using zh neurons. The underlined words represent the answers being analyzed.

3.3 Determining steering factors

We steer language-specific neurons in each layer with amplifying steering factors, which were determined in the identification stage (patched) and test-time intervention stage.

Patched Steering Factors. Let $v_{j,k}$ be the variable that contains the activation value of neuron j that is specific to language k . In the identification stage involving the multilingual FLORES-200 dataset, the steering factors are computed from the max and median of activation values across all sentences within the dataset S_k for language k . As a result, a steering factor is specific to neuron j and language k , to which we patch the language-specific neurons. This process is formally described as:

$$v_{j,k} \leftarrow \text{agg}\{s \in S_k \mid a_{j,s}\} ,$$

where $\text{agg} \in \{\text{max}, \text{median}\}$; and $a_{j,s}$ represents the activation value of neuron j averaged across all tokens in sentence s during the identification stage. We follow Kojima et al. (2024) in using the median and extend their approach by incorporating a stronger aggregation function, max. We refer to these steering factors as patch max (**pmax**) and median (**pmedian**).

Test-Time Intervention Steering Factors. Let $v_{j,k,m}$ be the variable containing the activation of neuron j for language k when processing sentence m , i.e., a sentence from the test dataset. We define the steering value as the maximum over activation values within the sentence. Using this value, we apply two operations to the language-specific neurons: addition (**+max**) to assign with a value surpassing its activation range, and replacement (**=max**) to assign the maximum activation. This computation is

formally represented as:

$$\text{steer}_{j,m} \leftarrow \max\{t \in T(m) | a_{j,m,t}\},$$

$$\text{=max} : v_{j,k,m} \leftarrow \text{steer}_{j,m},$$

$$\text{+max} : v_{j,k,m} \leftarrow v_{j,k,m} + \text{steer}_{j,m},$$

where $\text{steer}_{j,m}$ is the steering factor for neuron j during inference on sentence m ; $T(m)$ is the set of tokens in sentence m ; and $a_{j,m,t}$ is the activation value of neuron j associated with token t on sentence m . As a comparison to our amplifying factors, we also use deactivating factors 0 and 10p (10th percentile) as a replacement of max, which has been similarly demonstrated by Mondal et al. (2025) to deactivate neurons. We refer to these as =0 and =10p, respectively.

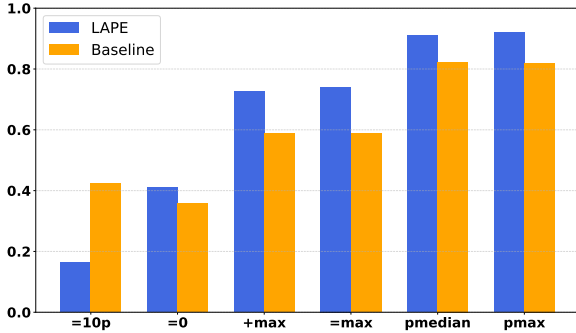


Figure 2: Comparison of averaged LSS scores of various steering factors across languages using LAPE and Baseline neurons.

4 Experimental Settings

This section discusses the experimental settings used throughout the methodology. It consists of three core stages. First, the identification of neurons specifically activated by language and those exhibiting language-specific responses. Second, the process of finding the optimal amplifying steering factor for influencing these neural activities. Finally, the evaluation stage to determine the impact of this optimal amplifying steering factor on various downstream tasks to assess its efficacy.

Identification of Language-Activated Neurons and Language-Specific Neurons. We identify language-activated (Baseline) and language-specific (LAPE) neurons by feeding the LLMs sentences from the dev split of FLORES-200 dataset (Team et al., 2022) containing 30,000 word-tokens per language. We assess 18 languages to ensure broader coverage of language specificity: en, nl, id, ms, vi, jp, zh, fr, pt, ru, et, ht, it, qu,

sw, ta, th, and tr in small and large models, namely Qwen2.5 Instruct (0.5B and 7B), Gemma2 Instruct(2B and 9B), and SeaLLMs3 Chat (1.5B and 7B), totaling in 6 models (see Appendix B for details). For identifying LAPE neurons, we use a language activation probability threshold (m) set at the 95th percentile and a LAPE score threshold (n) of 1%. All hyperparameter values are adopted from Tang et al. (2024) as they have been demonstrated to effectively capture language-specific neurons.

Finding the Optimal Steering Factor. We assess the determined steering factors on LSS score for the small models using a reconstructed version of MLAMA (Kassner et al., 2021), a multilingual fact-probing dataset. MLAMA is reconstructed such that each language question includes answers from all other languages and is filtered to ensure that no answer in different languages are the same. For example, if both Indonesian and Malaysian answers are “Jerman” (i.e., Germany in English), that row is discarded. The dataset contains 185 instances per language, with 15 of the 18 available languages are used due to limitation in the dataset.

Evaluating the Optimal Steering Factor on Downstream Tasks. We evaluate the model’s confidence and downstream task performances by conducting interventions using the optimal amplifying steering factors. We provide the deltas between the intervened and non-intervened performances to quantify the extent to which the interventions shift task outcomes. Interventions are performed in every layer of the model, with layer-wise amplification result and analysis provided in Appendix A.1.

We evaluate on commonsense reasoning task, which is considered one of the most difficult tasks for LLMs (Li et al., 2022), using XCOPA (Edoardo M. Ponti and Korhonen, 2020) and XWinograd (Muennighoff et al., 2022) dataset. For factual knowledge, we use Include-lite (Romanou et al., 2024) dataset. Building on Mondal et al. (2025), who focused only on self-language intervention, we extend the analysis of these task performances to cross-lingual intervention. The devtest split of FLORES-200 is used to measure perplexity and BLEU for translation from English (en) to every other languages.

Translation tasks are performed using 2 types of prompts: targeted and non-targeted. Targeted prompts explicitly mention the target language, whereas non-targeted prompts do not. For exam-

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
en	99.46	98.92	96.22	96.22	84.32	33.51	100.0	100.0	100.0	100.0	100.0	9.73	100.0	100.0	94.59	94.05	98.92	5.41
nl	100.0	100.0	92.97	90.81	67.57	32.97	100.0	100.0	100.0	100.0	68.65	15.14	100.0	100.0	93.51	89.73	61.08	16.22
ms	47.57	55.68	30.81	31.89	37.84	38.92	56.76	55.68	42.7	41.08	45.95	47.03	85.95	52.97	39.46	43.78	47.03	44.32
vi	100.0	100.0	80.0	75.14	23.78	18.92	100.0	100.0	96.76	96.22	14.05	10.81	100.0	100.0	87.57	76.76	14.59	8.65
jp	100.0	100.0	75.14	71.35	28.11	3.24	100.0	100.0	60.54	60.0	20.0	6.49	100.0	100.0	43.78	44.86	16.22	1.62
zh	100.0	100.0	48.11	43.78	14.05	5.41	100.0	100.0	87.03	81.08	33.51	9.73	100.0	100.0	56.76	56.76	49.19	1.08
fr	100.0	100.0	96.22	95.14	84.86	7.57	100.0	100.0	97.84	97.84	60.0	14.59	100.0	100.0	98.92	95.14	85.95	17.3
pt	99.46	100.0	96.76	94.05	72.43	9.73	100.0	100.0	98.92	98.92	77.3	9.19	100.0	100.0	96.76	91.35	82.7	4.32
ru	100.0	100.0	90.27	79.46	2.7	0.54	100.0	100.0	95.14	94.59	18.92	1.62	100.0	100.0	86.49	71.35	3.24	0.0
et	94.05	94.05	77.84	76.22	67.03	49.73	100.0	88.65	97.84	94.59	55.68	16.22	99.46	98.92	89.19	89.73	39.46	18.92
it	100.0	100.0	92.43	92.43	82.7	39.46	100.0	100.0	99.46	99.46	89.19	24.32	100.0	99.46	98.92	91.89	83.78	14.05
ta	100.0	100.0	1.08	1.08	0.0	0.0	100.0	98.38	50.81	45.95	0.0	0.0	100.0	100.0	1.08	1.08	0.0	0.0
th	100.0	100.0	41.62	41.62	0.54	0.54	100.0	100.0	65.41	64.32	1.08	0.0	100.0	100.0	25.41	22.7	0.54	0.0
tr	98.92	100.0	56.76	56.76	45.41	29.73	91.89	83.24	96.22	94.59	48.65	40.0	100.0	100.0	52.43	52.43	31.89	14.05
avg	89.30	89.91	65.08	63.06	40.76	18.02	89.91	88.40	79.24	77.91	42.20	13.66	92.36	90.09	64.32	61.44	40.97	9.73

Table 1: Language Steering Shift (LSS) scores from id to languages listed for amplifying (**pmax**, **pmedian**, **=max**, **+max**) and deactivating (**=0**, **=10p**) LAPE neurons across models. Bolded texts denote the highest values within a model.

ple, when translating to id, the targeted prompt begins with *"Translate from English to Indonesian."*, whereas the non-targeted prompt begins with *"Translate from English into the target language."*

5 Result and Analysis

This section presents the findings and analysis of the optimal amplifying steering factors and their evaluation on model confidence and downstream tasks.

5.1 LSS Scores and Optimal Amplifying Steering Factor

Result of the Language Steering Shift (LSS) score shows that LAPE neuron amplification yields high scores for many languages, often reaching 100%, as shown in Table 1 for the intervention on id. These strong results are consistent across other languages, as detailed in Appendix E, indicating successful language shifting. However, interventions involving certain language pairs may result in lower LSS scores. For example, translation between id and ms in either direction tends to produce scores capped at 50% in most models. This is likely due to the high similarity between the two languages, which makes it more difficult for the models to distinguish between them. Notably, en remains easy to steer to, suggesting a strong influence that

is possibly due to its dominance and widespread presence in LLM training data.

The optimal amplifying factors, defined as those that result in the highest aggregated LSS scores, are found to be **pmax** and **pmedian**. A one-way ANOVA followed by Tukey’s HSD test for post-hoc pairwise comparisons reveals that the patched steering factors **pmax** (mean = 0.9204) and **pmedian** (mean = 0.9095) significantly outperform the other factors ($p < 0.001$); Further results are provided in Appendix E. However, the difference between the two is not statistically significant ($p = 0.98$), indicating that both are equally effective as optimal amplifying factors.

Test-time intervention factors, **+max** and **=max**, perform well mainly for high-resource languages such as en, pt, it, and fr, but show limited effectiveness for other languages, with the lowest performance observed for ta in Qwen and Gemma despite its substantial neuron presence. The deactivating factor **=0** shows variation in results, with successful steering in some languages (e.g., en, nl, and it), while **=10p** demonstrates stronger deactivation capability, often resulting in LSS scores close to 0% across many languages, align with findings from Mondal et al. (2025).

Compared to Baseline neurons, LAPE neurons yield higher average LSS scores across languages as illustrated in Figure 2. This suggests that LAPE is more effective at steering language, likely due

to its stronger language specificity. In contrast, Baseline neurons show greater variability (see Appendix E); due to their lower specificity, deactivating these neurons with $=0$ and $=10p$ sometimes result in higher LSS scores than amplifying them. This contrast further supports LAPE neurons as more reliable and consistent for steering towards corresponding language.

Model	Include-lite		XWinograd		XCOPA	
	D	O	D	O	D	O
S-7B	0.38	-0.11	1.27	-6.71	-0.44	-5.49
S-1.5B	1.02	0.54	-0.76	-6.83	-0.18	-3.18
Q-7B	0.26	-0.17	-0.03	-7.19	-0.22	-5.98
Q-0.5B	0.16	-0.36	0.17	-3.38	-0.04	-1.69
G-9B	0.87	-0.77	0.29	-5.46	0.13	-2.24
G-2B	0.25	-1.03	0.15	-3.38	0.07	-1.43

Table 2: Average delta task accuracies after **pmax** intervention on LAPE neurons for self-intervention/diagonal (D) and cross-intervention/off-diagonal (O) neuron steering for every models used. Negative and positive values respectively represent decreased and increased accuracy from the baseline (non-intervention scenario). S denotes SeaLLMs3, Q = Qwen2.5 and G = Gemma2.

5.2 Impact of Neuron Amplification

nl	0.80	-0.80	-0.80	0.00	-0.40	0.80	-1.59	0.40	0.80	-1.59	-1.99	0.80	-1.19
id	3.23	-0.40	0.00	-0.40	-1.61	2.02	-1.21	0.81	1.21	1.62	0.00	0.00	0.41
ms	0.40	2.81	1.60	-2.81	0.80	-1.21	0.80	0.40	3.61	-2.81	0.80	-0.40	-1.61
vi	3.60	0.80	1.20	1.60	-0.80	0.40	1.20	0.00	2.00	0.40	0.80	0.40	2.00
jp	1.60	0.40	0.40	2.41	2.01	0.80	2.41	0.40	0.40	0.40	0.40	0.40	2.81
zh	-0.82	0.00	0.00	2.45	3.26	2.04	2.86	-0.82	0.41	0.00	-0.82	-0.82	0.81
fr	-1.20	-2.80	-3.60	-1.60	0.00	0.40	-1.20	-0.40	-0.40	0.40	-3.20	1.60	0.00
pt	0.00	-1.19	-1.19	1.18	-0.79	-2.38	0.79	0.79	-1.19	-0.40	-0.79	-0.40	-1.19
ru	0.79	0.79	0.79	0.79	1.19	1.98	-0.79	1.98	3.17	-0.79	-1.19	0.79	1.59
et	-10.4	-13.7	-14.8	-10.9	-8.20	-6.01	-16.9	-9.29	-1.09	0.55	-16.4	-3.83	-21.3
it	0.79	1.19	1.59	0.39	-0.80	0.39	-0.80	0.79	0.00	0.00	0.79	0.79	1.59
ta	0.00	0.00	1.60	0.00	0.00	-0.80	-1.60	-0.80	-2.00	-2.00	-2.80	-0.80	-2.80
tr	-1.21	0.40	0.00	0.80	0.80	0.00	-3.62	-1.21	0.40	-1.61	-1.21	0.00	0.00
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	it	ta	tr

Figure 3: Delta Include-lite accuracy under LAPE neurons on Gemma2 9B. Blue highlights reductions, whereas red highlights increases.

To further evaluate the optimal steering factors, we analyze their effectiveness in steering LAPE neurons using the most amplifying factor, **pmax**, across multiple downstream tasks and language modeling performance via perplexity. The next best factor, **pmedian**, shows comparable but less prominent results, while Baseline neuron amplification generally degrades performance in both self-

and cross-intervention settings (see Appendix B for further results).

en	0.04	-0.09	-10.16	-1.04	-11.70	-6.84
fr	-9.64	-1.21	-8.43	-10.84	-8.43	-15.66
jp	-10.95	-1.67	0.42	-0.83	2.61	-8.96
pt	-20.15	-4.56	-11.41	0.76	-6.85	-14.45
ru	-10.79	-2.54	-6.67	-0.95	0.63	-9.53
zh	-13.10	-1.39	-5.96	-0.80	-3.77	-0.80
en	-0.13	-0.56	-8.22	-0.86	-2.02	-5.04
fr	-3.62	-1.21	-4.82	2.41	-13.26	-8.44
jp	-5.32	-0.41	0.84	-3.33	1.36	-4.48
pt	-8.37	0.00	-6.85	1.14	-1.14	-9.51
ru	-7.31	-2.54	-4.77	-0.64	-1.59	-6.04
zh	-0.59	-0.79	0.40	-0.79	4.17	1.99
en	fr	jp	pt	ru	zh	

Figure 4: Delta XWinograd accuracy on Qwen2.5 7B (left) and Qwen2.5 0.5B (right) after amplification of LAPE neurons. Blue highlights reductions, whereas red highlights increases.

5.2.1 Knowledge and Reasoning Tasks

Overall performance of knowledge and reasoning tasks intervention as presented in Table 2 shows that **pmax** self-intervention can improve accuracy up to 1%, with consistent gains observed in Include-lite. XCOPA and XWinograd self-intervention may show less consistent improvements with only modest changes in some models. In contrast, cross-language interventions generally lead to performance drops on average. These results suggest that amplifying LAPE neurons in self-language on knowledge and reasoning tasks is potentially beneficial for enhancing performance, though their effectiveness in cross-lingual transfer appears limited.

Include-lite Evaluation. The result in Figure 3 reveals that self-intervention improves performance or modestly changes. As a comparison, cross-intervention shows stronger degradation, visibly captured in et. Some models capture increasing accuracy across similar languages such as id and ms, though the correlations are not always clear as they often show better performance than self-intervention.

XWinograd Evaluation. Figure 4 shows that self-intervention tends to be non-destructive, with some languages such as jp and pt, showing slight improvements across most models. However, the improved languages vary by models; for example, zh improves in Qwen2.5 0.5B but not in the 7B variant. Cross-lingual interventions degrade performance on average, particularly in larger models. Nevertheless, occasional improvements are observed in less related language pairs, such as zh

when intervened by ru in Qwen2.5 0.5B (see Appendix H for additional results). A consistent trend across models is that amplifying en neurons reliably reduces accuracy, suggesting they may play a role in reasoning rather than language transfer.

et	-0.40	0.20	-2.80	-2.20	-2.40	1.00	0.40	-0.80	-1.80	0.60	0.20
ht	0.40	0.60	0.20	1.60	1.40	0.80	0.40	0.80	2.00	-0.20	0.20
id	-13.6	-12.8	-1.80	-8.00	-13.0	-14.6	-13.8	-20.0	-12.8	-18.4	-22.2
it	-17.4	-13.0	-15.6	-1.00	-13.8	-14.2	-13.8	-16.8	-14.8	-18.4	-19.0
qu	0.60	1.40	1.00	1.80	0.00	0.20	-0.20	-0.40	-0.40	0.20	-0.40
sw	0.60	-1.20	-1.40	-2.20	0.80	1.00	1.80	0.40	-0.40	0.40	-1.40
ta	-3.20	-3.20	-1.60	-2.20	-3.60	-2.40	-0.60	-0.60	-2.60	-1.00	-3.00
th	-3.60	-4.00	-3.60	-2.80	-2.20	-4.60	-3.00	0.40	-2.60	-5.00	-5.00
tr	-5.40	-5.00	-5.20	-4.00	-5.00	-2.00	-4.40	-5.20	-1.00	-6.00	-6.00
vi	-15.2	-11.6	-14.4	-5.40	-12.6	-14.6	-14.4	-18.0	-14.0	-0.20	-20.6
zh	-5.80	-3.20	-2.20	-4.40	-4.40	-5.40	-7.00	-6.80	-7.60	-3.80	-1.80
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 5: Delta XCOPA accuracy after intervention of LAPE neurons for SeaLLMs3 7B (right). Row i depicts the initial language and column j is the intervention language.

XCOPA Evaluation. Figure 5 shows that amplifying LAPE neurons does not consistently improve in self-intervention, but changes are generally reserved, indicating non-destructive effects. In contrast, cross-language amplification often leads to larger drops, visibly captured on the intervention of it, id, th, tr, and vi neurons, degrading other languages more than their own. Low-resource languages like qu, et, sw, and ta show self-intervention gains, particularly in SeaLLMs3. SeaLLMs3 and Qwen2.5, having shared model architecture (Zhang et al., 2024), display very similar patterns, suggesting architecture may contribute to neuron behavior. Overall, XCOPA deltas reveal clearer signs of degradation than improvement.

5.2.2 Translation Task

Table 3 shows that BLEU scores increase after amplifying LAPE neurons with **pmax** for translating English to 17 other languages with non-targeted prompt. The results suggest that amplifying language-specific neurons enhances the model’s ability to generate text in the target language, including low-resource languages such as qu, ht, and sw, across most models. In contrast,

nl	-2.31	0.41	0.26	-0.02	-0.55	-0.52	0.30	0.26	-0.23	0.63	0.49	0.51	0.69	0.55	-0.40	-0.52	-0.02
id	-0.07	5.72	5.54	0.01	-0.65	-0.61	0.30	0.23	-0.29	0.53	0.42	0.51	0.58	0.45	-0.47	-0.60	-0.01
ms	-0.07	3.28	3.32	0.02	-0.61	-0.57	0.21	0.20	-0.26	0.42	0.43	0.44	0.57	0.46	-0.41	-0.56	-0.02
vi	-0.09	0.32	0.24	9.59	-0.52	-0.51	0.31	0.22	-0.23	0.39	0.34	0.48	0.53	0.42	-0.43	-0.51	-0.01
jp	-0.01	0.00	0.00	-0.01	0.09	0.01	0.00	0.01	-0.03	0.02	0.02	0.00	0.01	-0.01	-0.03	0.07	-0.00
zh	-0.05	0.01	0.01	-0.05	-0.01	-0.02	-0.02	-0.02	-0.07	0.04	0.05	0.03	0.06	0.03	0.01	0.38	-0.02
fr	-0.18	0.42	0.29	0.09	-0.51	-0.50	12.4	0.41	-0.24	0.57	0.76	0.53	0.64	0.50	-0.40	-0.50	0.01
pt	-0.11	0.41	0.33	0.03	-0.61	-0.59	9.45	-0.26	0.44	0.42	0.60	0.50	0.38	-0.49	-0.59	-0.00	-0.00
ru	-4.87	-4.78	-4.81	-4.89	-5.21	-5.18	-4.78	-4.75	-0.93	-4.73	-4.76	-4.76	-4.74	-4.76	-5.06	-5.17	-4.86
et	-0.01	0.21	0.08	-0.05	-0.48	-0.42	0.17	0.13	-0.17	0.40	0.17	0.20	0.18	0.19	-0.31	-0.41	-0.03
ht	-0.05	0.31	0.21	-0.01	-0.55	-0.53	0.27	0.27	-0.21	0.59	0.46	0.42	0.64	0.52	-0.40	-0.52	-0.06
it	-0.03	0.35	0.26	0.01	-0.63	-0.61	0.32	0.44	-0.29	0.43	0.25	1.74	0.49	0.41	-0.49	-0.61	-0.07
qu	-0.03	0.24	0.22	-0.02	-0.45	-0.39	0.19	0.21	-0.13	0.31	0.30	0.30	0.30	0.29	-0.22	-0.37	0.06
sw	-0.09	0.38	0.32	0.08	-0.51	-0.48	0.28	0.24	-0.19	0.33	0.34	0.38	0.40	0.35	-0.37	-0.47	0.03
ta	-0.02	0.01	0.04	-0.09	-0.39	-0.33	-0.00	0.05	-0.04	0.11	0.11	0.03	0.09	0.10	-0.17	-0.31	0.01
th	-0.00	0.05	0.01	-0.07	-0.08	-0.06	0.02	0.02	-0.08	0.07	0.07	0.10	0.06	0.00	0.78	-0.02	-0.02
tr	-0.00	0.21	0.13	-0.05	-0.47	-0.42	0.16	0.13	-0.17	0.28	0.24	0.19	0.30	0.21	-0.25	-0.40	0.71
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	ht	it	qu	sw	ta	th	tr

Figure 6: BLEU deltas for *non-targeted* **pmax** intervention in Qwen2.5 0.5B. Row i is the target language and column j is the intervention language. For example, row x col y represents BLEU score when translating English to language x intervened by y neurons. Blue highlights reductions, whereas red highlights increases.

the improvements from targeted prompts are less pronounced. Smaller models like Qwen2.5 0.5B and SeaLLMs3 1.5B tend to exhibit more consistent gains, which may indicate a higher degree of neuron locality in these models. The difference in performance between targeted and non-targeted prompts may suggest that amplifying these neurons improves the generation of the target language, though not always with high accuracy for translation. In contrast to amplification of Baseline neurons, which often degrades BLEU scores (Appendix G), LAPE demonstrates the ability to improve performance by targeting language-specific neurons.

To support the increasing BLEU findings, particularly in non-targeted prompt, we compare the performance with the cross-intervention setting presented in Figure 6. Self-intervention results in increasing BLEU scores compared to cross-intervention. Notably, using ru as the intervention language reduces performance in the cross-intervention case, while other languages show more modest increase compared to self-intervention. The observed improvements across languages may relate to linguistic similarity. For example, some increases occur when translating to id with ms as the intervention language, and vice versa; or when translating to pt with fr or it, and to fr with ht.

Lang	Targeted Prompt						Non-targeted Prompt					
	G-2B	G-9B	Q-0.5B	Q-7B	S-1.5B	S-7B	G-2B	G-9B	Q-0.5B	Q-7B	S-1.5B	S-7B
et	-0.44 (5.27)	+0.04 (5.06)	-0.04 (1.24)	-0.74 (4.09)	-0.15 (2.32)	-0.81 (3.76)	+0.07 (0.45)	-0.01 (0.24)	+0.40 (0.49)	+1.78 (0.76)	+1.01 (0.87)	+2.56 (0.43)
fr	-1.23 (26.49)	-4.19 (41.90)	+0.08 (13.93)	-0.16 (27.24)	+0.21 (18.23)	-1.36 (30.47)	+8.32 (0.76)	+3.97 (0.57)	+12.39 (0.52)	+14.40 (13.75)	+18.38 (1.43)	+24.69 (0.32)
ht	-0.94 (2.53)	-1.06 (2.21)	-0.08 (1.55)	-0.15 (3.35)	-0.19 (3.34)	-2.03 (5.12)	+0.01 (0.17)	-0.01 (0.13)	+0.78 (0.25)	+2.19 (0.35)	+1.53 (0.56)	+2.53 (0.43)
id	-0.53 (18.96)	-1.60 (17.89)	+0.49 (7.34)	+2.15 (21.29)	-0.64 (20.46)	+0.61 (30.02)	+1.09 (0.55)	+0.21 (0.29)	+5.72 (0.65)	+23.88 (1.13)	+14.07 (1.92)	+26.36 (0.52)
it	-0.67 (15.79)	-2.44 (25.30)	+0.03 (6.34)	-1.23 (15.90)	-1.04 (8.23)	-0.49 (17.42)	+3.66 (0.48)	+0.20 (0.25)	+1.74 (0.63)	+13.44 (1.16)	+9.62 (1.15)	+15.55 (0.44)
jp	+0.01 (0.11)	-0.38 (0.46)	+0.15 (0.05)	+0.20 (0.05)	+0.36 (0.14)	+0.06 (0.44)	+0.09 (0.02)	-0.02 (0.03)	+0.09 (0.06)	+0.06 (0.09)	+0.24 (0.11)	+0.06 (0.08)
ms	-0.38 (13.51)	-0.89 (35.14)	+0.08 (4.85)	-0.24 (16.13)	-0.25 (15.39)	-0.85 (24.82)	+0.69 (0.49)	+0.08 (0.31)	+3.32 (0.61)	-0.37 (1.01)	+9.05 (1.54)	+18.64 (0.50)
nl	+0.05 (13.08)	-2.72 (21.08)	+0.15 (4.52)	+1.01 (14.30)	+0.33 (8.57)	-0.66 (14.67)	+4.49 (0.45)	+0.06 (0.24)	+2.31 (0.55)	+9.84 (0.85)	+5.36 (1.20)	+11.01 (0.41)
pt	-2.42 (25.08)	-1.77 (44.80)	+0.41 (13.98)	-1.02 (36.48)	+0.39 (12.19)	+0.23 (30.99)	+10.11 (0.61)	+12.45 (0.44)	+9.45 (0.61)	+23.96 (1.33)	+15.32 (1.47)	+18.47 (0.58)
qu	-0.06 (0.34)	+0.39 (0.21)	+0.05 (0.46)	-0.12 (1.39)	-0.28 (1.05)	-0.30 (0.64)	+0.12 (0.44)	+0.16 (0.34)	+0.30 (0.46)	+0.38 (0.75)	-0.00 (0.93)	+0.11 (0.63)
ru	-0.33 (13.72)	+1.66 (16.16)	+0.26 (5.67)	-0.51 (16.88)	-0.99 (8.90)	-3.64 (17.02)	-0.02 (0.38)	+1.61 (0.24)	-0.93 (5.22)	+14.03 (2.20)	+9.62 (0.75)	+11.58 (4.77)
sw	-0.34 (7.62)	+2.71 (12.15)	-0.24 (0.58)	-0.48 (2.39)	-0.37 (1.56)	-0.06 (3.06)	+0.03 (0.45)	-0.05 (0.26)	+0.35 (0.51)	+1.01 (0.78)	-0.06 (1.07)	+1.59 (0.36)
ta	-0.66 (2.21)	-0.16 (5.58)	-0.01 (0.13)	-0.05 (0.28)	-0.05 (0.42)	-0.14 (0.59)	-0.01 (0.40)	-0.10 (0.23)	-0.17 (0.39)	-0.23 (0.45)	-0.21 (0.64)	+0.12 (0.29)
th	-0.45 (8.38)	-0.63 (19.93)	-0.49 (2.19)	-0.40 (8.08)	-0.10 (3.69)	-1.20 (6.61)	+0.06 (0.45)	+0.06 (0.28)	+0.71 (0.48)	+6.96 (0.77)	+2.36 (1.02)	+5.29 (0.43)
tr	-2.49 (22.69)	-5.95 (32.65)	+0.10 (10.71)	+4.73 (23.58)	+0.08 (18.12)	+0.89 (26.70)	+0.77 (0.33)	+4.42 (0.16)	+9.59 (0.52)	+25.26 (0.74)	+15.65 (1.65)	+21.40 (0.36)
vi	-0.35 (0.53)	-2.23 (2.89)	+0.10 (0.13)	-0.53 (1.53)	-0.17 (0.62)	-0.30 (0.56)	+0.00 (0.09)	+0.03 (0.09)	-0.02 (0.20)	+0.28 (0.19)	-0.10 (0.32)	+0.95 (0.24)

Table 3: BLEU score changes of self-language **pmax** intervention from en to target languages with Targeted and Non-targeted prompts on the FLORES-200 dataset. Each values are changes with increases shown in green and decreases in red. The subscripted values to the right of each number indicate the BLEU score without intervention. Q denotes Qwen2.5, G denotes Gemma2 and S denotes SeaLLMs-v3.

5.2.3 Perplexity Changes

Self-intervention of LAPE neurons using **pmax** can modestly reduce perplexity in most languages, showing success in improving language modeling performance as illustrated in Figure 7. Perplexity tends to increase more in cross-intervention settings than in self-intervention. However, the extent of the increase is not always related to the similarity of the language pair; For example, intervention between highly similar and same-family languages like id and ms does not lower perplexity, while less-related pairs like fr on en can. Interestingly, smaller models exhibit larger increases in perplexity, a pattern consistent with findings from the translation task intervention. This may suggest that smaller models contain more localized neurons responsible for language identity (see Appendix F for additional results).

The positive effects of perplexity reduction from LAPE neuron amplification are further supported by a comparison with Baseline neuron amplification. Intervention using Baseline neurons shows a similar pattern of increased perplexity in cross-intervention settings, although the increases are more pronounced. In self-intervention, amplification with Baseline neurons barely reduces perplexity, indicating no improvement in language modeling performance. These results highlight the advantage of LAPE neurons, whose specificity contributes to their ability to enhance language modeling performance. Detailed results for Baseline neuron interventions are provided in Appendix F.

5.3 Overlap Patterns in Language-Specific Neurons

As illustrated in Figure 8, although LAPE neurons are inherently tailored to individual languages, overlaps still exist across languages, albeit at significantly lower proportions. Compared with Baseline neurons with higher overlap proportions (Appendix D), the stronger specificity of LAPE neurons significantly improve self-intervention and capture patterns in cross-intervention for perplexity and task performance, indicating that better specificity leads to better performance for neuron-level intervention.

en	-0.00	0.02	0.02	0.02	0.19	2.74	2.61	-0.01	0.04	0.45	0.04	0.03	0.03	0.01	0.04	0.56	0.86	0.03
nl	-5.46	-0.07	0.08	0.08	1.36	8.68	8.16	3.53	2.58	3.14	0.12	0.22	0.27	0.33	0.06	1.80	2.98	0.62
id	-5.16	1.27	-0.02	0.01	2.70	9.60	8.05	3.41	2.41	4.74	0.25	0.23	0.19	0.31	0.07	3.09	4.62	1.19
ms	-3.96	1.05	0.00	-0.09	2.51	9.15	7.83	2.86	1.70	4.49	0.09	0.04	0.18	0.16	-0.03	3.07	4.56	0.97
vi	-6.83	0.93	0.08	0.06	-0.04	7.95	6.11	3.61	2.38	3.33	0.25	0.25	0.21	0.50	0.06	1.86	3.57	1.00
jp	9.43	2.00	0.34	0.29	0.75	0.01	1.40	6.02	5.38	2.01	0.62	2.09	3.53	1.57	0.81	0.06	0.30	0.35
zh	-8.79	0.54	0.18	0.16	0.54	2.59	0.14	3.46	2.72	1.46	0.27	0.92	1.58	0.87	0.34	0.14	0.37	0.33
fr	-6.20	0.43	0.06	0.06	0.85	9.65	9.29	0.00	0.65	4.36	0.13	0.10	0.09	0.15	0.06	2.51	3.81	0.40
pt	-5.96	1.10	0.09	0.09	0.85	9.25	8.69	2.73	-0.05	4.63	0.18	0.19	0.03	0.19	0.07	2.38	2.90	0.89
qu	-8.99	1.57	0.19	0.20	1.00	10.3	11.0	4.48	3.88	0.03	1.66	3.20	0.63	3.88	0.91	2.09	5.63	1.94
et	-3.99	1.04	0.33	0.31	2.22	8.84	8.18	3.05	2.24	4.04	-0.16	0.10	0.60	0.14	0.15	2.06	5.15	0.72
ht	-0.72	0.75	0.40	0.40	1.61	6.95	5.43	0.03	0.57	5.76	0.41	-0.09	0.37	0.07	0.36	4.23	4.78	0.63
it	-5.77	0.85	0.11	0.10	0.83	9.45	8.93	2.39	0.75	3.75	0.20	0.20	-0.02	0.25	0.08	2.91	3.75	1.12
ru	-0.31	0.44	0.27	0.33	1.13	6.32	4.74	0.44	0.34	5.03	0.10	0.03	0.22	-0.09	0.29	3.69	4.49	0.31
sw	-2.25	1.00	0.24	0.17	2.42	8.26	7.06	1.91	1.38	4.19	0.25	0.05	0.23	0.10	-0.09	2.91	5.11	0.91
ta	-8.53	5.98	1.03	0.83	3.22	9.49	7.76	8.39	8.16	4.96	3.81	5.90	5.60	6.04	3.28	0.05	1.88	4.71
th	-10.4	6.45	0.74	0.70	3.47	6.99	6.02	7.46	6.41	5.66	2.83	4.60	3.39	4.40	2.04	1.00	0.13	3.95
tr	-5.93	0.79	0.14	0.10	1.73	8.68	6.69	3.28	3.03	4.66	0.08	0.19	0.47	0.22	0.07	1.48	3.22	-0.07
	en	nl	id	ms	vi	jp	zh	fr	pt	ru	et	ht	it	qu	sw	ta	th	tr

Figure 7: PPL change with **pmax** steering factor under LAPE neurons on Gemma2 9B. Row i is the input language and column j is the language of intervention.

A strong overlap is captured in LAPE neurons across mutually intelligible languages like Indonesian (id) and Malaysian (ms), likely due to their linguistic similarity (Ahmad and Pustaka, 2011). Languages sharing writing systems, such as Chinese (zh) and Japanese (jp) (Taylor et al., 1995), also show overlap, consistent with Kojima et al. (2024). Romance languages, i.e., Italian (it), French (fr), and Portuguese (pt), share neurons. Vietnamese (vi) overlaps moderately with Thai (th) and slightly with id, ms, jp, and zh, likely due to regional or historical ties. Haitian Creole (ht) overlaps with French, reflecting shared roots (Prou et al., 2012). Surprisingly, low-resource languages like Quechua (qu), Haitian Creole (ht), and Swahili (sw) also show overlap, possibly indicating more general or shared neurons.

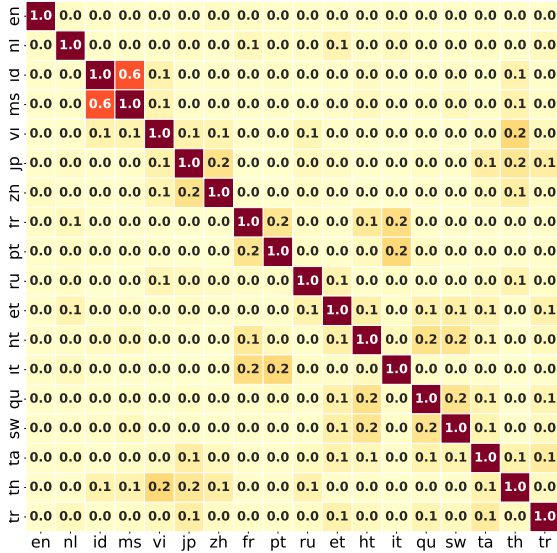


Figure 8: Symmetric heatmaps of overlap LAPE neuron proportions in SeaLLMs3 7B using jaccard distance.

6 Conclusion

This study proposes the Language Steering Shift (LSS) metric to compare steering factors of language-specific (LAPE) neuron intervention and identify the optimal one, then evaluates its impact in downstream tasks and language modeling. Optimal amplification steers outputs toward target languages with over 90% success on average. Self-language intervention modestly improves or maintains performance in language modeling measured by perplexity, translation, knowledge, and reasoning tasks. Cross-language intervention generally lowers performance, though gains can appear between closely related languages in perplexity and

BLEU scores for translation. Compared to less specific language-activated neurons, LAPE neurons yield clearer, more consistent improvements, suggesting that greater neuron specificity enhances multilingual performance when intervened. Overall, these findings suggest that amplifying LAPE neurons in self-language can enhance a model’s language capabilities, although not always improving task accuracy. Furthermore, the consistent degradation from cross-language interventions highlights their limited effectiveness for cross-lingual transfer.

Limitations

This study is limited to decoder-only models in 0.5B to 9B parameters and limited sets of tasks. Methods to identify language-specific neurons are also limited to LAPE, there are still gaps on the effect of amplification on neurons obtained by fine-grained attribution methods such as Integrated Gradients.

References

- I.H. Ahmad and Dewan Bahasa Pustaka. 2011. *Kesengangan leksikal bahasa Melayu Malaysia dan bahasa Indonesia*. Dewan Bahasa dan Pustaka.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. *Knowledge neurons in pretrained transformers*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.
- Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. 2024. *Explainable and interpretable multimodal large language models: A comprehensive survey*. *Preprint*, arXiv:2412.02104.
- Olga Majewska, Qianchu Liu, Ivan Vulic, Edoardo M. Ponti, Goran Glavač, and Anna Korhonen. 2020. *XCOPA: A multilingual dataset for causal common-sense reasoning*. *arXiv preprint*.
- Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. *Multilingual LAMA: investigating knowledge*

- in multilingual pretrained language models. *CoRR*, abs/2102.00894. To appear in EACL2021.
- Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. [Effective cross-lingual transfer of neural machine translation models without shared vocabularies](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.
- 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)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.
- Yongqi Leng and Deyi Xiong. 2025. [Towards understanding multi-task learning \(generalization\) of llms via detecting and exploring task-specific neurons](#). *Preprint*, arXiv:2407.06488.
- Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d’Autume, Phil Blunsom, and Aida Nematzadeh. 2022. [A systematic investigation of commonsense knowledge in large language models](#). *Preprint*, arXiv:2111.00607.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. [Locating and editing factual associations in gpt](#). *Preprint*, arXiv:2202.05262.
- Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. [Language-specific neurons do not facilitate cross-lingual transfer](#). In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. [Crosslingual generalization through multitask finetuning](#). *Preprint*, arXiv:2211.01786.
- M. Prou, A.K. Spears, and C.M.B. Joseph. 2012. *The Haitian Creole Language: History, Structure, Use, and Education*. Caribbean Studies. Bloomsbury Academic.
- 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, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. [Include: Evaluating multilingual language understanding with regional knowledge](#). *arXiv preprint arXiv:2411.19799*.
- Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024. [Does large language model contain task-specific neurons?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7113, Miami, Florida, USA. 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](#). *Preprint*, arXiv:2402.16438.
- I. Taylor, M.M. Taylor, and M.M. Taylor. 1995. *Writing and Literacy in Chinese, Korean and Japanese*. Studies in written language and literacy. John Benjamins Publishing Company.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024. [Sharing matters: Analysing neurons across languages and tasks in llms](#). *Preprint*, arXiv:2406.09265.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). *Preprint*, arXiv:2309.16042.
- Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. [Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages](#). *Preprint*, arXiv:2407.19672.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. [How do large language models handle multilingualism?](#) *Preprint*, arXiv:2402.18815.

A Ablation Study

A.1 Layer-Wise Amplification on LSS Score

We perform layer-wise amplification on Gemma2 2B model (26 layers) on LSS score using the reconstructed MLAMA dataset. Consistent throughout each language of intervention as captured by the average LSS scores across all languages in Figure 68, the highest average is found in the 19th layer, and the lowest is in the 10th layer.

Later layers are more likely to result in higher LSS score, especially from the 15th layer until just before the last layer. First few layers may also result in high LSS scores, though not as high. The middle layers result in relatively low LSS scores, often reaching 0 in most languages, with the exception of tr whose the lowest LSS is found in the 18th layer. We observe that some languages from the same family share similar layer-wise LSS scores, such as id, ms, vi and pt, it, fr.

This result aligns with findings from Tang et al. (2024) and Kojima et al. (2024) as language-specific neurons are mostly found in the first and last layers, excluding the very last one. Results for layer-wise amplification for each language are provided in Figure 69 (en), 71 (et), 72 (nl), 73 (id), 70 (ms), 74 (vi), 75 (jp), 76 (zh), 77 (fr), 78 (pt), 79 (ru), 80 (it), 81 (ta), 82 (th), 83 (tr).

A.2 Identifying Language-Specific Neurons from XWinograd to Evaluate XWinograd Performance Under Intervention

In this ablation study, we explore the behavior of amplification when the language-specific neurons are identified from the same dataset used for task evaluation. Instead of using FLORES-200 as in our original approach, we identify the language-specific neurons using the XWinograd dataset. Once identified, we amplify the neurons with **pmax** steering factor to evaluate the exact same XWinograd dataset from which the neurons were identified. The result in Figure 20 shows that there are no indication of improvement in the self-intervention nor cross-intervention in similar languages, leading to the conclusion that identifying and steering language-specific neurons from the same dataset may not better improve task performance in the respective languages. This result captures less visible patterns, which may be due to the fact that the XWinograd dataset contains significantly fewer tokens than FLORES-200. This scarcity makes LAPE less effective at identifying

language-specific neurons, and consequently, it struggles to improve task performance.

B Model and Language Details

B.1 Models and Language Neurons Used

We use models trained in different primary languages: Qwen2.5 Instruct (Qwen et al., 2025) 0.5B and 7B (primarily trained in Chinese and English), Gemma2 Instruct (Team et al., 2024) 2B and 9B (primarily trained in English), SeaLLMs-v3 Chat (Zhang et al., 2024) 1.5B and 7B (primarily trained in Southeast Asian languages). Qwen2.5 and SeaLLMs-v3 share the same model architecture, whereas Gemma2 has a different architecture. We employ neurons specific to 18 languages, English (en), Dutch (nl), Indonesian (id), Malay (ms), Vietnamese (vi), Japanese (jp), Chinese (zh), French (fr), Portuguese (pt), Russian (ru), Estonian (et), Haitian Creole (ht), Italian (it), Quechua (qu), Swahili (sw), Tamil (ta), Thai (th), and Turkish (tr). The selected languages account for different language domains associated with the models and represent both high and low-resource languages to further analyze their behavior.

B.2 Experimental Settings for Finding the Most Optimal Steering Factors

To find the most optimal steering factors, we reconstruct MLAMA dataset (Kassner et al., 2021) using 15 languages due to limitation in the dataset. The 15 languages are English (en), Dutch (nl), Indonesian (id), Malay (ms), Vietnamese (vi), Japanese (jp), Chinese (zh), French (fr), Portuguese (pt), Russian (ru), Estonian (et), Italian (it), Tamil (ta), Thai (th), and Turkish (tr). In the reconstructed MLAMA dataset, each question has 15 answers in different languages, with no lexical overlap.

We evaluate the proposed Language Steering Shift (LSS) scores on the reconstructed MLAMA dataset to measure how much an answer is shifted to the target language after intervened by the steering factors. This evaluation is only performed on the small versions of each model (Qwen2.5 0.5B, Gemma2 2B, and SeaLLMs-v3 1.5B) as they have already resulted in very high LSS scores, showing strong capability in steering to target languages.

B.3 Experimental Settings for Task Evaluation

We evaluate every models (6 in total) mainly using the **pmax** steering factor on several tasks. For translation task and perplexity, we use the devtest split of the FLORES-200 dataset on the same 18 languages the neurons were identified from. The translation evaluation is performed from English to every other 17 languages as English is considered the dominant language due to its widespread use as a multilingual benchmarks. Translation tasks are performed using 2 types of prompt: targeted and non-targeted. Targeted prompts explicitly mention the target language, whereas non-targeted prompts do not. For example, for translation to *id*, the targeted prompt is as follows,

```
"Translate from English to Indonesian.  
English: {question}  
Indonesian:"
```

whereas the non-targeted prompt is as follows,

```
"Translate from English into the  
target language.  
English: {question}  
Target language:"
```

For commonsense reasoning task, we evaluate XCOPA (Edoardo M. Ponti and Korhonen, 2020) and XWinograd (Muennighoff et al., 2022), each containing 2-option questions. XCOPA dataset aims to choose the more plausible cause or effect of a given premises with a focus on lower-resource languages including Estonian (*et*), Haitian Creole (*ht*), Indonesian (*id*), Italian (*it*), Quechua (*qu*), Swahili (*sw*), Tamil (*ta*), Thai (*th*), Turkish (*tr*), Vietnamese (*vi*), and Chinese (*zh*). XWinograd aims to resolve an ambiguous pronoun using commonsense knowledge, available in English (*en*), French (*fr*), Japanese (*jp*), Portuguese (*pt*), Russian (*ru*), and Chinese (*zh*).

For knowledge and reasoning task, we evaluate Include-lite dataset (Romanou et al., 2024), containing 4-option multiple-choice-questions (MCQ) extracted from academic and professional exams, covering 57 topics, including regional knowledge. The languages used for this dataset are limited to the available languages in the obtained language-specific neurons set, which are the following: Dutch (*nl*), Indonesian (*id*), Malay (*ms*), Vietnamese (*vi*), Japanese (*jp*), Chinese (*zh*), French (*fr*), Portuguese (*pt*), Russian (*ru*), Estonian (*et*), Italian (*it*), Tamil (*ta*), and Turkish (*tr*).

C Proportions of LAPE and Baseline neurons

Throughout all models, the proportions of LAPE neurons are presented in Table 27, Baseline neurons in Table 28. Proportions of LAPE neurons show a significant variability across languages and models. Some lower-resource languages (e.g. *qu*, *sw*, *ta*) show very high proportions for certain models, while others (e.g. *en*, *nl*, *ms*, *vi*) generally show lower proportions. This may suggest these languages need more dedicated neurons due to linguistic complexity or limited pre-training exposure. In contrast, *en* has the fewest neurons across models, consistent with Tang et al. (2024), likely because English’s dominance in pre-training leads to more distributed, broadly learned features.

D Neuron Overlaps

We calculate the Jaccard distance for neuron overlap between languages for LAPE neurons (Figure 10, 13, 14, 11, 12) and Baseline neurons (Figure 15, 18, 19, 16, 17).

E LSS Scores

The complete LSS scores for both Baseline and LAPE neurons are described in Table 5 to Table 18. The aggregated LSS scores for each steering factor is presented in Table 26. Further statistical testing to determine the significance of the optimal steering factor is provided in 19. Visualization of token log-probs for same-meaning answers in different languages are provided in both intervened and non-intervened LLMs in Figure 84 and 85.

F Perplexity Changes After Amplification

Perplexity changes for **pmax** intervention of LAPE neurons are provided in Figure 58, 59, 56, 57, 60, and 61; **pmedian** intervention are provided in Figure 64, 63, and 62. Perplexity changes for **pmax** intervention of Baseline neurons are provided in Figure 66, 65, and 67.

G BLEU Score Changes for Translation Tasks After Amplification

For the self-language intervention, BLEU score changes for **pmax** intervention of LAPE neurons with non-targeted prompt is provided in Table 21. Sample of BLEU delta heatmap (self- and cross-intervention) is provided in Figure 9. BLEU score

changes for **pmedian** intervention of LAPE neurons with targeted prompt is provided in Table 20. **pmax** intervention on Baseline neurons with targeted prompt is provided in Table 22. Samples of translation output for both targeted and non-targeted prompts are presented in Figure 86 and 87.

H Delta Accuracies of Reasoning and Knowledge Tasks After Amplification

Table 23 reveals the baseline (no intervention) XWinograd accuracy. Delta accuracies of XWinograd dataset under LAPE neurons when steered with **pmax** are provided in Figure 21, 22 and 23; intervention results for **pmedian** are provided in Figure 24, 25, and 26. Delta XWinograd accuracies of **pmax** intervention on Baseline neurons are provided in Figure 29, 30, and 31.

Table 24 reveals the baseline (no intervention) XCOPA accuracy. Delta accuracies of XCOPA dataset under LAPE neurons when steered with **pmax** are shown in Figure 32, 33, 34, and 36; intervention results for **pmedian** are shown in Figure 39, 38, and 40. Delta XCOPA accuracies of **pmax** intervention on Baseline neurons are provided in Figure 41, 42, and 43.

Table 25 reveals the baseline (no intervention) Include-lite accuracy. Delta accuracies of Include-lite dataset under LAPE neurons when steered with **pmax** are shown in Figure 47, 46, 45, 44, 49, and 48, intervention results for **pmedian** are provided in Figure 51, 50, and 52. Delta Include-lite accuracies of **pmax** intervention on Baseline neurons are provided in Figure 54, 53, and 55.

nl	-0.33	-3.92	-3.94	-4.26	-4.87	-4.87	-3.89	-4.06	-4.53	-4.18	-2.12	-3.79	-3.01	-2.79	-4.78	-4.86	-4.35
id	-6.73	0.75	0.47	-7.21	-8.01	-8.00	-7.11	-7.01	-7.59	-6.91	-4.73	-6.80	-5.71	-4.40	-7.90	-8.00	-7.46
ms	-4.01	0.36	0.20	-4.37	-5.08	-5.07	-4.26	-4.22	-4.68	-4.16	-2.68	-3.96	-3.38	-2.63	-4.89	-5.05	-4.50
vi	-11.8	-11.9	-12.0	-0.21	-12.8	-12.8	-11.9	-11.9	-12.4	-12.2	-10.8	-11.6	-11.9	-12.2	-12.7	-12.8	-12.3
jp	-0.01	0.03	0.01	-0.01	0.29	0.21	0.03	0.01	-0.01	-0.01	0.00	-0.00	-0.03	-0.04	-0.01	0.09	0.01
zh	-0.10	0.06	0.08	0.05	0.55	0.17	-0.01	0.03	0.04	0.16	0.02	0.01	0.12	0.03	0.15	0.64	0.06
fr	-13.1	-13.1	-13.1	-13.5	-14.1	-14.1	0.27	-13.0	-13.8	-13.4	-10.5	-12.9	-12.3	-13.2	-14.0	-14.1	-13.6
pt	-12.7	-12.6	-12.6	-13.0	-13.8	-13.8	-12.5	0.30	-13.4	-12.7	-11.7	-12.1	-12.5	-12.9	-13.7	-13.8	-13.1
ru	-6.64	-6.57	-6.55	-6.75	-7.08	-7.08	-6.44	-6.70	-0.01	-6.66	-6.62	-6.38	-6.74	-6.74	-6.98	-7.06	-6.75
et	-0.70	-0.60	-0.60	-0.84	-1.40	-1.40	-0.65	-0.51	-1.11	-0.18	-0.19	-0.15	-0.55	-0.76	-1.24	-1.37	-0.91
ht	-0.03	0.07	0.16	-0.16	-0.78	-0.78	-0.00	-0.07	-0.40	0.07	-0.36	0.13	-0.45	-0.49	-0.62	-0.76	-0.21
it	-5.61	-5.49	-5.56	-5.94	-6.63	-6.63	-5.63	-5.35	-6.20	-5.75	-4.69	-0.04	-5.40	-5.73	-6.56	-6.63	-6.06
qu	-0.43	0.32	0.43	0.15	-0.33	-0.32	0.31	0.13	0.05	0.39	-0.03	0.45	0.02	-0.19	-0.16	-0.27	0.26
sw	-0.31	0.49	0.38	0.13	-0.54	-0.54	0.07	0.05	-0.24	0.15	-0.06	0.30	-0.15	-0.29	-0.50	-0.52	-0.05
ta	-0.43	0.40	0.44	0.25	-0.08	-0.08	0.46	0.27	0.35	0.45	0.02	0.10	0.16	0.25	-0.01	-0.05	0.41
th	-1.19	-1.18	-1.16	-1.28	-1.30	-1.29	-0.16	-1.23	-1.28	-1.28	-1.26	-0.81	-1.35	-1.22	-1.18	-0.23	-1.23
tr	-1.41	-1.36	-1.36	-1.70	-2.18	-2.18	-1.47	-0.99	-1.75	-1.48	-0.30	-1.05	-0.75	-1.26	-1.82	-2.15	-0.38
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	ht	it	qu	sw	ta	th	tr

Figure 9: BLEU deltas for *targeted* **pmax** intervention in Qwen2.5 0.5B. Row i is the target language and column j is the intervention language. For example, row x col y represents BLEU score when translating English to language x intervened by y neurons.

en	1.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
nl	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1
id	0.0	0.0	1.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
ms	0.0	0.0	0.8	1.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
vi	0.0	0.0	0.1	0.1	1.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1
jp	0.0	0.0	0.0	0.0	0.1	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2
zh	0.1	0.0	0.0	0.0	0.1	0.3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
fr	0.0	0.1	0.0	0.0	0.0	0.0	0.0	1.0	0.2	0.0	0.0	0.1	0.3	0.0	0.0	0.0	0.0
pt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0
ru	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1
et	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0	0.1	0.0	0.1	0.1	0.1	0.0
ht	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	1.0	0.1	0.4	0.2	0.0	0.1
it	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.1	1.0	0.0	0.0	0.0	0.0
qu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.0	1.0	0.2	0.1	0.1
sw	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.2	1.0	0.1	0.0
ta	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	1.0	0.1
th	0.0	0.0	0.1	0.1	0.2	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	1.0
tr	0.0	0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.1	1.0
en	nl	id	ms	vi	jp	zh	fr	pt	ru	et	ht	it	qu	sw	ta	th	tr

Figure 10: Symmetric heatmaps of overlap neuron proportions in SeaLLMs3 1.5B using jaccard distance for LAPE neurons.

	Qwen							Gemma							SeaLLM						
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓		pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓		pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	
	Baseline																				
en	96.22	97.84	95.14	94.05	92.43	92.97		99.46	99.46	79.46	78.38	31.35	81.62		99.46	100.0	95.14	94.05	51.89	60.54	
nl	90.81	88.11	80.54	81.08	71.89	78.92		71.89	62.7	42.16	44.32	31.35	58.92		94.05	97.3	76.22	77.84	42.7	43.78	
ms	40.54	46.49	31.89	30.81	29.73	34.05		29.73	25.95	31.35	31.35	41.08	40.0		40.0	42.16	31.89	32.43	36.22	23.78	
vi	89.73	90.81	76.22	76.22	54.59	59.46		81.08	72.97	31.35	29.73	9.19	14.59		98.38	100.0	72.43	70.81	12.97	16.22	
jp	87.57	88.11	63.78	68.11	29.73	28.11		79.46	91.89	42.7	41.62	21.62	38.38		97.3	97.3	42.7	36.76	10.81	5.41	
zh	87.03	95.68	54.05	55.68	48.11	32.97		78.92	90.27	27.03	26.49	27.03	32.97		100.0	100.0	25.95	20.54	13.51	2.16	
fr	92.97	95.14	78.38	82.16	67.57	74.05		91.89	94.59	77.84	80.0	37.3	47.57		100.0	99.46	80.54	78.38	52.43	54.05	
pt	91.89	89.19	75.14	74.05	61.08	69.19		95.14	98.38	78.92	81.08	35.14	60.0		88.65	88.65	71.35	73.51	27.03	32.43	
ru	92.97	97.84	57.84	60.0	34.59	37.3		88.11	90.81	27.57	26.49	5.95	12.43		99.46	100.0	41.08	38.38	3.78	5.41	
et	94.05	92.97	72.97	73.51	60.54	68.11		61.08	62.7	32.43	30.81	20.0	21.08		92.97	95.14	69.73	69.73	30.27	35.68	
it	98.92	96.76	83.78	87.03	66.49	74.59		93.51	98.38	75.68	74.59	47.57	67.03		95.14	97.84	75.68	75.14	44.86	52.43	
ta	73.51	68.11	1.08	1.08	0.54	0.0		10.81	5.95	6.49	4.32	0.0	0.0		80.54	76.76	0.54	0.54	0.0	0.0	
th	95.14	95.68	14.05	14.59	7.57	4.86		52.97	60.54	11.35	11.89	0.54	0.54		99.46	100.0	5.41	8.11	1.62	1.08	
tr	92.97	91.89	56.22	55.68	41.08	45.95		45.95	52.97	24.32	29.19	37.84	37.84		94.59	96.76	45.95	48.11	19.46	28.11	
	LAPE																				
en	99.46	98.92	96.22	96.22	84.32	33.51		100.0	100.0	100.0	100.0	100.0	9.73		100.0	100.0	94.59	94.05	98.92	5.41	
nl	100.0	100.0	92.97	90.81	67.57	32.97		100.0	100.0	100.0	100.0	68.65	15.14		100.0	100.0	93.51	89.73	61.08	16.22	
ms	47.57	55.68	30.81	31.89	37.84	38.92		56.76	55.68	42.70	41.08	45.95	47.03		85.95	52.97	39.46	43.78	47.03	44.32	
vi	100.0	100.0	80.00	75.14	23.78	18.92		100.0	100.0	96.76	96.22	14.05	10.81		100.0	100.0	87.57	76.76	14.59	8.65	
jp	100.0	100.0	75.14	71.35	28.11	3.24		100.0	100.0	60.54	60.0	20.0	6.49		100.0	100.0	43.78	44.86	16.22	1.62	
zh	100.0	100.0	48.11	43.78	14.05	5.41		100.0	100.0	87.03	81.08	33.51	9.73		100.0	100.0	56.76	56.76	49.19	1.08	
fr	100.0	100.0	96.22	95.14	84.86	7.57		100.0	100.0	97.84	97.84	60.00	14.59		100.0	100.0	98.92	95.14	85.95	17.30	
pt	99.46	100.0	96.76	94.05	72.43	9.73		100.0	100.0	98.92	98.92	77.30	9.19		100.0	100.0	96.76	91.35	82.7	4.32	
ru	100.0	100.0	90.27	79.46	2.70	0.54		100.0	100.0	95.14	94.59	18.92	1.62		100.0	100.0	86.49	71.35	3.24	0.00	
et	94.05	94.05	77.84	76.22	67.03	49.73		100.0	88.65	97.84	94.59	55.68	16.22		99.46	98.92	89.19	89.73	39.46	18.92	
it	100.0	100.0	92.43	92.43	82.70	39.46		100.0	100.0	99.46	99.46	89.19	24.32		100.0	99.46	98.92	91.89	83.78	14.05	
ta	100.0	100.0	1.08	1.08	0.00	0.00		100.0	98.38	50.81	45.95	0.00	0.00		100.0	100.0	1.08	1.08	0.00	0.00	
th	100.0	100.0	41.62	41.62	0.54	0.54		100.0	100.0	65.41	64.32	1.08	0.00		100.0	100.0	25.41	22.70	0.54	0.00	
tr	98.92	100.0	56.76	56.76	45.41	29.73		91.89	83.24	96.22	94.59	48.65	40.00		100.0	100.0	52.43	52.43	31.89	14.05	

Table 4: Language Steering Shift (LSS) scores from id to languages listed for amplifying (pmax, pmedian, =max, +max) and deactivating factors (=0, =10p) across models. Bolded texts denote the highest values within a model.

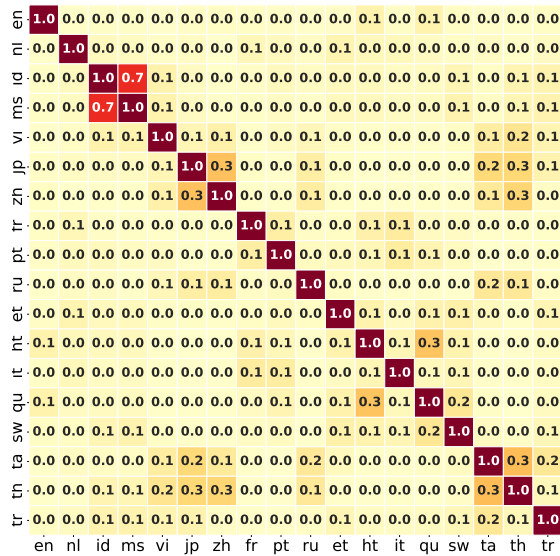


Figure 11: Symmetric heatmaps of overlap neuron proportions in Gemma2 2B using jaccard distance for LAPE neurons.

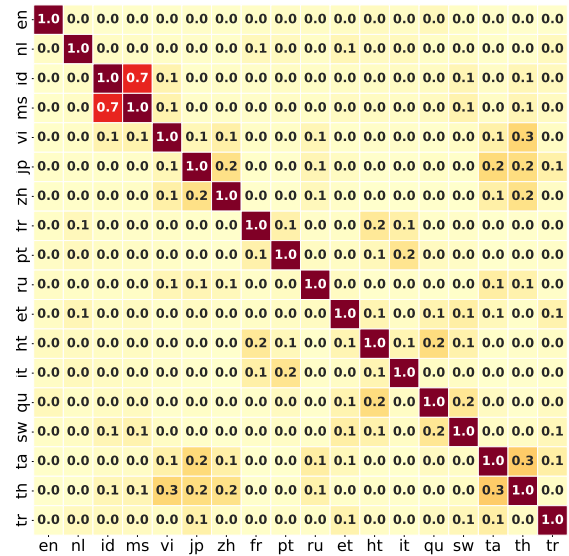


Figure 12: Symmetric heatmaps of overlap neuron proportions in Gemma2 9B using jaccard distance for LAPE neurons.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
nl	63.24	68.65	30.81	32.97	5.41	10.27	63.78	70.81	37.84	40.54	29.19	11.89	90.27	97.3	17.3	19.46	10.27	8.11
id	67.03	69.73	24.32	21.08	11.35	12.97	77.3	77.84	65.41	65.95	60.0	52.43	87.03	96.22	15.68	16.22	5.95	4.86
ms	47.57	47.57	24.86	23.78	8.65	6.49	54.59	42.7	28.11	28.65	28.65	24.86	77.84	87.03	12.43	7.03	5.41	3.78
vi	74.59	67.57	28.11	25.41	7.57	7.57	73.51	70.27	39.46	41.62	9.19	5.95	95.14	94.05	10.81	10.27	4.86	5.95
jp	34.59	45.95	7.03	6.49	1.08	1.08	57.84	66.49	29.19	28.11	27.03	7.03	84.86	97.3	0.54	0.54	1.62	0.54
zh	29.73	81.08	5.95	6.49	10.27	1.62	50.27	83.24	9.73	9.19	21.08	5.95	82.16	97.84	0.0	0.0	14.59	1.62
fr	67.57	71.89	47.03	46.49	38.92	37.84	85.41	88.65	58.92	58.92	56.22	22.16	88.11	98.92	37.3	35.14	21.62	18.92
pt	71.35	71.35	7.03	6.49	3.78	6.49	88.11	91.35	66.49	66.49	64.86	31.89	77.84	94.05	8.65	8.65	7.03	4.32
ru	58.38	70.27	4.32	4.86	2.7	3.78	78.38	83.24	20.0	21.08	4.86	4.32	92.97	97.84	9.19	5.41	1.62	1.08
et	94.59	94.59	17.84	13.51	12.97	10.81	58.38	62.16	30.27	23.78	5.41	3.78	96.22	96.22	6.49	7.03	6.49	2.16
it	74.59	75.14	30.27	29.73	23.24	20.0	89.19	91.35	62.16	61.62	58.38	38.38	87.57	95.14	18.92	21.62	12.43	3.24
ta	10.81	9.19	0.0	0.0	0.0	0.0	3.24	1.08	1.08	1.08	0.0	0.0	19.46	11.35	0.0	0.0	0.0	0.0
th	80.0	78.38	0.54	1.62	0.54	0.54	31.35	43.24	1.08	2.16	0.0	0.0	95.14	98.92	1.62	1.62	1.08	0.0
tr	71.35	69.73	18.92	16.76	3.78	7.03	35.68	35.14	8.11	6.49	16.76	9.73	83.78	98.38	4.86	7.03	4.32	3.24
	LAPE																	
nl	97.84	96.76	41.08	40.54	53.51	21.62	99.46	96.76	95.14	94.59	63.24	10.27	99.46	100.0	72.97	70.81	54.05	11.89
id	98.92	99.46	28.11	23.78	81.62	14.59	98.92	94.59	95.14	92.97	71.89	28.11	100.0	100.0	52.97	56.76	60.0	3.78
ms	98.38	98.92	17.84	19.46	45.95	10.27	97.3	94.59	94.59	94.05	46.49	14.59	98.92	99.46	32.43	33.51	32.97	4.32
vi	100.0	100.0	48.11	38.38	18.38	3.24	98.92	98.38	76.76	79.46	17.3	9.73	100.0	100.0	71.35	44.86	15.14	2.16
jp	99.46	99.46	8.65	8.11	25.95	0.54	100.0	99.46	41.08	40.54	35.68	8.65	100.0	100.0	1.08	0.54	8.65	0.0
zh	99.46	99.46	7.03	0.0	12.97	7.03	95.68	99.46	60.54	61.08	38.38	6.49	100.0	100.0	10.81	2.16	23.24	4.86
fr	96.22	96.76	77.3	74.05	70.81	7.57	100.0	98.92	79.46	77.84	70.81	16.22	100.0	100.0	76.76	70.27	65.95	1.62
pt	99.46	99.46	57.3	40.54	62.7	3.78	100.0	100.0	80.54	81.62	72.97	26.49	100.0	100.0	74.05	61.62	62.16	3.78
ru	100.0	100.0	16.22	11.89	2.7	1.62	99.46	99.46	83.24	80.0	10.81	2.16	100.0	100.0	22.7	17.84	5.41	0.54
et	96.76	94.59	15.68	14.05	56.76	18.38	98.92	96.76	95.68	95.68	39.46	5.95	97.84	97.84	31.89	31.89	57.84	25.95
it	99.46	99.46	60.0	44.86	68.11	13.51	100.0	99.46	87.57	87.03	69.73	9.73	99.46	99.46	61.62	58.92	76.76	3.24
ta	99.46	99.46	0.0	0.0	0.0	0.0	97.3	68.11	11.89	11.35	0.0	0.0	99.46	99.46	0.0	0.0	0.0	0.0
th	100.0	100.0	3.24	2.7	0.0	0.0	100.0	95.68	34.59	32.97	0.0	0.0	100.0	100.0	1.62	1.08	1.08	0.0
tr	98.92	98.92	15.14	13.51	34.59	9.73	86.49	80.54	72.97	67.57	42.16	13.51	97.84	97.84	19.46	19.46	32.97	3.24

Table 5: Language Steering Shift (LSS) scores from en to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	99.46	95.68	92.97	78.92	88.65	99.46	99.46	100.0	100.0	84.86	96.76	100.0	93.51	99.46	97.84	78.38	79.46
nl	69.19	65.95	66.49	69.19	65.41	70.81	91.35	91.35	78.92	84.32	74.59	81.62	83.78	71.35	74.59	74.05	58.38	69.19
id	63.78	64.86	75.68	62.16	60.0	67.57	96.76	95.68	94.05	94.59	84.32	92.97	70.81	67.57	71.89	72.97	59.46	63.78
ms	52.43	49.19	38.38	34.59	31.89	41.62	84.32	84.32	69.19	69.19	66.49	84.86	45.95	50.27	37.84	37.84	25.95	27.03
vi	89.73	89.73	63.24	66.49	37.84	54.59	98.92	98.38	86.49	86.49	30.81	71.89	99.46	98.92	76.76	80.0	27.03	29.19
jp	85.95	89.19	50.27	49.73	14.59	17.3	95.14	95.14	78.92	79.46	29.19	49.73	99.46	99.46	40.0	35.68	2.16	2.16
zh	83.78	96.76	42.7	52.97	41.62	27.03	97.3	97.3	78.92	81.62	28.65	45.95	98.92	100.0	27.03	25.95	24.32	2.7
fr	89.19	92.43	77.3	74.05	75.14	74.59	99.46	98.38	94.05	94.05	66.49	92.97	94.05	98.38	83.24	82.7	61.62	66.49
pt	77.84	83.78	64.32	65.41	58.92	65.41	98.38	98.38	97.84	97.84	76.22	88.11	84.86	91.89	69.73	71.35	41.08	48.11
ru	97.3	98.38	37.84	45.41	22.7	24.32	98.38	98.92	88.65	90.27	10.27	28.11	98.92	100.0	27.57	25.41	2.7	3.24
it	86.49	90.27	65.95	69.19	67.57	72.97	98.38	99.46	92.97	93.51	70.81	90.27	89.19	95.14	71.35	70.27	65.41	69.19
ta	39.46	34.05	0.0	0.0	0.0	0.0	43.78	32.43	15.68	16.76	0.54	1.62	81.62	76.22	0.0	0.0	0.0	0.0
th	99.46	99.46	12.43	18.92	5.95	3.78	81.08	77.84	24.32	27.03	2.16	14.05	100.0	100.0	2.7	2.7	1.08	1.08
tr	74.05	74.05	47.57	48.65	35.14	43.78	92.43	89.19	74.59	77.3	77.3	86.49	72.43	87.03	55.14	56.76	38.92	48.11
	LAPE																	
en	91.35	90.27	100.0	100.0	49.73	21.08	100.0	98.92	100.0	100.0	95.14	22.16	97.84	97.84	97.3	97.84	80.54	7.57
nl	95.68	92.97	82.7	78.38	49.19	38.92	99.46	96.76	100.0	100.0	75.68	65.41	100.0	99.46	88.11	83.78	42.16	30.81
id	95.68	96.22	63.78	61.62	57.84	34.59	99.46	95.14	99.46	99.46	45.95	14.05	99.46	96.22	90.27	91.35	44.86	17.84
ms	92.43	95.68	52.43	44.32	33.51	17.3	99.46	98.92	99.46	99.46	53.51	39.46	98.38	96.76	71.35	71.35	44.32	28.65
vi	100.0	100.0	88.11	80.0	14.59	1.08	100.0	100.0	99.46	99.46	18.92	5.95	100.0	100.0	91.89	88.11	37.3	7.57
jp	100.0	100.0	60.54	49.19	8.11	0.54	100.0	100.0	95.68	95.68	13.51	3.24	100.0	100.0	32.97	32.43	8.11	0.0
zh	100.0	100.0	49.19	42.7	15.14	7.57	100.0	100.0	96.22	94.05	15.68	4.32	100.0	100.0	62.7	55.14	40.54	0.0
fr	98.92	98.92	97.84	92.97	52.97	5.41	99.46	99.46	100.0	100.0	32.97	16.76	99.46	99.46	99.46	98.38	65.95	6.49
pt	98.92	98.92	90.27	88.65	44.86	10.27	100.0	100.0	98.38	98.38	49.73	21.62	99.46	98.92	94.05	91.89	67.03	14.05
ru	100.0	100.0	81.62	77.3	3.24	2.7	100.0	100.0	99.46	99.46	9.19	2.16	100.0	100.0	91.35	81.62	1.08	0.0
it	97.84	98.38	87.57	87.03	64.32	14.59	100.0	97.84	100.0	100.0	55.14	26.49	99.46	99.46	98.38	90.81	71.89	16.22
ta	100.0	100.0	0.0	0.0	0.0	0.0	100.0	99.46	85.41	85.95	0.0	0.0	100.0	100.0	0.0	0.0	0.54	0.0
th	100.0	100.0	52.43	53.51	0.0	0.0	100.0	100.0	85.41	84.86	0.0	0.0	100.0	100.0	41.08	18.38	0.54	0.0
tr	98.92	99.46	64.86	60.54	52.43	44.86	74.59	61.08	96.22	95.68	58.38	55.14	99.46	99.46	71.35	70.27	59.46	51.89

Table 6: Language Steering Shift (LSS) scores from et to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	98.38	97.84	92.43	95.14	43.24	69.73	98.92	99.46	58.92	58.92	49.73	94.59	98.38	98.38	72.97	73.51	54.05	51.89
nl	76.76	78.92	56.22	57.84	35.14	47.03	44.32	32.97	43.24	45.95	57.84	70.81	88.11	96.22	58.38	56.22	25.41	27.03
id	75.14	78.92	61.08	61.62	34.59	49.73	88.65	85.95	81.08	80.0	68.65	84.86	95.68	98.38	57.84	52.97	17.3	15.68
ms	69.19	69.19	57.3	56.22	31.89	39.46	47.57	37.3	38.92	40.0	40.54	58.38	75.14	93.51	42.7	42.16	14.05	14.59
vi	69.73	78.92	51.35	53.51	30.27	40.54	81.08	74.05	47.03	43.78	8.11	14.05	97.84	97.3	45.95	48.11	7.03	11.89
jp	58.38	60.0	18.92	23.24	11.35	14.59	76.76	87.03	38.38	38.92	31.35	48.11	100.0	100.0	15.68	16.76	2.7	1.62
zh	59.46	94.05	24.32	24.86	17.84	15.68	70.27	92.97	26.49	24.86	29.19	42.7	96.76	100.0	14.59	14.05	4.32	1.08
pt	89.19	88.11	46.49	41.08	25.41	38.38	95.68	97.84	75.68	75.68	66.49	77.84	90.81	91.89	45.95	49.19	10.27	20.54
ru	90.81	95.68	32.43	31.35	10.27	13.51	90.81	92.97	28.11	27.57	5.95	8.11	97.84	100.0	14.59	13.51	1.62	2.16
et	100.0	99.46	61.08	58.38	36.76	46.49	50.81	63.24	33.51	30.81	23.78	40.54	100.0	99.46	52.43	51.35	14.05	17.3
it	80.54	72.43	61.62	60.0	51.35	56.22	89.19	96.76	72.43	70.27	75.14	83.24	95.14	98.38	54.05	54.05	36.22	29.73
ta	45.95	45.41	0.54	0.54	0.0	0.0	8.11	2.16	2.16	2.7	0.0	0.0	63.78	60.54	0.0	0.0	0.0	0.0
th	91.35	91.89	18.38	14.05	3.78	6.49	37.3	49.73	8.65	8.11	0.0	3.78	99.46	100.0	6.49	7.57	1.08	0.54
tr	80.0	79.46	58.38	58.92	35.14	42.16	48.11	42.7	23.78	23.78	38.38	42.16	75.68	98.92	36.22	37.84	10.27	10.81
	LAPE																	
en	100.0	100.0	97.3	90.27	92.43	10.27	100.0	100.0	100.0	100.0	97.3	52.97	100.0	100.0	99.46	99.46	100.0	17.84
nl	100.0	100.0	74.59	61.62	60.0	27.03	99.46	96.76	88.65	88.65	64.32	37.84	100.0	98.92	94.59	88.65	68.65	10.81
id	100.0	100.0	70.27	68.11	78.92	11.89	100.0	92.43	98.38	99.46	73.51	58.92	99.46	97.84	78.38	76.76	70.27	9.19
ms	100.0	100.0	55.14	51.89	42.16	12.97	100.0	96.22	98.38	94.59	39.46	23.24	100.0	100.0	70.27	71.35	31.35	4.32
vi	100.0	100.0	55.68	54.05	14.59	5.41	100.0	98.92	88.11	87.03	16.76	8.65	100.0	100.0	93.51	84.86	17.3	8.11
jp	100.0	100.0	47.03	44.32	10.27	0.54	100.0	100.0	56.22	56.76	30.81	29.73	100.0	100.0	18.92	19.46	10.27	0.0
zh	100.0	100.0	26.49	24.32	15.68	0.0	98.92	100.0	59.46	66.49	18.38	3.24	100.0	100.0	34.59	30.81	42.7	2.7
pt	100.0	100.0	93.51	90.81	72.43	12.43	100.0	100.0	96.76	97.3	80.0	40.54	99.46	98.92	91.35	84.86	54.05	16.22
ru	100.0	100.0	79.46	61.08	3.24	0.0	100.0	100.0	89.73	89.19	5.41	0.0	100.0	100.0	54.05	40.0	1.08	0.54
et	100.0	100.0	53.51	48.65	72.43	41.08	100.0	94.59	99.46	98.92	48.65	13.51	100.0	100.0	61.08	60.54	81.08	24.86
it	100.0	100.0	66.49	65.41	74.05	7.57	96.76	95.14	97.3	97.84	76.76	63.24	100.0	100.0	85.95	80.54	57.3	34.05
ta	100.0	100.0	0.54	0.0	0.0	0.0	100.0	89.19	31.89	32.97	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0
th	100.0	100.0	43.78	44.32	0.54	0.0	100.0	97.3	60.54	51.89	1.08	0.54	100.0	100.0	16.76	5.41	0.0	0.0
tr	100.0	100.0	52.97	50.81	49.73	14.05	75.14	77.3	96.76	96.22	65.41	42.7	100.0	100.0	51.35	51.35	35.68	5.41

Table 7: Language Steering Shift (LSS) scores from fr to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax ↑	pmed ↑	=max ↑	+max ↑	=0 ↓	=10p ↓	pmax ↑	pmed ↑	=max ↑	+max ↑	=0 ↓	=10p ↓	pmax ↑	pmed ↑	=max ↑	+max ↑	=0 ↓	=10p ↓
	Baseline																	
en	96.22	95.14	94.05	93.51	75.68	88.65	98.92	99.46	64.86	66.49	46.49	85.95	98.38	99.46	91.89	92.43	55.14	57.84
nl	90.27	94.05	78.38	80.54	56.22	79.46	40.54	38.38	50.27	48.11	47.57	68.65	94.59	97.3	80.0	77.84	25.95	23.78
id	90.81	89.73	74.59	70.27	48.65	71.89	73.51	71.89	76.22	76.76	76.76	87.03	96.22	98.38	67.03	67.03	22.16	15.14
ms	72.43	74.59	53.51	54.05	38.38	39.46	34.59	29.19	40.0	39.46	38.92	54.05	87.03	92.97	36.22	40.0	14.59	11.89
vi	90.27	93.51	68.11	68.11	44.32	66.49	80.54	73.51	52.97	54.05	12.43	21.62	99.46	98.92	70.81	69.73	7.03	9.19
jp	86.49	91.35	48.11	43.24	12.97	16.76	62.16	78.38	48.11	47.03	32.43	42.7	100.0	98.92	35.68	26.49	3.24	1.62
zh	85.95	95.68	30.27	28.65	29.19	27.57	63.78	90.27	26.49	24.86	25.95	33.51	97.3	100.0	18.92	16.76	9.19	1.08
fr	94.59	96.76	83.24	81.08	79.46	82.16	88.11	92.97	71.89	71.89	51.89	73.51	97.3	97.84	81.62	82.16	29.73	46.49
pt	78.38	75.68	68.11	71.89	51.89	62.7	84.86	82.16	76.76	74.05	63.78	84.32	78.92	83.24	60.0	62.16	22.7	24.32
ru	94.59	96.22	45.41	47.57	34.59	33.51	89.19	90.27	31.89	31.35	8.11	13.51	97.84	99.46	23.78	21.08	1.08	2.16
et	92.97	94.59	71.89	72.97	49.73	58.92	55.14	52.97	36.22	35.68	18.38	35.68	98.38	98.92	56.22	56.22	13.51	23.78
ta	54.59	54.59	1.62	1.62	0.0	0.0	8.11	2.16	1.62	1.08	0.0	0.0	76.22	71.89	0.54	0.54	0.0	0.0
th	98.38	98.38	18.38	20.54	7.57	8.11	34.05	46.49	11.89	14.05	1.62	7.03	100.0	100.0	5.95	5.41	1.08	1.08
tr	85.41	82.7	53.51	56.22	42.16	46.49	29.73	29.19	31.89	27.57	38.38	35.14	93.51	97.3	38.92	43.24	12.97	14.05
	LAPE																	
en	100.0	98.92	94.05	93.51	74.59	22.16	100.0	100.0	100.0	100.0	97.84	6.49	99.46	98.38	95.68	95.14	92.97	14.59
nl	100.0	100.0	87.03	86.49	74.59	37.3	100.0	96.76	98.92	98.92	67.03	22.7	100.0	100.0	97.84	95.68	61.08	11.89
id	100.0	100.0	78.92	70.81	76.76	10.81	99.46	93.51	99.46	99.46	74.05	38.38	100.0	100.0	88.11	94.05	65.41	3.24
ms	100.0	100.0	49.19	48.65	41.08	15.14	99.46	97.3	99.46	99.46	38.92	21.08	100.0	100.0	70.81	69.19	23.24	6.49
vi	100.0	100.0	80.0	78.92	21.08	3.24	100.0	98.92	98.38	93.51	21.08	12.97	100.0	100.0	91.89	88.65	18.38	5.95
jp	100.0	100.0	71.35	66.49	21.62	1.62	100.0	100.0	48.65	53.51	25.95	9.19	100.0	100.0	18.92	21.08	11.89	0.0
zh	100.0	100.0	41.08	39.46	16.22	4.86	98.92	100.0	55.14	58.92	24.86	3.78	100.0	100.0	40.0	36.76	48.65	0.54
fr	100.0	100.0	95.68	94.05	59.46	17.84	100.0	100.0	97.3	96.22	81.62	47.57	100.0	99.46	97.84	96.76	64.86	32.43
pt	98.92	99.46	86.49	83.24	52.43	5.41	100.0	100.0	96.76	96.76	85.95	52.97	100.0	100.0	92.43	90.27	56.22	25.95
ru	100.0	100.0	86.49	78.92	5.41	1.62	100.0	100.0	92.97	93.51	12.43	2.16	100.0	100.0	83.78	55.14	2.7	0.0
et	100.0	98.92	74.05	57.3	74.05	44.32	100.0	95.14	99.46	99.46	61.08	20.0	99.46	99.46	85.41	84.32	63.78	14.05
ta	100.0	100.0	0.54	0.54	0.0	0.0	100.0	93.51	27.03	26.49	0.0	0.0	100.0	100.0	0.54	0.54	0.0	0.0
th	100.0	100.0	44.86	45.95	1.08	0.0	100.0	100.0	61.62	51.35	1.62	1.62	100.0	100.0	12.43	10.27	0.54	0.0
tr	98.92	100.0	49.19	51.89	51.35	15.68	89.73	82.16	98.38	96.22	60.0	35.68	100.0	100.0	49.73	49.19	31.35	10.27

Table 8: Language Steering Shift (LSS) scores from it to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	100.0	100.0	95.14	96.22	100.0	98.92	97.84	97.3	58.92	91.35	100.0	100.0	100.0	99.46	94.05	91.35
nl	99.46	98.92	96.22	97.3	60.54	72.43	89.73	92.43	69.19	73.51	52.97	65.41	100.0	100.0	92.43	92.97	14.59	17.84
id	100.0	99.46	92.43	94.05	38.38	67.57	95.68	92.97	95.14	93.51	58.38	79.46	100.0	100.0	90.27	90.27	16.76	16.22
ms	99.46	99.46	75.14	75.68	32.97	40.0	76.76	71.35	63.78	62.16	28.65	52.97	100.0	99.46	75.14	75.68	12.97	9.73
vi	100.0	100.0	97.3	97.84	78.38	83.24	95.68	94.05	76.76	76.76	6.49	21.08	100.0	100.0	87.03	89.73	16.76	18.92
zh	81.62	88.11	75.14	77.84	27.57	53.51	82.16	91.89	59.46	58.38	56.22	68.11	91.35	91.89	57.3	58.92	24.86	32.97
fr	98.92	98.38	92.97	92.43	69.73	77.84	96.22	98.38	86.49	86.49	44.32	78.38	100.0	100.0	91.89	91.89	36.22	40.54
pt	99.46	100.0	97.84	97.3	38.92	64.86	100.0	100.0	84.32	87.03	50.81	80.54	99.46	100.0	93.51	92.43	14.59	10.81
ru	98.92	99.46	69.19	69.19	49.19	54.59	88.11	89.73	56.76	56.76	24.86	38.92	99.46	100.0	78.92	79.46	15.14	14.05
et	100.0	100.0	94.59	96.22	38.92	68.65	75.68	76.22	54.05	52.43	7.03	20.0	100.0	100.0	85.95	87.03	15.68	14.05
it	99.46	96.22	93.51	94.59	56.22	78.38	96.22	100.0	84.32	84.86	58.92	85.95	99.46	96.76	91.35	91.89	36.22	27.57
ta	77.84	77.3	1.08	1.08	0.0	0.0	13.51	2.16	6.49	5.95	0.0	0.54	75.14	67.57	1.08	1.08	0.0	0.0
th	99.46	100.0	32.97	37.3	24.32	32.43	38.38	34.59	14.59	14.59	6.49	11.89	100.0	100.0	44.32	44.32	12.43	12.43
tr	98.92	91.35	80.0	79.46	34.05	44.32	82.7	77.84	45.95	44.86	18.92	34.59	100.0	100.0	70.27	70.27	18.38	38.92
	LAPE																	
en	94.05	92.97	100.0	100.0	74.05	26.49	100.0	100.0	100.0	100.0	79.46	7.03	92.43	90.81	100.0	100.0	75.14	17.84
nl	100.0	99.46	96.22	96.22	27.03	12.97	100.0	96.76	100.0	100.0	50.27	9.19	100.0	100.0	98.38	97.3	69.73	24.86
id	100.0	100.0	96.76	96.22	74.05	9.19	99.46	98.38	99.46	99.46	41.62	8.65	100.0	100.0	98.92	99.46	78.92	18.38
ms	100.0	99.46	75.14	74.59	35.68	12.97	99.46	98.92	100.0	100.0	26.49	7.57	100.0	100.0	75.14	76.76	39.46	11.89
vi	100.0	100.0	98.38	98.38	54.59	43.78	100.0	97.3	99.46	99.46	14.59	5.95	100.0	100.0	97.3	87.03	27.03	15.68
zh	92.97	96.22	78.38	79.46	32.97	30.27	88.65	94.05	87.03	87.03	67.03	51.89	92.97	92.43	76.76	75.68	74.59	52.43
fr	98.92	98.92	97.84	93.51	50.81	18.92	100.0	96.76	100.0	100.0	66.49	20.54	100.0	100.0	98.92	94.59	44.32	21.08
pt	99.46	98.92	98.92	97.84	56.76	17.84	100.0	99.46	100.0	100.0	64.86	21.62	100.0	99.46	90.81	89.73	58.38	7.03
ru	100.0	100.0	91.89	92.97	22.7	20.54	100.0	98.92	97.3	96.76	23.24	8.65	100.0	100.0	93.51	89.19	32.43	28.65
et	100.0	100.0	96.22	90.27	57.3	22.7	98.92	95.14	100.0	100.0	36.76	9.19	100.0	100.0	95.68	95.68	59.46	21.08
it	98.38	98.92	94.59	92.43	78.92	16.76	100.0	98.38	100.0	100.0	61.62	22.7	99.46	98.92	97.3	96.22	54.59	13.51
ta	100.0	100.0	1.62	1.62	0.0	0.0	75.14	12.97	55.68	56.76	1.08	0.54	100.0	99.46	0.54	0.54	0.0	0.0
th	100.0	100.0	61.08	61.08	9.73	7.57	100.0	81.08	60.0	58.92	7.03	5.95	100.0	100.0	42.7	44.32	8.65	5.95
tr	100.0	100.0	73.51	69.73	24.32	5.41	94.59	82.7	99.46	99.46	38.38	20.54	100.0	100.0	74.05	75.14	24.86	17.84

Table 9: Language Steering Shift (LSS) scores from jp to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	98.92	98.92	93.51	96.76	96.22	97.84	92.97	93.51	41.08	67.03	100.0	99.46	98.38	98.92	52.43	51.89
nl	96.22	96.22	83.78	83.78	81.62	77.3	82.16	77.84	73.51	74.05	37.3	55.68	97.3	99.46	81.62	84.86	42.16	48.11
id	61.08	46.49	67.57	67.03	70.81	71.35	71.89	74.05	69.73	69.73	57.84	62.16	65.41	49.19	66.49	67.03	64.32	69.19
vi	96.76	98.38	87.57	88.11	72.43	71.35	79.46	76.76	54.05	57.84	16.76	43.24	99.46	99.46	77.84	80.0	18.92	34.59
jp	92.97	94.59	70.81	72.97	35.68	35.68	85.41	88.65	68.65	69.73	29.19	38.92	100.0	100.0	57.3	55.68	6.49	5.95
zh	94.59	99.46	60.54	63.24	53.51	42.7	89.19	92.43	68.11	67.57	27.57	39.46	100.0	100.0	36.76	37.3	5.95	0.0
fr	95.14	94.05	85.41	88.65	82.16	84.32	92.97	93.51	80.54	80.54	45.95	61.62	98.92	99.46	88.65	87.57	54.05	52.43
pt	88.11	86.49	69.73	69.73	68.11	65.41	97.3	97.84	86.49	85.95	48.65	65.95	95.68	98.38	65.95	68.65	32.43	47.03
ru	96.76	98.38	65.95	63.78	41.08	43.24	93.51	94.59	69.73	68.65	4.86	11.35	100.0	100.0	42.7	41.62	3.24	7.03
et	96.22	95.68	83.24	84.32	77.3	78.92	85.41	89.19	68.11	64.86	15.14	45.41	97.84	98.38	83.78	82.16	35.68	37.3
it	97.3	93.51	85.41	87.57	80.0	83.24	95.14	96.76	84.32	84.86	56.76	69.73	94.59	97.3	80.0	77.84	54.05	55.68
ta	82.16	80.0	2.16	2.16	0.0	0.0	36.76	32.43	30.81	28.65	0.0	0.54	84.86	83.78	0.0	0.0	0.0	0.0
th	100.0	100.0	35.14	34.59	18.38	17.3	55.68	59.46	25.95	28.11	0.54	1.08	100.0	100.0	13.51	14.05	1.08	1.62
tr	98.92	98.92	83.24	82.16	53.51	67.03	76.22	83.24	63.24	63.24	43.78	67.57	98.38	98.92	77.84	77.84	20.0	35.14
	LAPE																	
en	100.0	98.92	100.0	100.0	92.97	27.03	100.0	98.92	100.0	99.46	95.68	23.24	100.0	100.0	99.46	100.0	94.05	18.38
nl	100.0	100.0	90.27	88.11	67.03	52.43	100.0	100.0	100.0	100.0	68.11	26.49	100.0	100.0	97.3	96.76	65.95	12.43
id	52.43	50.81	64.86	64.86	60.54	60.0	44.32	38.92	60.54	61.08	42.7	43.78	65.41	69.19	64.86	64.86	54.05	52.97
vi	100.0	100.0	90.81	87.03	29.19	17.84	100.0	99.46	91.89	92.97	19.46	10.27	100.0	100.0	97.84	92.43	30.27	20.54
jp	100.0	100.0	77.3	75.68	23.24	0.0	100.0	100.0	82.16	83.24	10.27	5.95	100.0	100.0	50.81	49.73	11.35	0.54
zh	100.0	100.0	72.43	63.78	12.97	6.49	100.0	100.0	97.3	96.22	37.84	5.95	100.0	100.0	57.84	54.05	42.16	2.7
fr	99.46	99.46	99.46	98.38	81.62	16.22	100.0	100.0	95.68	97.3	65.95	38.92	100.0	100.0	100.0	96.76	87.03	27.03
pt	97.3	97.84	98.92	95.14	72.97	21.08	100.0	100.0	97.3	97.3	82.7	48.65	100.0	100.0	100.0	95.14	83.24	16.76
ru	100.0	100.0	95.14	93.51	2.7	1.62	100.0	100.0	97.3	98.38	16.76	2.7	100.0	100.0	91.89	83.24	5.41	1.08
et	96.22	94.05	86.49	85.95	77.3	67.03	99.46	88.65	100.0	98.92	55.68	34.05	100.0	98.92	88.11	88.11	54.59	34.05
it	98.38	98.38	95.14	94.05	80.0	35.14	100.0	100.0	96.22	94.59	73.51	40.54	99.46	99.46	99.46	96.22	85.95	25.41
ta	100.0	100.0	2.16	2.16	0.0	0.0	100.0	89.19	75.14	77.84	0.0	0.0	100.0	100.0	1.62	1.62	0.0	0.0
th	100.0	100.0	67.03	64.86	0.54	0.54	100.0	100.0	69.19	69.73	1.08	0.0	100.0	100.0	48.65	41.08	0.0	0.0
tr	100.0	100.0	88.65	84.32	46.49	30.81	77.3	77.3	94.59	89.73	54.05	46.49	99.46	99.46	85.95	85.41	35.68	19.46

Table 10: Language Steering Shift (LSS) scores from ms to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	99.46	99.46	99.46	100.0	84.32	96.76	98.38	99.46	81.08	82.7	57.3	82.7	100.0	98.92	99.46	98.92	55.14	75.68
id	94.05	95.68	56.22	57.84	54.59	57.3	94.05	90.81	89.73	90.81	68.65	87.03	93.51	95.68	53.51	43.78	31.89	35.14
ms	80.0	87.57	55.68	57.3	40.0	49.19	68.65	60.54	57.84	56.22	54.59	59.46	62.7	86.49	45.95	42.7	18.38	21.62
vi	88.11	90.27	80.54	78.92	60.54	67.03	91.89	86.49	65.41	70.81	10.27	24.86	100.0	99.46	71.89	71.89	11.35	16.76
jp	94.59	95.14	42.16	39.46	18.92	16.76	89.19	89.73	55.68	57.3	31.35	44.32	98.38	98.38	22.16	21.62	2.7	1.62
zh	89.19	96.76	38.92	37.3	43.24	25.95	90.81	97.3	45.95	45.95	31.89	46.49	100.0	100.0	14.05	15.14	7.03	1.62
fr	86.49	89.19	74.05	75.14	80.54	77.84	95.68	94.59	83.78	83.78	63.24	78.92	94.59	98.92	76.22	78.38	32.43	48.11
pt	90.27	93.51	72.97	74.59	68.65	70.81	97.3	98.38	90.27	90.81	59.46	85.95	91.89	93.51	74.59	74.59	20.0	31.89
ru	95.68	98.38	40.54	40.0	25.41	28.11	96.22	95.68	54.05	52.43	16.22	19.46	99.46	100.0	21.08	23.24	3.78	3.78
et	95.14	95.68	49.19	49.73	48.65	51.89	68.65	64.32	49.19	50.27	29.73	43.24	98.92	98.92	47.03	47.57	32.43	35.14
it	97.84	95.68	83.24	85.95	68.11	74.05	98.38	99.46	83.24	84.86	64.86	84.86	97.3	96.76	77.84	75.68	56.22	50.81
ta	45.95	42.7	0.0	0.0	0.0	0.0	23.78	17.84	13.51	13.51	0.0	0.0	81.62	64.32	0.0	0.0	0.0	0.0
th	99.46	100.0	16.22	16.76	10.81	11.89	52.97	63.24	16.22	16.76	1.62	7.03	100.0	100.0	14.05	14.59	1.08	1.08
tr	85.41	91.35	52.97	52.97	42.7	46.49	76.22	57.84	40.54	42.7	52.43	47.57	92.97	96.76	50.27	52.43	23.78	27.03
	LAPE																	
en	98.92	99.46	98.92	98.92	81.62	17.3	100.0	99.46	99.46	98.92	96.76	3.24	95.14	95.14	97.84	99.46	94.05	7.57
id	98.92	99.46	63.24	56.22	81.08	54.59	100.0	98.92	100.0	100.0	67.57	41.62	99.46	99.46	86.49	91.35	64.32	24.86
ms	100.0	100.0	54.59	52.43	42.16	20.0	100.0	96.76	100.0	99.46	36.22	14.59	99.46	100.0	62.7	62.16	26.49	12.97
vi	100.0	100.0	90.27	87.57	18.38	2.7	100.0	98.38	97.84	98.92	17.84	5.95	100.0	100.0	91.35	85.41	29.19	8.11
jp	100.0	100.0	75.68	72.43	10.27	0.0	100.0	100.0	82.7	86.49	29.73	9.19	100.0	100.0	17.84	20.54	11.89	1.62
zh	100.0	100.0	36.76	36.22	11.89	0.54	99.46	100.0	91.35	90.27	34.59	10.27	100.0	100.0	44.86	36.76	49.19	3.78
fr	98.38	97.3	92.43	90.27	61.62	35.14	100.0	98.92	95.68	95.14	62.7	21.62	100.0	100.0	98.38	96.76	75.68	15.14
pt	100.0	100.0	97.3	95.14	80.0	7.03	100.0	100.0	94.05	94.05	78.38	21.62	100.0	100.0	97.84	94.59	80.54	3.24
ru	100.0	100.0	83.78	80.0	3.24	0.54	100.0	100.0	96.22	96.22	23.24	11.35	100.0	100.0	91.35	69.73	3.78	1.08
et	98.92	98.38	63.78	50.27	52.97	46.49	95.68	81.62	98.38	97.84	34.59	22.7	100.0	100.0	81.08	81.62	63.24	49.19
it	98.92	99.46	85.95	85.41	77.3	26.49	98.92	97.84	98.38	98.92	63.24	21.08	100.0	100.0	94.05	96.22	82.7	12.43
ta	100.0	100.0	1.08	1.08	0.0	0.0	100.0	98.92	52.97	56.76	0.0	0.0	100.0	100.0	1.08	1.08	0.0	0.0
th	100.0	100.0	33.51	34.59	0.54	0.0	100.0	100.0	83.24	83.24	1.08	0.0	100.0	100.0	25.95	21.08	0.0	0.0
tr	100.0	100.0	60.54	61.08	44.86	23.78	91.89	81.62	90.81	89.73	60.0	40.0	100.0	100.0	57.84	57.84	44.32	23.24

Table 11: Language Steering Shift (LSS) scores from nl to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	99.46	99.46	100.0	100.0	72.97	95.14	96.76	96.22	68.65	65.95	52.43	84.86	100.0	100.0	99.46	98.92	65.41	65.41
nl	97.84	97.3	87.57	88.11	63.24	69.73	54.05	40.0	48.11	48.65	61.62	70.27	98.92	100.0	80.54	80.0	47.03	47.03
id	97.84	99.46	71.35	68.65	56.22	59.46	76.22	58.38	72.97	72.97	71.35	88.65	100.0	100.0	72.43	70.81	38.38	29.19
ms	81.08	96.22	56.76	56.76	45.95	47.03	36.22	26.49	38.92	37.84	37.84	55.68	91.89	98.92	47.57	47.03	23.24	20.0
vi	88.65	93.51	58.38	58.38	51.89	53.51	81.08	75.68	49.19	50.27	8.11	16.22	100.0	100.0	77.3	78.38	22.7	22.7
jp	94.59	95.14	44.32	47.03	16.76	24.32	69.19	82.16	45.41	44.32	30.27	40.0	100.0	100.0	25.95	27.03	3.24	2.16
zh	90.27	97.3	40.0	50.27	28.11	24.32	56.22	82.7	21.08	17.3	23.78	40.0	99.46	100.0	22.16	23.24	12.97	0.0
fr	93.51	96.22	79.46	80.0	83.78	83.24	88.11	90.81	74.59	74.05	58.92	76.76	98.92	100.0	80.54	81.62	53.51	59.46
ru	98.38	100.0	52.97	58.38	23.78	30.81	78.92	84.32	23.78	21.62	10.81	10.27	100.0	100.0	25.95	26.49	1.62	3.24
et	97.84	97.84	66.49	71.35	52.43	58.92	54.05	60.54	41.62	35.68	25.95	29.19	99.46	99.46	65.41	64.86	30.27	36.22
it	90.81	88.65	78.38	78.92	67.03	67.57	88.65	87.03	65.95	67.03	63.24	70.81	94.59	98.92	82.7	84.32	68.65	64.32
ta	48.11	48.11	1.62	1.62	0.0	0.0	8.65	4.86	3.24	3.24	0.0	0.0	74.59	66.49	0.0	0.0	0.0	0.0
th	98.38	98.38	17.3	18.38	11.35	12.97	40.0	56.76	4.86	7.57	2.7	3.24	100.0	100.0	12.97	12.97	2.16	1.08
tr	89.19	90.81	62.16	61.08	48.11	52.43	35.14	28.65	23.24	22.7	39.46	36.22	99.46	100.0	46.49	48.11	29.73	31.89
	LAPE																	
en	100.0	98.38	100.0	100.0	82.7	30.27	100.0	100.0	100.0	100.0	99.46	28.65	97.3	95.68	98.92	100.0	91.89	27.03
nl	100.0	100.0	97.84	96.22	71.35	8.11	100.0	98.92	91.89	89.19	68.65	41.62	100.0	100.0	98.38	97.3	74.59	7.03
id	100.0	100.0	94.05	92.43	73.51	29.73	100.0	96.76	95.68	96.22	71.89	45.95	100.0	100.0	99.46	99.46	63.78	8.65
ms	100.0	100.0	51.35	54.05	47.03	18.92	96.76	95.14	92.97	93.51	45.41	30.81	100.0	100.0	72.43	73.51	25.95	1.62
vi	100.0	100.0	73.51	61.62	21.62	8.65	100.0	99.46	92.43	89.73	19.46	12.97	100.0	100.0	91.35	86.49	24.86	10.27
jp	100.0	100.0	78.38	69.19	17.3	1.08	100.0	100.0	54.59	54.59	36.76	10.27	100.0	100.0	29.73	29.19	14.05	0.54
zh	100.0	100.0	41.62	37.3	22.16	2.16	99.46	99.46	63.78	65.41	35.14	7.03	100.0	100.0	39.46	31.35	51.35	5.95
fr	100.0	100.0	96.76	94.05	77.84	29.73	99.46	99.46	95.68	96.22	77.84	59.46	100.0	100.0	99.46	99.46	70.81	30.81
ru	100.0	100.0	85.95	77.3	3.24	3.24	100.0	100.0	90.81	91.89	11.35	1.08	100.0	100.0	83.24	63.24	2.16	2.16
et	99.46	99.46	75.68	55.68	73.51	24.32	100.0	93.51	99.46	98.38	71.35	51.89	100.0	100.0	87.57	88.11	65.41	13.51
it	97.3	97.3	89.19	85.41	75.14	26.49	100.0	99.46	92.43	91.89	82.16	74.05	100.0	100.0	97.84	94.59	73.51	47.57
ta	100.0	100.0	0.54	0.54	0.0	0.0	100.0	97.3	35.14	35.68	0.0	0.0	100.0	100.0	1.08	1.08	0.0	0.0
th	100.0	100.0	52.97	57.84	1.08	0.0	100.0	100.0	66.49	51.89	2.16	2.16	100.0	100.0	24.32	12.97	1.08	0.0
tr	100.0	100.0	62.16	62.16	57.84	32.97	80.0	85.41	96.76	95.14	74.05	54.05	100.0	100.0	71.35	69.19	35.68	4.86

Table 12: Language Steering Shift (LSS) scores from pt to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	100.0	100.0	80.0	94.05	100.0	100.0	97.84	97.84	53.51	75.14	100.0	100.0	99.46	100.0	70.81	54.59
nl	99.46	100.0	95.14	94.59	30.81	65.95	96.22	94.59	84.86	86.49	49.19	67.03	100.0	100.0	93.51	92.97	19.46	19.46
id	100.0	100.0	96.76	96.76	30.81	55.68	96.76	94.05	94.59	93.51	64.86	87.57	100.0	100.0	90.81	90.81	21.62	15.68
ms	100.0	100.0	78.38	77.3	24.86	47.57	75.68	68.65	68.11	68.11	39.46	66.49	100.0	100.0	71.89	71.35	12.97	9.73
vi	99.46	98.38	89.73	91.89	49.73	90.27	91.35	88.65	68.65	69.73	15.68	67.03	100.0	100.0	87.57	86.49	25.95	25.95
jp	95.68	96.22	83.78	83.24	6.49	64.32	91.89	90.81	69.19	69.19	49.73	78.92	100.0	100.0	62.16	57.3	19.46	29.19
zh	96.76	98.38	82.16	86.49	18.38	43.24	92.97	93.51	72.97	74.05	46.49	77.3	100.0	100.0	43.24	40.0	31.35	1.62
fr	100.0	99.46	97.84	97.84	54.05	70.27	100.0	100.0	94.05	94.05	55.14	77.84	100.0	100.0	96.22	96.76	48.11	40.54
pt	98.92	97.3	98.92	98.92	27.03	63.78	100.0	100.0	97.84	98.38	67.57	84.32	98.92	99.46	95.68	94.59	17.3	16.76
et	100.0	100.0	95.14	96.76	23.24	52.43	89.73	62.16	75.68	74.05	19.46	43.24	100.0	99.46	71.35	81.08	14.05	16.76
it	98.92	98.92	98.38	98.38	31.89	69.73	99.46	100.0	95.14	96.22	61.08	76.76	100.0	100.0	96.76	96.76	51.35	42.16
ta	87.03	84.32	2.7	3.24	0.54	1.08	12.97	4.86	10.81	11.89	0.54	1.62	94.59	81.08	1.62	1.62	0.0	0.0
th	100.0	100.0	28.65	31.89	29.73	28.11	38.38	38.92	20.0	21.62	11.89	16.76	100.0	100.0	30.81	31.89	9.73	8.65
tr	96.22	91.89	80.0	80.0	27.57	51.35	87.57	76.76	64.32	63.78	45.41	71.89	100.0	100.0	70.81	71.89	14.59	12.43
	LAPE																	
en	100.0	100.0	99.46	99.46	81.08	29.19	100.0	100.0	100.0	100.0	95.14	15.14	100.0	100.0	100.0	100.0	92.43	8.65
nl	100.0	100.0	97.3	96.76	65.41	1.08	99.46	91.89	100.0	100.0	37.84	7.57	100.0	100.0	100.0	98.38	68.65	9.19
id	100.0	100.0	99.46	99.46	84.32	20.54	98.92	92.43	97.3	98.38	62.16	15.14	100.0	100.0	100.0	100.0	73.51	8.65
ms	100.0	100.0	84.32	78.38	46.49	6.49	98.92	97.84	100.0	100.0	47.57	14.59	100.0	100.0	100.0	100.0	34.59	7.57
vi	100.0	100.0	100.0	100.0	29.73	14.59	100.0	98.92	100.0	99.46	19.46	4.86	100.0	100.0	96.22	98.38	28.11	9.19
jp	99.46	99.46	95.14	94.05	16.76	10.27	99.46	97.3	93.51	95.14	28.11	16.76	100.0	100.0	94.59	90.27	11.89	4.32
zh	100.0	100.0	99.46	98.92	28.65	17.84	98.38	98.38	97.3	96.22	48.11	22.7	100.0	100.0	96.76	89.73	50.27	6.49
fr	100.0	100.0	100.0	97.84	71.35	9.73	98.38	98.38	98.38	98.38	69.73	17.3	100.0	100.0	97.3	96.76	69.19	7.03
pt	100.0	100.0	96.22	88.11	73.51	8.11	98.92	99.46	99.46	99.46	65.41	15.68	100.0	100.0	100.0	98.92	75.68	4.32
et	100.0	98.92	97.84	97.84	52.97	34.59	87.03	75.68	99.46	99.46	48.65	19.46	94.59	94.59	96.76	96.76	41.62	28.11
it	100.0	100.0	99.46	98.92	74.59	6.49	89.19	86.49	88.11	86.49	69.19	36.22	100.0	99.46	100.0	100.0	70.81	7.57
ta	98.92	98.38	0.0	1.08	0.54	0.0	62.7	20.0	31.89	25.95	0.54	0.0	99.46	99.46	18.38	15.68	0.0	0.0
th	100.0	100.0	99.46	99.46	2.16	0.0	100.0	79.46	95.68	84.32	3.78	3.78	100.0	100.0	94.05	92.97	0.0	0.0
tr	100.0	100.0	85.95	85.95	30.81	7.03	54.59	56.22	68.11	65.95	31.35	23.78	97.84	97.84	97.3	94.59	35.14	2.16

Table 13: Language Steering Shift (LSS) scores from ru to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.84	100.0	100.0	100.0	100.0	100.0	97.3	96.76
nl	100.0	100.0	100.0	100.0	95.68	100.0	100.0	100.0	98.92	98.92	88.65	100.0	100.0	100.0	100.0	100.0	84.86	81.08
id	100.0	100.0	100.0	100.0	93.51	99.46	100.0	100.0	99.46	98.92	94.59	99.46	100.0	98.38	100.0	100.0	80.0	84.86
ms	100.0	100.0	100.0	100.0	75.68	100.0	100.0	100.0	100.0	99.46	80.0	100.0	100.0	100.0	100.0	100.0	60.0	60.54
vi	100.0	98.92	100.0	100.0	100.0	100.0	99.46	98.92	98.92	98.92	77.84	98.92	100.0	100.0	100.0	100.0	73.51	78.38
jp	97.84	98.92	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.84	100.0	100.0	100.0	100.0	100.0	65.41	96.22
zh	100.0	100.0	100.0	100.0	98.92	100.0	100.0	100.0	98.92	98.38	99.46	100.0	100.0	100.0	100.0	100.0	65.95	25.95
fr	100.0	100.0	100.0	100.0	94.59	100.0	100.0	100.0	100.0	100.0	91.35	100.0	100.0	100.0	100.0	100.0	90.81	83.78
pt	100.0	100.0	100.0	100.0	87.57	100.0	100.0	100.0	100.0	100.0	95.68	100.0	100.0	100.0	100.0	100.0	84.32	81.62
ru	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.92	98.38	88.65	99.46	100.0	98.92	100.0	100.0	23.24	32.43
et	100.0	99.46	100.0	100.0	96.76	100.0	100.0	100.0	98.38	98.92	69.73	99.46	100.0	100.0	100.0	100.0	87.57	92.43
it	100.0	100.0	100.0	100.0	96.76	100.0	100.0	100.0	100.0	100.0	98.38	100.0	100.0	100.0	100.0	100.0	92.43	89.73
th	100.0	100.0	98.38	98.38	98.92	98.92	97.84	98.38	76.22	76.76	87.57	93.51	99.46	100.0	97.84	97.3	24.32	79.46
tr	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.46	98.92	82.7	98.92	100.0	99.46	100.0	100.0	58.92	60.0
	LAPE																	
en	97.84	97.3	100.0	100.0	56.76	12.97	100.0	100.0	100.0	100.0	98.92	21.08	96.22	96.22	100.0	100.0	87.57	24.32
nl	100.0	100.0	100.0	100.0	69.73	32.97	100.0	94.59	100.0	100.0	37.84	7.57	100.0	97.84	100.0	100.0	54.05	11.89
id	100.0	100.0	100.0	100.0	76.76	48.65	100.0	99.46	100.0	100.0	74.05	55.14	100.0	100.0	100.0	100.0	68.11	34.05
ms	100.0	100.0	100.0	100.0	45.95	30.81	100.0	99.46	100.0	100.0	30.81	17.84	99.46	99.46	100.0	100.0	39.46	26.49
vi	100.0	99.46	100.0	100.0	28.11	27.57	99.46	99.46	100.0	100.0	21.62	10.27	100.0	100.0	100.0	100.0	40.0	23.24
jp	100.0	100.0	100.0	100.0	40.0	44.86	100.0	100.0	100.0	100.0	29.73	23.24	100.0	100.0	100.0	100.0	7.57	2.16
zh	100.0	100.0	100.0	100.0	14.05	8.11	100.0	100.0	100.0	100.0	62.7	42.16	100.0	100.0	100.0	100.0	25.41	0.0
fr	99.46	99.46	100.0	100.0	71.35	41.62	100.0	100.0	100.0	100.0	47.03	46.49	100.0	100.0	100.0	100.0	71.89	16.22
pt	100.0	100.0	100.0	100.0	57.3	35.14	100.0	100.0	100.0	100.0	71.35	46.49	100.0	100.0	100.0	100.0	69.73	18.38
ru	100.0	100.0	100.0	100.0	5.41	5.41	100.0	100.0	100.0	100.0	46.49	26.49	100.0	100.0	100.0	100.0	2.16	2.16
et	99.46	99.46	100.0	100.0	88.11	84.86	100.0	97.3	100.0	100.0	57.3	28.65	100.0	98.92	100.0	100.0	54.05	48.65
it	99.46	99.46	100.0	100.0	42.16	11.89	100.0	100.0	100.0	100.0	88.65	88.11	99.46	99.46	100.0	100.0	78.38	21.08
th	100.0	100.0	99.46	99.46	12.97	21.08	100.0	100.0	100.0	99.46	91.35	94.05	100.0	100.0	99.46	99.46	1.62	2.16
tr	100.0	100.0	100.0	100.0	47.03	39.46	100.0	95.68	100.0	100.0	87.57	84.86	100.0	100.0	100.0	100.0	23.78	16.76

Table 14: Language Steering Shift (LSS) scores from ta to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	100.0	100.0	98.92	99.46	100.0	100.0	99.46	100.0	97.84	100.0	100.0	100.0	100.0	100.0	65.95	74.05
nl	100.0	100.0	99.46	99.46	88.11	92.43	100.0	97.84	95.68	94.59	81.08	91.89	100.0	100.0	98.92	98.92	32.97	31.89
id	100.0	100.0	99.46	99.46	75.68	94.59	99.46	95.14	99.46	99.46	95.68	99.46	100.0	100.0	99.46	98.38	27.03	30.81
ms	100.0	100.0	97.84	97.84	52.43	70.81	99.46	78.92	94.05	96.22	93.51	99.46	100.0	99.46	97.84	95.68	22.16	33.51
vi	99.46	96.76	99.46	99.46	87.03	99.46	100.0	99.46	98.38	98.38	87.03	96.22	100.0	100.0	99.46	100.0	51.35	69.19
jp	98.92	98.38	91.35	91.89	75.68	87.57	97.3	98.38	94.05	93.51	89.19	95.14	100.0	100.0	93.51	93.51	23.24	50.27
zh	100.0	100.0	97.84	98.38	65.95	64.32	100.0	100.0	97.84	97.3	95.68	98.92	100.0	100.0	87.03	87.03	9.19	5.95
fr	100.0	97.3	100.0	100.0	83.24	89.19	100.0	100.0	99.46	100.0	93.51	100.0	100.0	100.0	100.0	100.0	45.41	38.38
pt	100.0	100.0	100.0	100.0	77.3	81.62	100.0	100.0	100.0	100.0	96.76	100.0	100.0	100.0	99.46	99.46	41.08	32.43
ru	100.0	100.0	95.14	95.14	69.73	88.11	98.92	100.0	92.43	91.35	62.16	92.43	100.0	100.0	95.68	95.68	9.73	14.59
et	100.0	100.0	100.0	100.0	86.49	90.27	99.46	98.92	98.92	98.92	64.86	96.22	100.0	100.0	99.46	99.46	18.92	18.92
it	100.0	100.0	100.0	100.0	85.41	92.97	100.0	100.0	100.0	100.0	91.35	100.0	100.0	100.0	95.68	96.76	45.95	34.59
ta	97.84	93.51	7.03	7.57	4.86	3.24	72.97	52.43	63.78	61.08	56.76	55.14	99.46	98.92	5.95	5.95	0.0	1.08
tr	100.0	100.0	100.0	100.0	72.97	86.49	100.0	95.14	98.38	98.38	87.03	100.0	100.0	100.0	99.46	99.46	24.86	35.14
	LAPE																	
en	100.0	96.22	100.0	100.0	46.49	45.95	100.0	100.0	100.0	100.0	97.84	11.35	100.0	100.0	100.0	100.0	86.49	9.19
nl	100.0	100.0	99.46	99.46	61.62	10.27	100.0	97.84	100.0	100.0	71.89	21.62	100.0	100.0	100.0	100.0	64.32	4.86
id	100.0	100.0	97.84	97.84	11.89	13.51	100.0	99.46	100.0	100.0	61.08	15.68	100.0	100.0	100.0	100.0	58.92	14.59
ms	100.0	100.0	95.14	97.84	12.97	10.27	100.0	100.0	100.0	100.0	28.11	9.19	100.0	100.0	100.0	100.0	25.95	2.16
vi	100.0	100.0	100.0	100.0	8.65	0.0	100.0	100.0	100.0	100.0	20.0	0.54	100.0	100.0	100.0	100.0	19.46	1.08
jp	100.0	100.0	100.0	100.0	10.27	4.86	100.0	100.0	100.0	100.0	23.78	20.54	100.0	100.0	100.0	100.0	4.32	1.08
zh	97.84	85.95	100.0	100.0	20.54	14.59	100.0	100.0	100.0	100.0	51.35	18.92	100.0	100.0	99.46	98.92	61.08	5.95
fr	99.46	99.46	100.0	100.0	51.35	10.27	94.59	94.05	98.92	98.92	73.51	35.14	100.0	100.0	97.84	95.68	77.3	5.41
pt	100.0	100.0	98.92	98.92	78.92	7.03	100.0	100.0	99.46	99.46	62.16	20.54	100.0	100.0	100.0	100.0	83.24	1.62
ru	100.0	100.0	100.0	100.0	37.84	3.78	100.0	100.0	100.0	100.0	60.0	2.16	100.0	100.0	100.0	100.0	11.35	4.32
et	97.3	97.3	100.0	100.0	72.43	47.03	77.3	75.14	90.27	89.73	50.81	34.05	100.0	100.0	99.46	99.46	81.08	42.7
it	100.0	100.0	100.0	100.0	73.51	10.81	96.76	96.22	96.22	95.68	67.03	23.24	100.0	100.0	100.0	100.0	73.51	2.7
ta	99.46	98.92	4.86	4.86	0.0	0.0	91.35	53.51	92.97	90.27	0.54	0.0	100.0	100.0	60.0	56.76	0.0	0.0
tr	100.0	99.46	100.0	100.0	61.62	35.68	76.22	70.81	91.89	91.35	29.73	22.16	100.0	100.0	100.0	100.0	44.86	4.32

Table 15: Language Steering Shift (LSS) scores from th to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	100.0	100.0	91.35	100.0	100.0	100.0	97.84	97.3	29.19	77.84	100.0	100.0	100.0	100.0	79.46	64.86
nl	99.46	98.92	84.86	87.03	69.73	77.3	89.73	81.62	83.78	86.49	24.86	77.3	100.0	99.46	84.86	81.62	60.0	53.51
id	95.68	95.68	83.24	85.95	77.84	79.46	95.14	94.05	90.81	89.19	49.73	87.03	97.3	97.3	76.76	78.92	38.92	34.59
ms	84.32	84.86	75.14	74.05	44.32	54.05	66.49	63.24	69.19	69.19	37.84	70.27	89.73	91.35	65.41	64.86	23.78	28.11
vi	96.76	97.3	86.49	87.03	63.24	67.03	97.3	91.35	82.16	82.7	8.65	47.03	99.46	100.0	79.46	80.0	23.78	28.11
jp	94.59	91.35	58.92	62.16	34.05	36.22	90.81	90.27	78.92	80.0	16.22	37.3	98.92	98.92	52.97	50.27	3.24	2.16
zh	90.27	98.38	48.65	54.59	44.86	38.92	97.84	98.92	80.54	80.0	25.41	41.08	100.0	100.0	41.62	38.38	23.78	3.78
fr	98.92	98.38	92.97	92.43	81.62	84.86	99.46	99.46	90.81	90.81	21.62	85.95	98.92	100.0	91.35	91.89	58.38	56.76
pt	97.84	97.84	89.73	86.49	67.03	77.3	100.0	100.0	95.14	95.14	38.92	87.57	94.05	96.76	90.27	91.89	27.03	34.05
ru	98.38	98.38	64.86	63.78	38.38	38.92	97.3	98.92	78.38	73.51	2.7	18.92	99.46	100.0	55.68	55.14	1.62	4.32
et	97.3	94.05	83.24	83.24	78.92	81.62	88.11	71.89	72.43	68.11	8.65	50.27	99.46	98.38	79.46	82.16	61.08	55.14
it	96.76	95.68	94.05	92.97	78.38	83.78	100.0	100.0	93.51	94.05	39.46	91.35	99.46	100.0	93.51	91.35	58.92	55.68
ta	62.7	57.3	0.0	0.0	0.0	0.0	22.7	11.35	11.35	10.81	0.0	0.0	86.49	83.24	0.0	0.0	0.0	0.0
th	100.0	100.0	38.38	41.62	18.92	11.89	81.08	79.46	27.57	27.57	3.78	7.03	100.0	100.0	13.51	10.81	0.54	1.08
	LAPE																	
en	97.84	95.68	100.0	100.0	61.62	34.05	99.46	99.46	100.0	100.0	99.46	23.24	98.92	97.84	100.0	100.0	88.11	16.22
nl	100.0	100.0	93.51	90.81	72.43	25.95	100.0	98.38	100.0	100.0	45.95	6.49	100.0	100.0	100.0	98.38	70.81	38.38
id	98.38	98.92	88.11	84.32	62.7	44.32	99.46	95.14	99.46	99.46	77.3	63.78	100.0	100.0	96.76	97.3	25.41	16.22
ms	98.38	98.92	72.97	75.14	37.84	27.03	100.0	91.35	99.46	98.92	27.57	17.3	98.92	100.0	93.51	92.97	18.92	12.43
vi	100.0	100.0	93.51	89.73	9.73	4.86	100.0	99.46	99.46	99.46	18.92	10.81	100.0	100.0	99.46	96.76	13.51	4.86
jp	100.0	100.0	81.62	76.76	19.46	2.16	100.0	100.0	89.19	89.19	36.22	21.08	100.0	100.0	60.54	60.0	12.97	5.95
zh	100.0	100.0	61.62	53.51	14.05	7.57	100.0	100.0	98.38	99.46	36.76	8.11	100.0	100.0	75.68	65.95	48.11	4.86
fr	98.92	99.46	95.68	95.14	69.73	33.51	100.0	100.0	100.0	99.46	63.78	50.27	100.0	98.92	97.3	97.3	68.11	8.65
pt	97.84	96.76	95.68	95.14	70.27	25.41	100.0	100.0	100.0	100.0	86.49	58.38	100.0	100.0	96.76	92.97	72.97	4.86
ru	100.0	100.0	94.05	88.65	2.7	1.08	100.0	100.0	100.0	100.0	11.89	0.54	100.0	100.0	92.97	76.22	5.41	0.0
et	97.84	97.3	85.41	81.62	85.95	87.03	97.3	83.24	100.0	100.0	37.3	24.32	97.84	96.22	89.19	89.73	82.7	74.05
it	97.84	97.3	96.76	97.3	57.3	11.89	100.0	97.84	99.46	99.46	68.65	36.22	98.38	97.84	98.92	97.3	68.11	9.73
ta	100.0	100.0	1.08	1.08	0.0	0.0	100.0	96.22	80.54	86.49	0.0	0.0	100.0	100.0	0.0	1.08	0.0	0.0
th	100.0	100.0	74.59	71.89	0.0	0.0	100.0	100.0	93.51	94.05	7.03	5.41	100.0	100.0	51.89	29.73	0.54	0.0

Table 16: Language Steering Shift (LSS) scores from tr to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	99.46	100.0	99.46	99.46	93.51	95.68	100.0	100.0	94.59	94.59	69.19	95.68	100.0	100.0	99.46	99.46	77.84	62.16
nl	100.0	100.0	95.14	96.76	64.32	63.78	88.11	78.92	85.95	84.32	59.46	76.22	100.0	100.0	75.14	79.46	40.54	55.14
id	98.38	100.0	67.57	71.89	58.38	61.62	90.27	87.03	89.73	90.81	78.92	94.05	100.0	100.0	76.76	78.92	51.35	59.46
ms	87.57	94.59	63.24	61.62	37.3	38.38	63.78	56.22	56.76	57.84	53.51	72.97	98.92	100.0	48.65	50.27	28.65	38.38
jp	89.19	89.73	52.43	54.05	30.81	17.84	99.46	98.92	78.92	78.92	44.32	81.62	99.46	100.0	40.54	34.59	17.84	15.14
zh	84.32	94.59	50.81	49.73	38.38	4.32	90.81	98.38	83.24	82.16	51.35	75.14	100.0	100.0	17.3	14.05	22.7	8.11
fr	90.81	89.73	83.78	82.7	70.27	74.59	98.38	98.92	90.81	90.27	70.27	90.81	100.0	99.46	83.24	83.24	42.16	52.97
pt	83.78	80.0	82.16	80.0	60.54	65.41	96.76	97.3	93.51	94.59	72.97	91.35	96.22	96.76	82.7	81.08	27.57	36.22
ru	97.84	99.46	63.24	64.32	44.32	44.86	92.97	95.14	70.81	63.78	28.65	47.57	99.46	100.0	51.35	49.73	4.32	6.49
et	98.92	98.92	68.11	67.57	47.57	58.38	88.11	81.62	80.0	81.08	28.65	54.59	100.0	100.0	77.3	77.84	28.11	33.51
it	94.05	90.81	89.19	89.73	44.86	64.86	98.92	99.46	88.65	87.57	75.68	83.24	95.14	96.22	81.62	83.24	39.46	48.65
ta	71.35	71.35	0.0	0.0	0.0	0.0	16.76	3.24	3.24	2.7	0.0	1.08	76.22	76.22	0.0	0.0	0.0	0.0
th	98.38	97.84	17.3	17.84	11.89	8.65	52.97	66.49	23.24	24.86	5.41	15.68	100.0	100.0	18.92	20.0	2.16	2.16
tr	100.0	97.3	65.95	64.32	45.41	47.03	82.7	70.27	58.92	55.14	61.08	80.54	100.0	100.0	47.57	48.65	36.76	47.03
	LAPE																	
en	100.0	100.0	98.38	98.92	92.43	16.22	100.0	100.0	100.0	100.0	99.46	19.46	100.0	100.0	100.0	100.0	97.3	15.14
nl	100.0	100.0	97.84	96.76	79.46	8.11	100.0	100.0	100.0	100.0	67.57	11.35	100.0	100.0	98.38	98.38	84.32	7.03
id	100.0	100.0	89.19	80.0	92.43	88.11	100.0	99.46	99.46	100.0	48.11	35.68	100.0	100.0	98.92	98.38	74.05	47.03
ms	99.46	99.46	52.43	49.19	86.49	64.32	100.0	99.46	100.0	100.0	31.89	21.08	100.0	100.0	90.81	84.86	33.51	22.7
jp	100.0	100.0	62.7	61.62	9.19	5.95	100.0	100.0	98.38	98.92	24.86	10.27	100.0	100.0	48.11	47.03	15.68	9.19
zh	100.0	100.0	39.46	33.51	20.0	1.62	100.0	100.0	93.51	94.59	52.43	14.05	100.0	100.0	61.62	48.65	43.78	3.78
fr	99.46	98.92	98.92	93.51	70.81	21.08	100.0	100.0	99.46	100.0	55.14	16.76	100.0	100.0	100.0	100.0	60.54	11.35
pt	100.0	100.0	94.59	91.89	54.05	16.76	98.92	98.38	99.46	99.46	78.38	18.38	99.46	99.46	96.22	93.51	71.89	3.78
ru	100.0	100.0	85.41	71.35	6.49	3.24	100.0	100.0	100.0	100.0	19.46	1.62	100.0	100.0	87.03	79.46	2.7	0.0
et	100.0	100.0	89.19	84.86	80.54	25.95	99.46	82.16	100.0	100.0	48.65	10.27	100.0	100.0	90.27	90.81	73.51	11.89
it	100.0	100.0	98.38	97.3	63.78	7.57	100.0	100.0	100.0	100.0	87.57	14.05	100.0	100.0	100.0	99.46	83.24	14.59
ta	100.0	100.0	0.0	0.0	0.0	0.0	100.0	90.81	48.65	50.81	0.0	0.0	100.0	100.0	1.62	1.62	0.0	0.0
th	100.0	100.0	32.43	35.14	4.86	3.24	100.0	100.0	79.46	75.68	3.78	2.7	100.0	100.0	24.32	16.22	1.62	1.08
tr	100.0	100.0	61.62	61.08	83.24	71.89	81.08	74.59	100.0	100.0	60.0	52.43	100.0	100.0	67.57	68.11	41.62	32.97

Table 17: Language Steering Shift (LSS) scores from vi to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

	Qwen						Gemma						SeaLLM					
	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓	pmax↑	pmed↑	=max↑	+max↑	=0↓	=10p↓
	Baseline																	
en	100.0	100.0	100.0	100.0	84.86	96.76	100.0	100.0	95.14	95.14	51.35	71.35	100.0	100.0	100.0	100.0	58.92	43.24
nl	100.0	100.0	96.76	95.14	38.92	67.57	90.27	92.43	81.08	79.46	35.68	42.16	100.0	100.0	93.51	95.14	5.95	5.41
id	100.0	100.0	95.68	93.51	31.35	56.22	97.3	96.22	95.68	95.68	48.11	75.14	100.0	100.0	91.35	91.35	11.89	5.95
ms	100.0	99.46	77.84	76.22	24.86	45.41	76.76	71.89	70.27	69.73	27.03	54.59	100.0	100.0	69.19	67.57	5.95	4.86
vi	100.0	100.0	98.38	97.84	42.7	85.41	91.89	82.7	67.57	70.81	10.81	17.84	100.0	100.0	88.11	94.05	9.19	6.49
jp	89.19	90.27	72.43	73.51	47.57	58.38	79.46	79.46	63.78	61.08	35.14	63.24	91.35	92.43	65.95	67.03	25.95	26.49
fr	100.0	100.0	93.51	93.51	65.95	73.51	97.3	100.0	90.27	90.27	41.62	59.46	100.0	100.0	94.05	93.51	16.22	9.73
pt	100.0	100.0	98.92	99.46	33.51	65.41	98.92	100.0	92.97	92.97	52.97	67.57	100.0	100.0	96.22	96.22	5.41	3.78
ru	99.46	99.46	76.22	77.84	35.68	46.49	85.41	90.27	62.7	60.0	23.78	32.43	99.46	100.0	80.54	82.16	3.24	3.24
et	100.0	100.0	99.46	98.92	37.84	52.97	89.19	75.14	71.89	67.03	9.19	21.62	100.0	100.0	88.11	91.35	2.7	1.08
it	100.0	99.46	96.76	95.68	51.35	74.05	97.84	100.0	91.35	90.27	36.76	79.46	100.0	100.0	89.73	89.73	21.62	9.19
ta	85.95	85.41	3.24	3.24	0.0	0.0	7.57	3.24	3.78	3.78	0.0	1.62	82.16	69.19	0.0	0.0	0.0	0.0
th	100.0	98.92	42.7	43.78	14.05	19.46	52.43	52.43	14.05	14.59	2.7	7.57	100.0	100.0	32.43	34.05	1.62	1.08
tr	100.0	100.0	78.38	78.92	23.24	48.11	89.19	82.16	64.32	64.86	27.03	35.14	100.0	100.0	70.81	69.19	4.32	3.24
	LAPE																	
en	63.24	72.43	100.0	100.0	27.03	29.73	100.0	99.46	100.0	100.0	90.81	56.76	61.62	65.41	90.81	93.51	45.95	35.68
nl	100.0	100.0	92.97	94.05	40.54	9.73	99.46	89.73	100.0	100.0	46.49	11.35	100.0	100.0	97.84	97.3	57.3	8.65
id	96.76	97.3	87.57	92.97	83.78	32.43	100.0	92.43	99.46	99.46	42.16	9.19	100.0	100.0	95.68	94.05	67.57	7.57
ms	95.68	96.22	65.41	65.95	47.57	23.24	100.0	92.43	100.0	100.0	28.11	7.03	98.38	99.46	70.27	69.19	30.81	3.24
vi	96.22	92.97	97.3	95.68	27.57	18.38	97.3	89.73	94.59	93.51	15.14	7.57	98.38	95.68	96.76	93.51	14.59	5.95
jp	91.35	90.27	76.76	76.22	30.27	31.35	98.92	92.97	84.32	87.03	45.41	41.08	91.89	90.81	60.54	50.27	13.51	10.27
fr	92.43	91.89	95.68	91.35	60.54	27.03	98.92	98.92	100.0	100.0	65.95	41.08	98.38	98.92	94.59	92.97	50.27	8.65
pt	92.43	91.89	98.38	95.68	47.03	9.19	99.46	92.97	100.0	100.0	71.35	49.73	96.76	99.46	94.05	94.05	63.24	10.81
ru	98.92	98.92	97.3	94.59	2.16	2.16	94.59	91.35	96.22	97.3	23.78	14.05	98.38	99.46	88.65	85.95	8.11	7.03
et	98.92	98.92	97.84	94.05	69.73	28.65	98.38	89.73	100.0	99.46	31.89	4.32	99.46	98.92	94.59	94.05	74.05	31.89
it	95.14	95.68	95.14	97.3	63.24	16.22	99.46	98.38	100.0	100.0	58.92	42.16	99.46	99.46	94.59	93.51	65.95	7.03
ta	100.0	99.46	1.08	1.08	0.0	0.0	62.7	10.81	18.92	24.86	0.0	0.0	95.68	97.84	0.54	0.54	0.0	0.0
th	100.0	100.0	64.86	82.16	1.08	0.54	91.89	65.41	48.11	45.95	3.78	3.24	100.0	100.0	29.73	27.03	0.54	0.54
tr	99.46	100.0	68.11	67.57	30.81	17.3	87.03	85.41	99.46	97.3	47.03	30.81	96.76	98.38	71.89	73.51	29.73	15.68

Table 18: Language Steering Shift (LSS) scores from zh to languages listed for different steering factors across models. Bolded texts denote the highest values within a model.

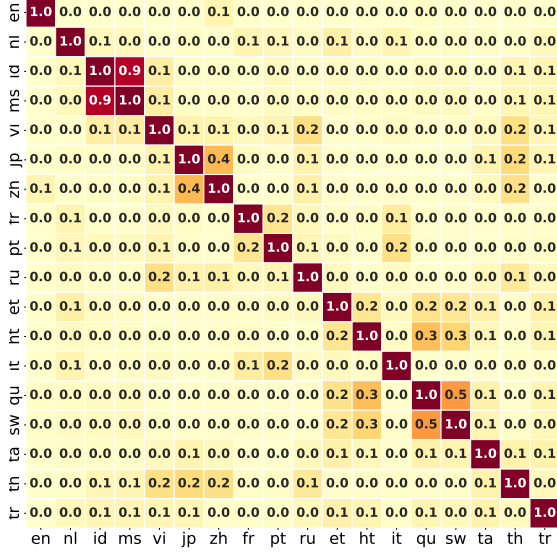


Figure 13: Symmetric heatmaps of overlap neuron proportions in Qwen2.5 0.5B using jaccard distance for LAPE neurons.

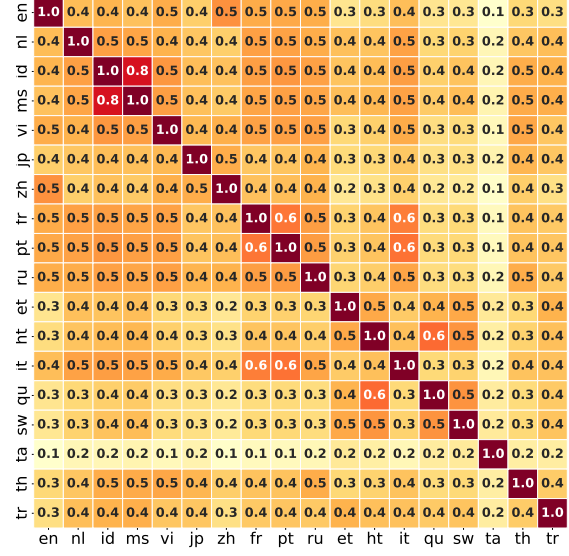


Figure 15: Symmetric heatmaps of overlap neuron proportions in SeaLLMs3 1.5B using jaccard distance for Baseline neurons.

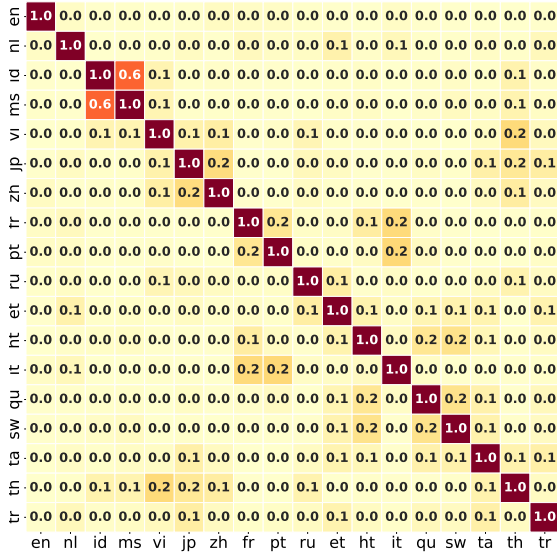
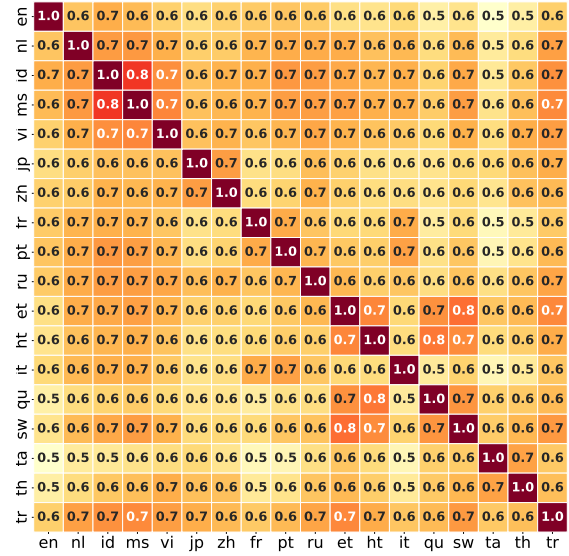


Figure 14: Symmetric heatmaps of overlap neuron proportions in Qwen2.5 7B using jaccard distance for LAPE neurons.



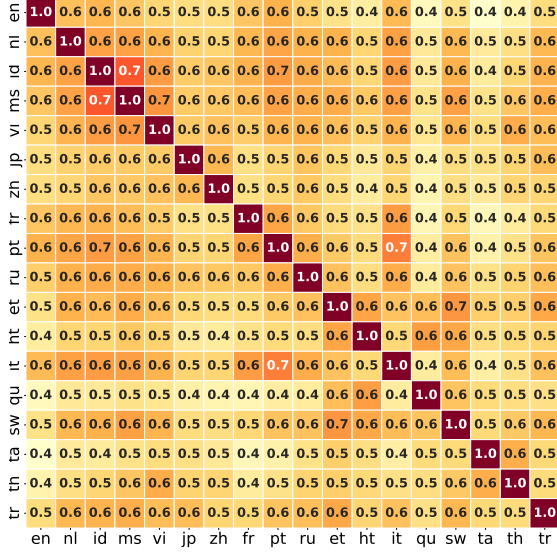


Figure 17: Symmetric heatmaps of overlap neuron proportions in Gemma2 9B using jaccard distance for Baseline neurons.

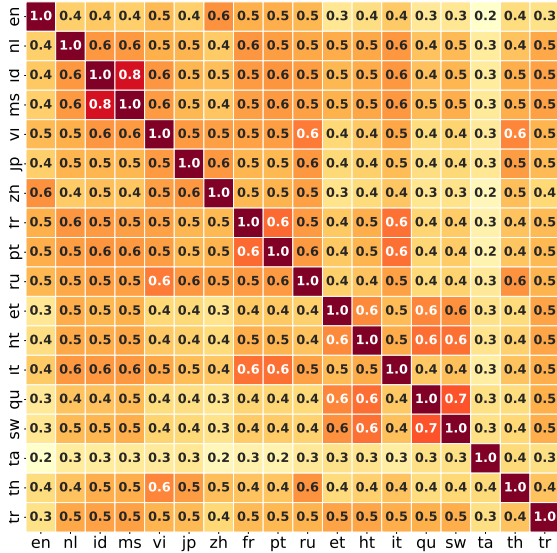


Figure 18: Symmetric heatmaps of overlap neuron proportions in Qwen2.5 0.5B using jaccard distance for Baseline neurons.

Factor 1	Factor 2	Δ Mean	p -value	Sig.
pmax	pmed	-0.0109	0.9813	No
pmax	=max	-0.1789	0.0000	Yes
pmax	+max	-0.1946	0.0000	Yes
pmed	=max	-0.1680	0.0000	Yes
pmed	+max	-0.1837	0.0000	Yes
=max	+max	-0.0157	0.9117	No

Table 19: Tukey HSD pairwise comparisons between LSS scores of steering factors on LAPE neurons. *Sig.* denotes the significance of Factor2 compared to Factor1.

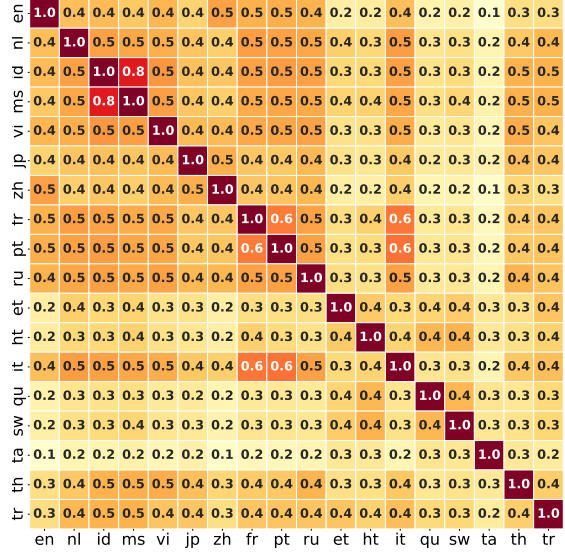


Figure 19: Symmetric heatmaps of overlap neuron proportions in Qwen2.5 7B using jaccard distance for Baseline neurons.

Lang	SeaLLMs3-1.5B	Qwen2.5-0.5B	Gemma2-2B
et	0.87 → 1.50	0.49 → 0.68	0.45 → 0.46
fr	1.43 → 18.49	0.52 → 12.94	0.76 → 2.96
ht	1.01 → 0.12	0.55 → 0.89	0.42 → 0.45
id	1.92 → 15.95	0.65 → 5.52	0.55 → 0.55
it	1.15 → 10.53	0.63 → 1.26	0.48 → 1.69
jp	0.11 → 0.26	0.06 → 0.15	0.02 → 0.02
ms	1.54 → 9.70	0.61 → 3.83	0.49 → 0.60
nl	1.20 → 2.25	0.55 → 2.79	0.45 → 0.62
pt	1.47 → 15.57	0.61 → 8.36	0.61 → 2.42
qu	0.93 → 0.24	0.46 → 0.76	0.44 → 0.46
ru	0.75 → 10.06	5.22 → 4.76	0.38 → 0.48
sw	1.07 → 0.17	0.51 → 0.89	0.45 → 0.46
ta	0.64 → 0.51	0.39 → 0.18	0.40 → 0.38
th	0.56 → 2.02	0.25 → 1.44	0.17 → 0.23
tr	1.02 → 3.44	0.48 → 1.62	0.45 → 0.44
vi	1.65 → 18.38	0.52 → 10.72	0.33 → 0.66
zh	0.32 → 0.33	0.20 → 0.22	0.09 → 0.11

Table 20: BLEU score changes of self-language intervention with **pmedian** from en to target listed on FLORES-200 dataset. The translation prompts are *non-targeted*.

Lang	Qwen2.5-7B	Qwen2.5-0.5B	SeaLLMs3-7B	SeaLLMs3-1.5B	Gemma2-2B	Gemma2-9b
et	0.76 → 2.54	0.49 → 0.89	0.43 → 2.99	0.87 → 1.87	0.45 → 0.52	0.24 → 0.23
fr	13.75 → 28.15	0.52 → 12.90	0.32 → 25.01	1.43 → 19.82	0.76 → 9.08	0.57 → 4.54
ht	0.92 → 1.68	0.55 → 1.02	0.34 → 1.35	1.01 → 0.89	0.42 → 0.50	0.24 → 0.26
id	1.13 → 25.01	0.65 → 6.37	0.52 → 26.88	1.92 → 15.99	0.55 → 1.64	0.29 → 0.50
it	1.16 → 14.60	0.63 → 2.37	0.44 → 15.99	1.15 → 10.77	0.48 → 4.14	0.25 → 0.45
jp	0.09 → 0.15	0.06 → 0.15	0.08 → 0.14	0.11 → 0.35	0.02 → 0.11	0.03 → 0.01
ms	1.01 → 0.63	0.61 → 3.93	0.50 → 19.14	1.54 → 10.59	0.49 → 1.18	0.31 → 0.39
nl	0.85 → 10.69	0.55 → 2.86	0.41 → 11.42	1.20 → 6.56	0.45 → 4.95	0.24 → 0.30
pt	1.33 → 25.29	0.61 → 10.06	0.58 → 19.05	1.47 → 16.80	0.61 → 10.72	0.44 → 12.90
qu	0.75 → 1.12	0.46 → 0.76	0.63 → 0.74	0.93 → 0.92	0.44 → 0.56	0.34 → 0.50
ru	2.20 → 16.23	5.22 → 4.28	4.77 → 16.36	0.75 → 10.37	0.38 → 0.35	0.24 → 1.85
sw	0.78 → 1.79	0.51 → 0.86	0.36 → 1.95	1.07 → 1.02	0.45 → 0.48	0.26 → 0.20
ta	0.45 → 0.22	0.39 → 0.22	0.29 → 0.41	0.64 → 0.43	0.40 → 0.39	0.23 → 0.13
th	0.35 → 2.53	0.25 → 1.03	0.43 → 2.96	0.56 → 2.09	0.17 → 0.18	0.13 → 0.12
tr	0.77 → 7.73	0.48 → 1.19	0.43 → 5.72	1.02 → 3.38	0.45 → 0.51	0.28 → 0.34
vi	0.74 → 26.00	0.52 → 10.11	0.36 → 21.76	1.65 → 17.31	0.33 → 1.10	0.16 → 4.57
zh	0.19 → 0.47	0.20 → 0.17	0.24 → 1.18	0.32 → 0.22	0.09 → 0.09	0.09 → 0.12

Table 21: BLEU score changes of self-language intervention from en to targets listed on FLORES-200 dataset. The translation prompts are *non-targeted*.

Lang	SeaLLMs3-1.5B	Qwen2.5-0.5B	Gemma2-2B
et	2.86 → 0.35	1.46 → 0.18	0.45 → 0.01
fr	23.58 → 0.16	14.02 → 0.13	0.76 → 0.15
ht	1.28 → 0.12	0.82 → 0.09	0.42 → 0.13
id	22.06 → 0.14	7.93 → 0.03	0.55 → 0.14
it	12.44 → 0.18	6.51 → 0.09	0.48 → 0.09
jp	0.13 → 0.00	0.07 → 0.01	0.02 → 0.00
ms	17.05 → 0.13	5.04 → 0.08	0.49 → 0.14
nl	10.72 → 0.23	4.77 → 0.12	0.45 → 0.12
pt	21.99 → 0.24	13.76 → 0.07	0.61 → 0.14
qu	1.18 → 0.01	0.39 → 0.05	0.44 → 0.06
ru	13.48 → 0.32	6.96 → 0.09	0.38 → 0.17
sw	1.67 → 0.14	0.61 → 0.00	0.45 → 0.06
ta	0.33 → 0.00	0.12 → 0.00	0.40 → 0.02
th	3.59 → 0.00	1.54 → 0.03	0.17 → 0.00
tr	4.56 → 0.15	2.21 → 0.05	0.45 → 0.14
vi	22.37 → 1.42	12.74 → 0.06	0.33 → 0.14
zh	0.66 → 0.00	0.12 → 0.09	0.09 → 0.02

Table 22: BLEU score changes of self-language **pmax** intervention on Baseline neurons from en to target listed on FLORES-200 dataset. The translation prompts are *targeted*.

Lang	Gemma2		SeaLLMs3		Qwen2.5	
	9B	2B	7B	1.5B	7B	0.5B
en	70.58	68.34	86.88	82.67	85.51	75.96
fr	65.06	56.63	66.27	77.11	72.29	61.45
jp	66.11	59.02	68.82	64.44	72.99	57.14
pt	66.54	61.60	77.57	68.44	79.85	62.36
ru	71.11	65.08	73.02	67.62	70.16	58.10
zh	70.24	66.27	80.95	76.39	81.75	64.88

Lang	Gemma2		SeaLLMs3		Qwen2.5	
	9B	2B	7B	1.5B	7B	0.5B
et	69.4	33.3	38.8	23.5	42.1	32.8
fr	66.4	47.6	53.6	40.0	57.6	41.6
id	65.3	51.6	57.3	50.4	59.7	41.1
it	70.9	54.6	61.8	51.8	72.9	41.4
jp	71.9	53.8	63.9	55.0	69.1	48.2
ms	59.4	44.6	49.0	38.6	53.0	33.7
nl	72.1	54.2	60.6	44.6	64.1	32.7
pt	63.6	47.8	56.9	42.3	61.7	38.3
ru	55.2	45.6	52.8	42.5	55.6	40.1
ta	45.6	32.8	30.0	33.6	32.4	28.0
tr	57.0	39.8	45.8	31.7	51.0	30.5
vi	62.4	46.0	58.0	43.6	56.8	40.8
zh	55.9	45.7	73.5	73.1	77.9	55.5

Table 23: Baseline performance of XWinograd across models.

Lang	Gemma2		SeaLLMs3		Qwen2.5	
	9B	2B	7B	1.5B	7B	0.5B
et	57.8	49.4	50.8	51.2	50.4	48.8
ht	52.4	46.8	50.4	51.4	51.8	51.2
id	69.4	62.0	73.6	65.4	73.4	57.2
it	69.6	63.6	72.0	62.0	76.6	52.2
qu	51.0	50.2	50.0	51.2	49.0	50.4
sw	61.8	54.0	53.4	53.4	53.4	54.4
ta	57.8	54.2	57.6	53.2	56.2	56.6
th	56.8	53.0	59.0	57.4	61.6	55.6
tr	64.2	54.4	61.2	54.8	61.4	53.6
vi	67.6	59.2	74.4	65.8	78.0	58.2
zh	72.6	67.2	78.0	70.0	79.8	63.6

Table 24: Baseline performance of XCOPA across models.

Table 25: Baseline performance of Include-lite across models.

Factor	Mean LSS
pmax	0.9204
pmedian	0.9095
=max	0.7415
+max	0.7258
=0	0.4109
=10p	0.1640

Table 26: Mean LSS throughout all languages and models for each steering factor.

Lang	Gemma2		SeaLLMs3		Qwen2.5	
	9B	2B	7B	1.5B	7B	0.5B
en	9	65	49	97	30	114
nl	61	223	352	672	126	238
id	43	126	108	382	56	137
ms	48	178	131	377	71	203
vi	52	214	216	435	86	224
jp	139	495	351	690	309	573
zh	72	362	299	887	196	436
fr	42	204	343	897	109	220
pt	33	189	354	633	75	195
ru	64	262	226	444	135	256
et	156	638	176	504	355	806
ht	71	968	73	383	210	832
it	45	185	327	637	107	203
qu	172	1644	110	434	333	1426
sw	177	1231	184	551	435	1297
ta	550	1127	325	457	1085	1053
th	155	311	267	507	164	292
tr	132	286	159	668	219	355

Table 27: Proportions of LAPE neurons obtained from FLORES-200 dataset.

Lang	Gemma2		SeaLLMs3		Qwen2.5	
	9B	2B	7B	1.5B	7B	0.5B
en	6	3995	8450	13838	2759	3653
nl	32	7051	10085	17462	5214	6570
id	22	6730	9724	16126	4864	6318
ms	27	7393	10188	17815	5324	6847
vi	19	6271	10828	20036	4303	5878
jp	62	7119	11032	19553	4996	6624
zh	38	4268	10246	19029	2935	3799
fr	29	6592	10608	17814	4683	6153
pt	21	6407	10692	17069	4547	6036
ru	32	5903	10554	17496	4231	5420
et	58	12607	11542	19546	7718	11544
ht	16	12497	11896	21920	6583	11363
it	28	7173	10945	17064	5352	6615
qu	35	14061	13129	25239	7579	12479
sw	47	14225	11856	19813	8511	13050
ta	120	21817	14465	27403	19088	19088
th	33	8352	12832	24646	5573	7692
tr	68	9139	10740	19227	6905	8502

Table 28: Proportions of baseline neurons obtained from FLORES-200 dataset.

en	0.13	-7.79	-8.05	-5.68	-2.15	-3.06
fr	2.41	0.00	1.20	-8.44	-6.03	-1.21
jp	-6.67	-1.77	0.11	-4.27	1.78	-4.79
pt	-8.37	-0.76	-6.09	-1.14	-0.76	-7.99
ru	-3.81	-2.86	-6.04	-4.45	-1.91	-3.50
zh	-2.18	-4.17	-5.75	-4.17	1.79	2.58
	en	fr	jp	pt	ru	zh

Figure 20: Delta XWinograd accuracy after steering LAPE neurons (obtained from the XWinograd dataset) with **pmax** for Qwen2.5 0.5B.

en	0.17	-1.12	-4.26	-1.16	-13.47	-9.55
fr	-19.28	-6.03	-22.89	-10.84	-12.05	-24.10
jp	-8.13	-0.73	1.04	-0.31	0.94	-4.69
pt	-7.98	-4.94	-7.22	0.76	-1.52	-14.07
ru	-6.99	-4.45	-6.35	-6.67	-0.32	-10.79
zh	-0.20	-0.79	-0.99	0.20	-0.60	-0.20
	en	fr	jp	pt	ru	zh

(a) SeaLLMv3 1.5B

en	-0.04	-0.77	-16.21	-0.47	-13.80	-9.72
fr	-6.03	4.81	-7.23	-3.62	-2.41	-13.26
jp	-1.98	3.55	2.30	-0.62	3.65	-3.44
pt	-19.40	-7.23	-15.21	0.00	-7.99	-20.16
ru	-15.24	0.63	-6.99	-2.54	0.95	-11.75
zh	-11.70	-0.99	-3.37	-2.78	-4.16	-0.39
	en	fr	jp	pt	ru	zh

(b) SeaLLMv3 7B

Figure 21: Delta XWinograd accuracy after steering LAPE neurons with **pmax** for SeaLLMv3 1.5B and SeaLLMv3 7B. Row i is the initial language, column j is the language of intervention.

en	-0.13	-0.56	-8.22	-0.86	-2.02	-5.04
fr	-3.62	-1.21	-4.82	2.41	-13.26	-8.44
jp	-5.32	-0.41	0.84	-3.33	1.36	-4.48
pt	-8.37	0.00	-6.85	1.14	-1.14	-9.51
ru	-7.31	-2.54	-4.77	-0.64	-1.59	-6.04
zh	-0.59	-0.79	0.40	-0.79	4.17	1.99
	en	fr	jp	pt	ru	zh

(a) Qwen2.5 0.5B

en	0.04	-0.09	-10.16	-1.04	-11.70	-6.84
fr	-9.64	-1.21	-8.43	-10.84	-8.43	-15.66
jp	-10.95	-1.67	0.42	-0.83	2.61	-8.96
pt	-20.15	-4.56	-11.41	0.76	-6.85	-14.45
ru	-10.79	-2.54	-6.67	-0.95	0.63	-9.53
zh	-13.10	-1.39	-5.96	-0.80	-3.77	-0.80
	en	fr	jp	pt	ru	zh

(b) Qwen2.5 7B

Figure 22: Delta XWinograd accuracy after steering LAPE neurons with **pmax** for Qwen2.5 0.5B and Qwen2.5 7B. Row i is the initial language, column j is the language of intervention.

en	-0.17	-0.17	-0.30	-0.56	1.29	0.78
fr	-1.21	1.20	-2.41	0.00	4.82	-1.21
jp	-6.57	-3.23	-0.52	-5.42	-2.09	-1.36
pt	-5.71	-3.04	-4.57	0.00	-3.81	-1.90
ru	-8.57	-6.67	-6.67	-9.21	0.00	-4.76
zh	-9.92	-2.98	-3.77	-5.56	-6.55	0.40
	en	fr	jp	pt	ru	zh

(a) Gemma2 2B

en	-0.09	-0.09	0.00	-0.04	1.16	0.52
fr	-6.02	1.21	-2.41	-6.02	-1.20	-7.23
jp	-13.24	-9.49	-1.04	-10.64	-4.69	-4.27
pt	-5.70	-2.66	-5.70	2.28	-3.42	-6.46
ru	-13.01	-7.62	-10.79	-8.57	-0.63	-11.11
zh	-12.30	-5.95	-4.57	-3.18	0.99	0.00
	en	fr	jp	pt	ru	zh

(b) Gemma2 9B

Figure 23: Delta XWinograd accuracy after steering LAPE neurons with **pmax** for Gemma2 2B and 9B. Row i is the initial language, column j is the language of intervention.

en	-0.08	-0.34	0.31	-0.34	0.78	0.86
fr	2.41	0.00	-2.41	-2.41	3.61	0.00
jp	-6.26	-0.42	-1.36	-2.40	-0.83	-0.52
pt	-4.95	-2.66	-4.57	0.00	-2.28	-0.38
ru	-7.30	-0.32	-1.59	-5.08	-0.32	-0.64
zh	-8.53	-0.99	-3.37	0.00	0.00	-0.20
	en	fr	jp	pt	ru	zh

Figure 24: Delta XWinograd accuracy after steering LAPE neurons with **pmedian** for Gemma2 2B. Row i is the initial language, column j is the language of intervention.

en	0.08	0.21	-5.94	-0.82	-0.18	-6.11
fr	-1.21	1.20	-6.03	2.41	-6.03	-2.41
jp	-4.59	0.21	-0.21	-1.56	0.52	-1.04
pt	-7.23	0.00	-6.47	-1.14	0.38	-5.71
ru	-4.77	0.95	-3.50	1.58	-2.23	-7.31
zh	-0.20	-0.99	0.99	-0.40	3.37	1.59
	en	fr	jp	pt	ru	zh

Figure 25: Delta XWinograd accuracy after steering LAPE neurons with **pmedian** for Qwen2.5 0.5B. Row i is the initial language, column j is the language of intervention.

en	-0.13	0.09	0.26	0.13	0.82	0.60
fr	-6.02	-1.20	2.41	-2.41	-3.61	1.21
jp	-10.7	-2.81	-0.73	-0.94	-0.83	-2.08
pt	-4.94	0.38	-2.66	4.18	0.00	-6.08
ru	-12.7	-0.63	-2.86	-0.63	-0.63	-8.25
zh	-11.1	-0.20	-1.39	0.59	1.19	1.19
	en	fr	jp	pt	ru	zh

Figure 27: Delta XWinograd accuracy after steering LAPE neurons with **pmedian** for Gemma2 9B. Row i is the initial language, column j is the language of intervention.

en	0.08	-0.73	-3.44	-0.56	-5.59	-10.1
fr	-21.7	-4.82	-25.3	-12.0	-14.5	-16.9
jp	-4.90	-0.31	0.84	-0.31	-0.52	-3.54
pt	-8.74	-4.56	-5.32	-0.38	-3.42	-13.3
ru	-8.89	-2.54	-4.45	-0.32	0.00	-8.25
zh	-0.40	-0.60	-0.60	0.40	0.00	-0.60
	en	fr	jp	pt	ru	zh

Figure 26: Delta XWinograd accuracy after steering LAPE neurons with **pmedian** for SeaLLMs 1.5B. Row i is the initial language, column j is the language of intervention.

en	-0.26	-0.26	-9.64	-0.48	-11.6	-6.20
fr	-7.23	-3.62	-6.02	-3.62	-4.82	-14.5
jp	-3.23	-1.35	0.42	-0.73	2.30	1.46
pt	-15.6	-2.66	-9.13	-1.90	-3.04	-12.9
ru	-5.72	0.63	-6.35	-0.32	1.27	-6.99
zh	-11.1	-1.19	-3.77	-0.40	-3.97	-0.20
	en	fr	jp	pt	ru	zh

Figure 28: Delta XWinograd accuracy after steering LAPE neurons with **pmedian** for Qwen2.5 7B. Row i is the initial language, column j is the language of intervention.

en	-5.94	-15.4	-18.2	-17.3	-11.2	-7.44
fr	-27.7	-14.5	-20.5	-22.9	-18.1	-19.3
jp	-14.5	-13.3	-4.69	-11.9	-6.05	-6.88
pt	-17.5	-13.7	-15.2	-12.5	-14.4	-12.9
ru	-11.1	-13.3	-13.7	-12.4	-6.03	-11.4
zh	-16.3	-22.0	-14.7	-18.3	-13.9	-8.93
	en	fr	jp	pt	ru	zh

Figure 29: Delta XWinograd accuracy after steering baseline neurons with **pmax** for SeaLLMv3 1.5B. Row i is the initial language, column j is the language of intervention.

en	-13.1	-24.5	-23.8	-21.8	-21.3	-12.8
fr	-8.44	-7.23	-8.44	-8.44	-6.03	-18.1
jp	-6.36	-7.61	-3.23	-6.57	-5.42	-8.03
pt	-15.2	-12.2	-12.5	-11.4	-9.89	-12.2
ru	-3.81	-6.35	-6.04	-4.77	-6.99	-6.04
zh	-8.73	-12.9	-14.3	-12.7	-11.3	-9.13
	en	fr	jp	pt	ru	zh

Figure 30: Delta XWinograd accuracy after steering baseline neurons with **pmax** for Qwen2.5 0.5B. Row i is the initial language, column j is the language of intervention.

en	-15.9	-18.1	-16.6	-17.5	-18.0	-16.4
fr	-7.23	-6.03	-9.64	-7.23	-8.44	-7.23
jp	-10.7	-9.38	-8.03	-10.5	-8.66	-9.28
pt	-9.51	-11.4	-10.3	-11.0	-10.3	-12.6
ru	-14.9	-14.0	-12.4	-18.4	-18.7	-15.6
zh	-16.3	-16.5	-14.5	-15.9	-16.3	-13.7
	en	fr	jp	pt	ru	zh

Figure 31: Delta XWinograd accuracy after steering baseline neurons with **pmax** for Gemma2 2B. Row i is the initial language, column j is the language of intervention.

et	2.20	-1.40	-1.40	-3.80	-4.20	-1.40	-2.80	-4.60	-5.20	-4.20	-4.60
ht	-2.40	0.60	-1.80	-0.60	2.20	-1.00	-2.80	-2.20	-3.80	-0.80	-4.00
id	0.40	0.00	0.20	-1.00	-2.00	1.00	-3.40	-2.20	-5.00	-4.40	-4.80
it	-0.80	-1.20	0.00	0.60	-0.60	-1.40	-2.00	-3.60	-4.40	-2.60	-8.80
qu	-2.20	-1.00	-1.60	-1.40	-1.00	-0.40	-1.60	-2.40	-1.20	-2.40	-4.80
sw	-3.20	-0.60	-1.60	-0.20	-0.20	-0.80	-2.00	-5.80	-3.40	-6.80	-7.40
ta	-5.00	-2.80	-2.80	-5.20	-4.20	-5.40	-0.80	-0.20	-6.40	-4.00	-5.80
th	-0.60	-0.20	-1.60	-2.80	-2.00	-0.80	-2.60	1.00	-0.20	-2.80	0.60
tr	-1.40	0.20	-3.80	-3.20	-0.80	-0.80	-0.40	-1.20	-1.00	-3.00	-3.00
vi	-0.40	-0.80	-1.40	-0.40	0.00	0.60	-2.40	0.00	-4.00	1.40	-4.40
zh	0.00	-3.00	-0.40	-2.80	-2.60	-1.20	-0.40	-0.40	-1.20	-0.80	-1.00
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 32: Delta XCOPA accuracy after steering LAPE neurons with **pmax** for Gemma2 9B.

et	0.20	0.60	0.00	0.80	0.80	-0.60	1.00	0.60	0.00	0.00	2.80
ht	-1.20	0.20	-1.00	-0.40	0.20	0.60	-1.20	0.20	1.60	-1.60	-1.80
id	-6.40	-4.60	-1.00	-5.40	-5.40	-1.60	-8.60	-8.40	-3.40	-5.60	-8.20
it	-5.00	-4.40	-2.00	1.80	-7.00	-2.20	-2.20	-1.20	-2.80	-2.00	-0.20
qu	0.20	0.20	1.40	0.20	0.40	0.00	0.60	1.20	1.60	-0.20	-0.60
sw	-1.40	-0.60	-0.80	0.00	-0.80	0.00	-0.60	0.00	-2.20	-1.40	1.80
ta	-1.60	-2.00	-2.20	-0.20	-2.60	0.20	-1.80	-2.40	-1.20	-0.60	0.20
th	-1.40	-1.80	-4.00	-1.20	-2.00	-3.20	-0.60	-0.40	-3.20	-0.40	-2.60
tr	0.80	-0.60	-2.20	0.20	0.40	0.20	1.00	0.80	-0.20	-1.40	0.80
vi	-4.60	-3.20	-6.20	-4.20	-4.80	-4.20	-4.60	-5.80	-5.40	0.20	-5.60
zh	-1.80	-0.60	-2.40	-0.20	0.40	-2.60	-2.00	-1.80	-3.20	-3.60	0.20
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 34: Delta XCOPA accuracy after steering LAPE neurons with **pmax** for Qwen2.5 0.5B.

et	0.60	-1.20	0.20	-1.40	-1.80	-0.20	1.40	1.20	2.00	-0.60	1.40
ht	1.40	0.60	2.00	0.60	1.20	0.00	-1.00	1.40	-1.40	0.40	-0.80
id	-1.40	-1.40	-0.20	-8.20	-3.40	-3.00	-1.60	-0.20	-1.40	-3.60	-2.60
it	-4.20	-1.80	-3.80	0.00	-3.20	-4.00	-6.60	-4.80	-5.00	-7.80	-7.60
qu	0.20	0.60	1.40	1.00	0.80	-1.60	-1.60	-1.00	-0.20	-0.80	0.80
sw	1.00	0.20	0.40	-0.80	1.60	1.60	0.20	-1.20	-1.60	-1.20	-3.60
ta	-1.80	-5.60	-1.60	-5.40	-4.00	-4.80	-1.00	-0.60	-0.60	-3.00	-2.00
th	-0.40	2.20	0.60	2.20	2.60	0.20	-0.60	-0.80	0.80	-0.60	1.40
tr	-2.60	-2.00	-2.20	-1.80	-1.60	-2.20	2.00	1.00	-0.80	-1.80	-1.20
vi	-0.80	-0.80	-2.80	-7.40	-2.40	-3.20	-0.40	2.60	-1.00	-0.20	-4.20
zh	-3.20	-3.40	-0.80	-4.40	-3.80	-1.80	-2.60	-0.80	-3.40	-2.60	0.20
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 33: Delta XCOPA accuracy after steering LAPE neurons with **pmax** for Gemma2 2B.

et	0.20	0.40	-0.40	-2.60	-1.20	-1.60	-0.40	1.00	-2.00	-1.00	1.60
ht	-1.40	0.00	-1.40	-1.60	0.40	-3.60	-2.00	-0.20	-0.80	0.00	-1.80
id	-16.0	-13.4	-0.20	-9.20	-11.8	-10.4	-14.6	-19.0	-15.6	-18.6	-21.6
it	-18.4	-17.6	-15.0	-0.40	-15.8	-18.0	-15.4	-23.0	-20.8	-22.2	-24.4
qu	1.60	2.00	2.80	-1.60	1.00	3.60	-0.20	2.60	1.60	0.80	1.20
sw	1.60	0.60	-1.40	-0.60	0.40	-0.40	0.60	-1.60	-2.00	0.00	-0.40
ta	1.00	0.40	0.40	-1.40	-1.20	-1.00	-0.60	0.20	-0.60	-1.20	0.00
th	-6.20	-6.20	-5.00	-5.80	-7.00	-5.60	-5.40	-1.60	-3.20	-4.00	-7.40
tr	-7.00	-6.60	-2.60	-6.40	-4.40	-4.60	-3.40	-3.80	-0.80	-6.00	-5.40
vi	-16.8	-14.4	-16.4	-9.20	-12.4	-12.6	-16.0	-19.2	-16.6	-0.40	-25.2
zh	-6.40	-5.60	-1.40	-1.00	-4.60	-4.80	-4.00	-2.20	-9.00	-2.60	0.80
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 35: Delta XCOPA accuracy after steering LAPE neurons with **pmax** for Qwen2.5 7B.

et	-1.00	-1.00	-0.40	-0.80	0.40	1.40	-1.20	-2.00	-1.40	0.40	0.20
ht	-1.40	0.00	2.80	-0.80	0.60	-0.20	-2.20	0.80	1.20	-1.40	0.40
id	-11.2	-6.80	0.20	-8.00	-6.20	-6.80	-14.8	-14.8	-11.2	-12.2	-14.6
it	-12.0	-8.60	-7.40	-0.80	-5.20	-8.00	-8.80	-10.6	-11.4	-9.40	-11.8
qu	-1.40	0.20	0.40	0.40	0.00	0.60	-1.40	-1.20	0.20	0.20	-0.40
sw	-0.40	0.40	0.40	-0.20	-0.80	-1.80	-1.20	-0.20	-1.40	-0.40	-1.00
ta	0.40	2.20	-0.20	3.00	1.40	1.40	-0.20	2.20	0.80	1.40	1.40
th	0.20	-2.00	-2.80	-2.40	-0.80	-2.00	0.20	0.00	-3.40	-4.00	-4.20
tr	0.40	-4.60	-3.00	-3.40	-3.40	-1.60	0.80	-0.40	0.40	0.00	-2.00
vi	-9.60	-5.80	-12.4	-4.80	-5.00	-9.00	-12.2	-12.4	-12.2	0.40	-14.4
zh	-1.20	-0.80	0.00	-3.40	-1.80	-2.00	-2.60	0.40	-2.60	-1.60	0.80
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 36: Delta XCOPA accuracy after steering LAPE neurons with **pmax** for SeaLLM3 1.5B.

et	0.40	0.20	-0.20	-1.00	-0.80	1.40	-1.20	-1.00	-1.20	0.20	-1.00
ht	0.00	0.80	2.00	-0.40	-0.20	-0.20	-2.00	0.20	2.40	-1.60	0.20
id	-6.00	-2.40	-0.40	-7.00	-2.20	-4.80	-13.8	-13.6	-8.40	-9.00	-15.0
it	-7.80	-0.80	-6.80	0.40	-0.80	-1.80	-9.40	-11.4	-9.20	-6.60	-12.4
qu	-1.20	0.40	0.20	0.60	-0.20	0.60	-1.60	-2.00	0.20	-1.00	-1.40
sw	0.60	-1.00	1.20	0.20	0.00	-1.00	-0.60	-0.20	-0.80	0.80	-0.20
ta	1.00	-0.20	0.20	1.80	0.40	1.60	1.00	0.60	2.00	1.20	-0.60
th	0.00	0.00	-2.60	-1.00	-1.40	-1.80	0.00	-0.20	-3.60	-4.00	-4.00
tr	-1.60	-2.60	-2.00	-2.20	-2.40	-3.00	0.60	0.40	-1.40	0.80	-2.80
vi	-6.60	0.80	-10.4	-3.60	1.00	-2.20	-11.6	-12.8	-11.6	-0.60	-12.4
zh	-0.40	0.20	0.80	-2.20	0.60	-2.00	-1.60	1.00	-0.20	-1.00	0.00
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 38: Delta XCOPA accuracy after steering LAPE neurons with **pmedian** for SeaLLMs 1.5B. Row i is the initial language, column j is the language of intervention.

et	-0.40	0.20	-2.80	-2.20	-2.40	1.00	0.40	-0.80	-1.80	0.60	0.20
ht	0.40	0.60	0.20	1.60	1.40	0.80	0.40	0.80	2.00	-0.20	0.20
id	-13.6	-12.8	-1.80	-8.00	-13.0	-14.6	-13.8	-20.0	-12.8	-18.4	-22.2
it	-17.4	-13.0	-15.6	-1.00	-13.8	-14.2	-13.8	-16.8	-14.8	-18.4	-19.0
qu	0.60	1.40	1.00	1.80	0.00	0.20	-0.20	-0.40	-0.40	0.20	-0.40
sw	0.60	-1.20	-1.40	-2.20	0.80	1.00	1.80	0.40	-0.40	0.40	-1.40
ta	-3.20	-3.20	-1.60	-2.20	-3.60	-2.40	-0.60	-0.60	-2.60	-1.00	-3.00
th	-3.60	-4.00	-3.60	-2.80	-2.20	-4.60	-3.00	0.40	-2.60	-5.00	-5.00
tr	-5.40	-5.00	-5.20	-4.00	-5.00	-2.00	-4.40	-5.20	-1.00	-6.00	-6.00
vi	-15.2	-11.6	-14.4	-5.40	-12.6	-14.6	-14.4	-18.0	-14.0	-0.20	-20.6
zh	-5.80	-3.20	-2.20	-4.40	-4.40	-5.40	-7.00	-6.80	-7.60	-3.80	-1.80
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 37: Delta XCOPA accuracy after steering LAPE neurons with **pmax** for SeaLLM3 7B.

et	1.80	-0.60	-0.40	-3.60	-0.20	0.20	2.40	2.40	0.80	-0.80	0.20
ht	0.60	0.00	-0.40	-0.60	0.20	0.60	0.00	0.80	-0.60	0.00	-0.80
id	-0.40	-0.60	-0.80	-3.20	-0.60	-1.20	-1.60	-1.80	-0.40	-3.00	-3.60
it	-1.00	-1.00	-1.80	-0.40	-2.40	-3.20	-3.00	-3.40	-2.80	-6.00	-5.20
qu	0.80	0.60	0.60	0.60	0.80	0.60	-3.00	-1.60	-0.20	-1.00	0.80
sw	0.60	0.60	0.80	-0.20	1.60	2.60	-1.40	0.80	-0.40	-0.40	-2.40
ta	-0.80	-2.60	1.60	-1.80	-3.80	0.20	0.40	-0.20	0.40	-0.20	-1.40
th	-1.00	0.40	-1.00	1.20	1.40	-0.80	-0.40	-0.60	-0.60	-0.40	-0.40
tr	-2.00	0.00	-1.80	0.00	-0.40	-0.60	2.00	0.60	-3.20	-1.00	-0.60
vi	-0.80	-1.60	-1.20	-4.40	-1.60	-1.20	0.60	1.20	-1.80	1.20	-5.80
zh	-1.00	-1.60	-0.40	-1.80	-2.40	0.00	-2.60	-1.60	-1.40	-2.40	0.20
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 39: Delta XCOPA accuracy after steering LAPE neurons with **pmedian** for Gemma2 2B. Row i is the initial language, column j is the language of intervention.

en	0.08	0.21	-5.94	-0.82	-0.18	-6.11
fr	-1.21	1.20	-6.03	2.41	-6.03	-2.41
jp	-4.59	0.21	-0.21	-1.56	0.52	-1.04
pt	-7.23	0.00	-6.47	-1.14	0.38	-5.71
ru	-4.77	0.95	-3.50	1.58	-2.23	-7.31
zh	-0.20	-0.99	0.99	-0.40	3.37	1.59
	en	fr	jp	pt	ru	zh

Figure 40: Delta XCOPA accuracy after steering LAPE neurons with **pmedian** for Qwen2.5 0.5B. Row i is the initial language, column j is the language of intervention.

et	1.00	-0.20	-1.00	-1.60	0.40	-1.20	2.00	1.00	-2.80	-0.80	2.40
ht	0.80	0.00	1.20	-1.00	1.00	0.20	0.80	1.00	1.40	-2.80	0.20
id	-4.40	-6.00	-6.60	-7.20	-6.40	-6.00	-8.80	-7.60	-7.00	-6.40	-6.20
it	-0.80	-0.60	-1.40	-1.80	-1.60	-1.20	-0.60	-1.80	-0.40	-2.40	-0.60
qu	0.60	-0.20	1.00	0.00	-0.60	0.40	0.80	-0.40	0.80	0.80	0.40
sw	-2.00	-1.80	-0.20	-1.80	-2.20	-1.80	1.20	0.40	-1.00	-0.80	-0.40
ta	1.00	1.80	0.60	1.80	0.60	0.80	0.40	-0.40	2.60	1.00	0.00
th	-3.20	-1.80	-2.60	-2.80	-3.20	-3.60	-2.60	-2.20	-2.40	-2.40	-3.00
tr	0.60	-2.40	-0.40	-0.80	-1.80	-2.40	0.00	-0.80	-2.60	-0.60	0.40
vi	-7.00	-8.60	-7.80	-8.60	-6.80	-8.00	-8.00	-9.20	-9.20	-6.60	-8.60
zh	-9.40	-8.00	-8.00	-8.80	-7.40	-9.20	-7.20	-8.00	-7.20	-8.40	-8.00
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 42: Delta XCOPA accuracy after steering base-line neurons with **pmax** for Qwen2.5 0.5B.

et	1.00	-0.40	-1.40	0.40	-1.20	-1.20	0.40	1.00	0.40	-0.20	-1.00
ht	0.80	2.20	1.00	0.60	0.60	0.80	0.40	1.80	1.40	1.80	1.20
id	-6.20	-8.00	-10.8	-7.80	-7.20	-8.80	-12.4	-8.60	-9.40	-11.6	-7.00
it	-10.8	-9.40	-11.6	-13.4	-10.8	-10.6	-9.80	-10.2	-10.4	-9.80	-10.2
qu	-1.20	-0.60	-2.00	-2.40	1.00	-0.20	-2.40	0.00	-0.80	-1.60	-2.00
sw	-0.60	0.60	0.60	-1.20	-0.60	-1.60	-1.80	0.60	2.80	0.20	0.40
ta	-1.20	1.20	-3.00	-2.40	-1.80	-3.40	-4.00	-3.20	-0.60	-2.40	-2.20
th	-1.00	-2.60	1.20	-2.80	-3.80	-1.60	-1.60	1.20	-0.20	-2.40	-2.40
tr	-2.60	-4.20	-4.00	-4.40	-2.40	-3.00	-4.80	-3.80	-6.60	-5.60	-5.40
vi	-5.00	-5.20	-5.40	-5.20	-7.00	-5.00	-6.60	-7.20	-4.80	-10.8	-5.80
zh	-10.0	-9.60	-9.20	-10.0	-10.0	-9.20	-9.00	-10.0	-10.6	-10.4	-11.6
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 41: Delta XCOPA accuracy after steering base-line neurons with **pmax** for Gemma2 2B.

et	-2.00	-2.20	-0.80	-4.60	-3.20	-3.00	-1.20	0.40	-3.00	-3.80	0.80
ht	-1.00	-0.60	1.20	0.60	-0.20	-0.20	-0.80	-0.40	-1.00	-0.80	0.20
id	-13.6	-12.4	-10.0	-12.2	-12.8	-12.8	-15.4	-14.6	-13.8	-13.0	-14.0
it	-12.0	-10.6	-12.0	-9.20	-11.8	-10.8	-9.60	-10.6	-12.8	-11.4	-9.80
qu	-1.80	0.00	-0.80	-0.60	-0.60	-1.20	-0.60	0.20	0.20	-1.40	0.00
sw	0.20	0.80	-0.60	0.40	0.00	1.60	0.00	-0.60	4.80	-0.60	0.60
ta	3.60	3.00	1.40	4.00	3.00	4.40	2.40	4.60	3.00	4.80	3.40
th	-2.60	-0.60	-3.40	-4.40	-2.00	-2.80	-5.60	-1.40	-3.00	-5.20	-5.60
tr	-1.00	-1.80	-2.20	-2.80	-3.20	-4.40	-1.40	-1.40	-1.80	0.60	0.00
vi	-16.4	-13.8	-14.6	-14.8	-13.4	-14.6	-17.2	-15.0	-17.0	-7.60	-15.8
zh	-16.0	-14.0	-11.6	-13.2	-15.0	-15.2	-15.0	-9.60	-14.4	-7.00	-14.0
	et	ht	id	it	qu	sw	ta	th	tr	vi	zh

Figure 43: Delta XCOPA accuracy after steering base-line neurons with **pmax** for SeaLLM3 1.5B.

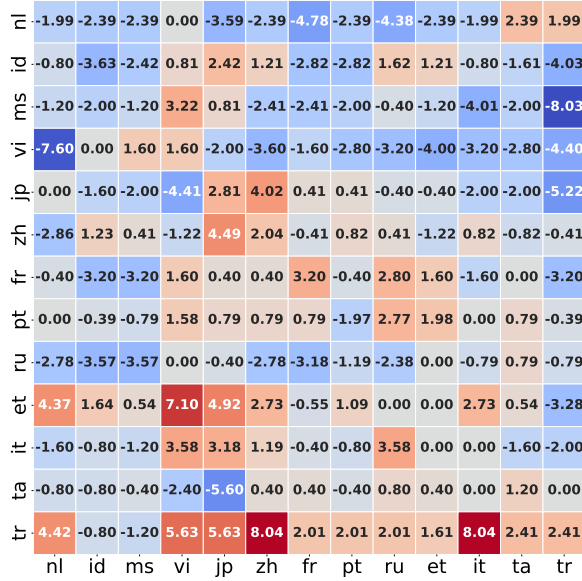


Figure 44: Delta Include-lite accuracy after steering LAPE neurons with **pmax** for Qwen2.5 0.5B.

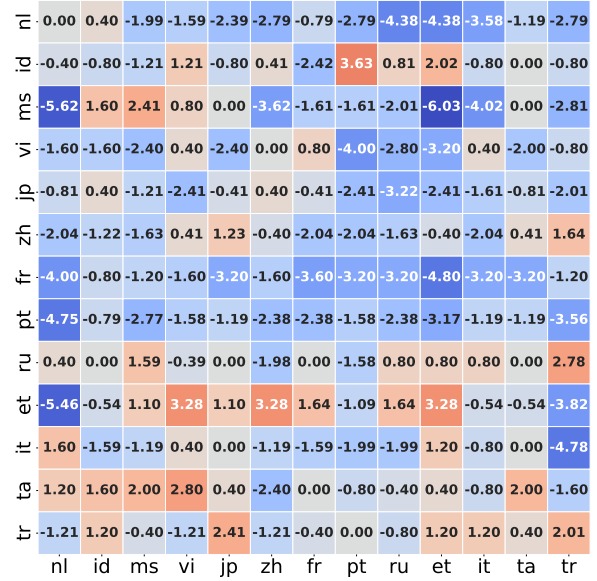


Figure 46: Delta Include-lite accuracy after steering LAPE neurons with **pmax** for Gemma2 2B.

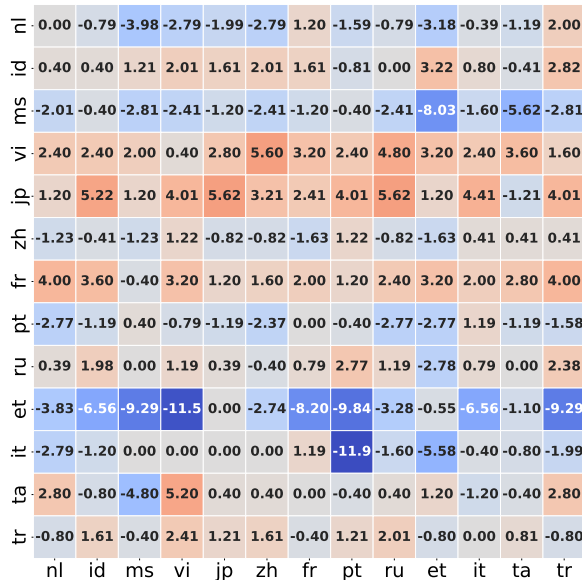


Figure 45: Delta Include-lite accuracy after steering LAPE neurons with **pmax** for Qwen2.5 7B.

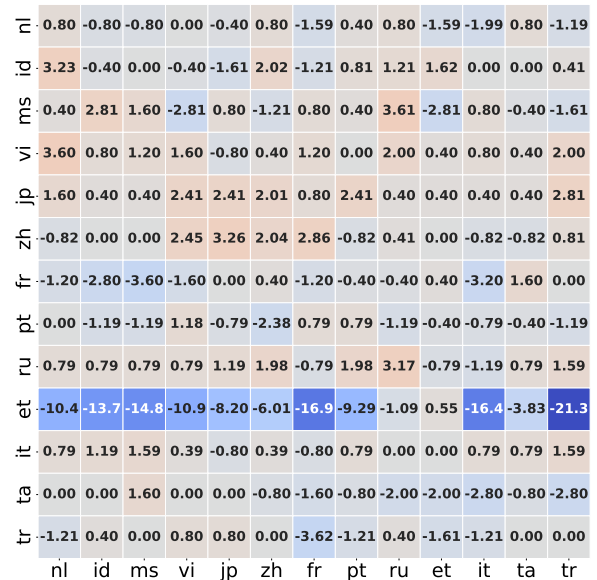


Figure 47: Delta Include-lite accuracy after steering LAPE neurons with **pmax** for Gemma2 9B.

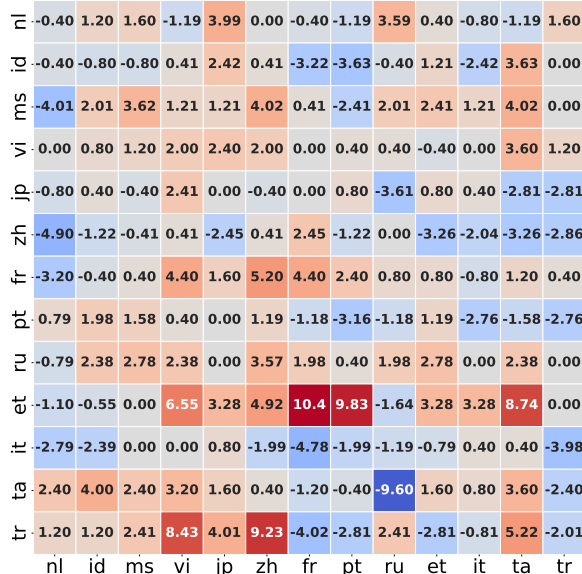


Figure 48: Delta Include-lite accuracy after steering LAPE neurons with **pmax** for SeaLLMv3 1.5B.

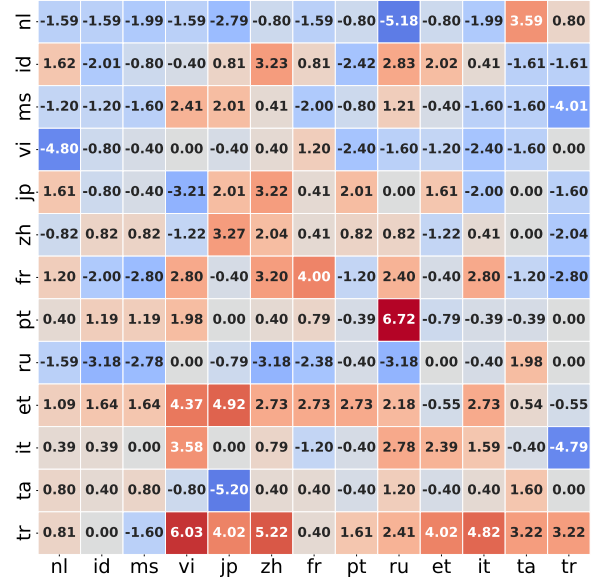


Figure 50: Delta Include-lite accuracy after steering LAPE neurons with **pmedian** for Qwen2.5 0.5B.

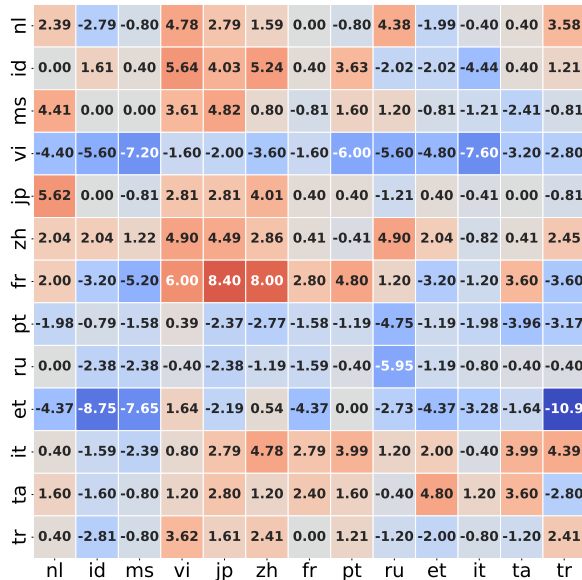


Figure 49: Delta Include-lite accuracy after steering LAPE neurons with **pmax** for SeaLLMv3 7B.

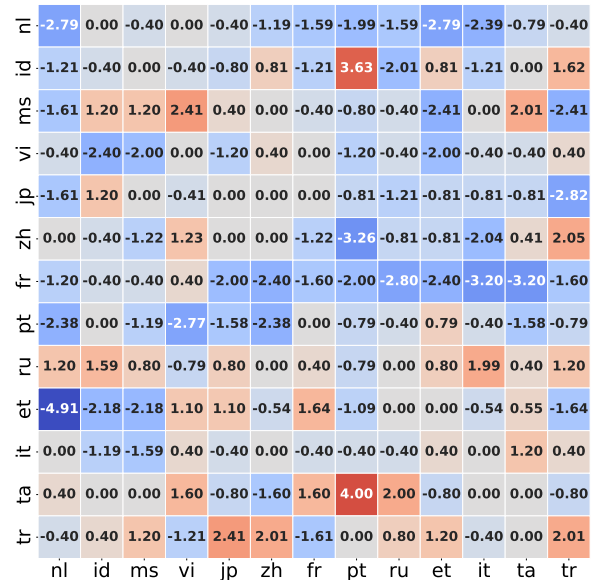


Figure 51: Delta Include-lite accuracy after steering LAPE neurons with **pmedian** for Gemma2 2B.

nl	-0.80	1.60	1.60	1.99	1.99	2.79	1.99	1.60	3.59	0.80	1.20	2.79	0.40
id	-2.82	0.00	-0.80	2.83	0.00	1.62	-1.61	-0.80	0.41	1.21	-0.40	3.23	-2.01
ms	0.00	1.61	1.21	2.82	4.02	5.23	2.82	1.61	4.02	1.21	3.62	3.22	0.00
vi	0.80	1.20	1.60	1.60	3.60	1.20	0.00	0.00	1.20	-0.80	1.20	2.80	1.60
jp	-1.20	0.00	0.80	0.40	0.40	-0.80	0.80	0.00	-1.61	0.80	-0.40	-0.40	-1.20
zh	-4.49	-0.82	-1.22	0.82	-0.82	0.00	1.23	-0.82	0.41	-1.63	-1.22	-0.82	-2.04
fr	-3.20	0.00	0.00	4.00	3.20	4.80	4.80	4.00	0.00	-1.60	1.20	1.60	1.60
pt	1.19	1.98	3.56	1.19	0.00	1.58	-3.95	-2.37	-0.79	1.19	-1.58	-3.95	-3.95
ru	-0.40	2.38	2.78	3.17	0.00	4.37	2.78	1.59	2.38	2.38	3.17	1.19	-0.79
et	1.64	3.28	2.73	5.46	3.82	1.09	8.74	7.65	-1.64	3.28	2.73	6.55	1.09
it	-1.59	-2.39	-2.39	0.80	0.80	1.60	-2.79	-0.79	-0.40	0.00	0.00	1.60	-4.38
ta	2.40	4.00	4.00	0.80	3.20	1.60	-0.80	-0.40	-2.80	2.00	0.40	2.40	-2.00
tr	2.81	2.41	2.41	4.82	2.81	4.82	-4.42	-1.21	2.81	0.40	-0.81	1.60	-3.62
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	it	ta	tr

Figure 52: Delta Include-lite accuracy after steering LAPE neurons with **pmedian** for SeaLLMv3 1.5B.

nl	-26.7	-27.1	-21.9	-23.9	-28.3	-20.3	-24.7	-30.3	-22.7	-29.1	-27.9	-23.1	-27.9
id	-29.8	-28.6	-29.0	-26.2	-30.2	-30.6	-27.4	-28.2	-29.0	-27.8	-27.4	-31.1	-30.2
ms	-18.1	-20.5	-15.3	-20.9	-20.5	-18.1	-20.9	-19.7	-18.5	-22.1	-23.3	-22.5	-23.7
vi	-20.0	-23.2	-22.4	-16.8	-20.8	-18.4	-23.6	-20.8	-26.0	-22.8	-22.0	-21.6	-17.2
jp	-22.5	-29.7	-27.7	-32.9	-32.5	-29.7	-30.5	-29.7	-30.1	-34.5	-30.5	-28.5	-30.1
zh	-24.9	-19.6	-19.2	-17.1	-17.6	-18.8	-18.8	-22.9	-18.8	-21.2	-20.0	-22.4	-22.0
fr	-22.0	-23.2	-24.4	-20.4	-25.2	-22.0	-20.0	-28.8	-20.0	-30.4	-25.2	-22.0	-23.6
pt	-16.2	-21.3	-19.0	-23.7	-23.7	-25.7	-22.1	-21.7	-25.7	-26.9	-20.9	-22.5	-21.7
ru	-22.2	-22.6	-22.6	-18.3	-19.8	-23.0	-21.0	-23.8	-20.2	-26.6	-21.8	-22.2	-23.8
et	-12.6	-11.5	-6.55	-14.2	-13.1	-11.5	-7.65	-12.6	-3.28	-15.8	-14.8	-10.9	-8.74
it	-31.5	-29.5	-29.1	-25.9	-27.5	-30.7	-28.3	-31.5	-29.1	-33.9	-32.7	-28.3	-28.3
ta	-2.00	-9.20	-7.60	-2.40	-6.40	-12.0	-6.40	-14.0	-3.20	-14.0	-6.80	-16.4	-15.6
tr	-16.9	-15.3	-14.1	-13.7	-18.1	-14.1	-20.1	-12.0	-18.9	-14.9	-18.1	-15.7	-16.5
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	it	ta	tr

Figure 54: Delta Include-lite accuracy after steering Baseline neurons with **pmax** for Gemma2 2B.

nl	-19.1	-11.6	-6.77	-17.1	-19.9	-13.9	-15.5	-9.96	-16.3	-22.3	-17.9	-20.7	-23.5
id	-5.64	-15.7	-14.9	-11.3	-10.5	-8.87	-14.1	-14.9	-13.7	-11.7	-10.1	-10.5	-6.85
ms	-14.9	-11.2	-8.43	-9.63	-14.4	-11.2	-17.3	-11.2	-12.9	-12.4	-10.0	-12.0	-15.3
vi	-15.2	-13.6	-14.0	-17.2	-12.4	-13.2	-14.4	-16.0	-15.6	-14.4	-17.2	-14.8	-12.4
jp	-20.1	-18.9	-16.9	-16.5	-18.5	-18.5	-19.3	-17.7	-20.1	-20.5	-20.5	-20.1	-21.7
zh	-29.4	-30.2	-29.4	-30.2	-29.8	-30.6	-28.6	-29.4	-29.8	-28.6	-29.4	-29.0	-27.8
fr	-15.6	-7.60	-9.60	-10.8	-12.8	-10.8	-12.8	-12.4	-14.0	-8.40	-9.60	-13.2	-14.0
pt	-11.5	-10.7	-12.7	-10.3	-13.0	-7.90	-13.4	-9.88	-14.6	-9.88	-11.1	-13.0	-11.1
ru	-9.52	-13.1	-13.9	-14.3	-10.7	-15.1	-14.3	-15.1	-15.5	-14.3	-12.3	-15.1	-11.1
et	-5.47	-1.64	-1.64	-1.64	-2.19	-4.92	2.18	-4.92	-9.29	1.09	-1.10	-0.55	0.54
it	-12.4	-12.4	-11.6	-10.8	-12.8	-16.7	-13.9	-12.8	-17.5	-16.3	-13.9	-12.4	-13.1
ta	-10.0	-9.60	0.80	-2.80	-11.6	-11.2	-11.2	-11.6	-8.40	-12.4	-11.6	-10.4	-11.2
tr	2.81	3.22	5.22	3.62	6.83	8.44	9.64	1.21	4.82	4.82	4.42	3.22	5.63
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	it	ta	tr

Figure 53: Delta Include-lite accuracy after steering Baseline neurons with **pmax** for Qwen2.5 0.5B.

nl	-19.5	-17.5	-19.9	-16.7	-21.5	-26.3	-20.3	-19.5	-27.9	-29.1	-24.3	-28.3	-12.7
id	-25.0	-22.6	-26.2	-25.4	-25.0	-26.6	-20.6	-26.2	-27.4	-25.8	-25.8	-23.4	-29.0
ms	-14.0	-11.2	-13.2	-11.2	-11.6	-15.3	-14.0	-16.1	-16.5	-14.9	-12.8	-17.3	-7.63
vi	-21.2	-16.4	-22.8	-18.0	-22.4	-16.8	-24.8	-14.8	-14.0	-20.4	-21.6	-19.6	-16.4
jp	-30.1	-24.9	-29.3	-27.3	-24.5	-29.7	-25.7	-32.5	-24.9	-30.5	-26.5	-29.7	-28.1
zh	-48.2	-49.4	-46.9	-45.3	-45.3	-42.5	-46.5	-48.6	-48.6	-45.3	-44.1	-49.8	-51.0
fr	-14.4	-12.0	-13.2	-11.2	-12.0	-15.2	-13.6	-12.4	-11.2	-13.6	-8.80	-14.4	-13.6
pt	-13.4	-15.4	-19.8	-12.6	-15.4	-19.0	-14.6	-15.8	-22.1	-20.2	-17.8	-19.8	-21.7
ru	-17.1	-18.7	-16.7	-20.6	-15.5	-16.3	-23.4	-19.4	-15.1	-13.9	-18.7	-16.3	-14.7
et	-1.64	-1.10	4.37	2.73	-3.28	9.83	-1.64	4.37	4.37	8.19	3.28	4.37	9.83
it	-28.7	-24.7	-24.7	-23.5	-28.3	-25.5	-25.9	-24.3	-25.1	-25.1	-28.7	-26.3	-20.3
ta	-7.20	-10.0	-6.40	-6.40	-5.20	-17.6	-10.8	-8.80	-12.4	-12.0	-10.8	-12.0	-12.8
tr	-10.4	-8.44	-10.0	-7.23	-8.04	-2.41	-7.63	-5.63	-2.01	-2.41	-2.81	-2.81	-11.7
	nl	id	ms	vi	jp	zh	fr	pt	ru	et	it	ta	tr

Figure 55: Delta Include-lite accuracy after steering Baseline neurons with **pmax** for SeaLLMv3 1.5B.



Figure 56: PPL changes after steering LAPE neurons with **pmax** for Qwen2.5 0.5B.



Figure 58: PPL changes after steering LAPE neurons with **pmax** for Gemma2 2B.



Figure 57: PPL changes after steering LAPE neurons with **pmax** for Qwen2.5 7B.

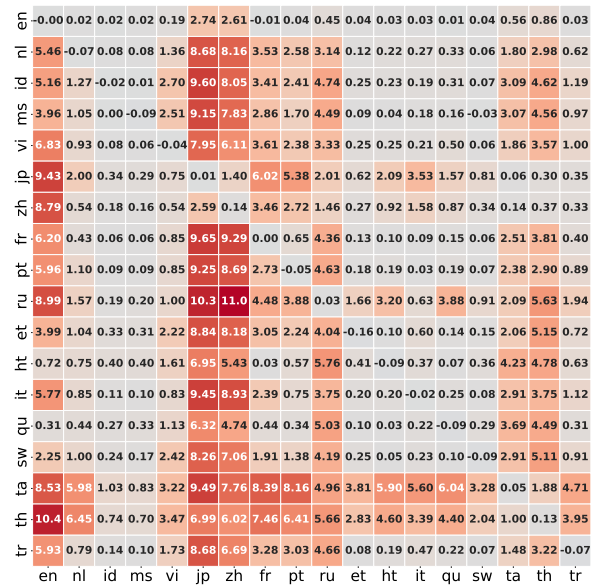




Figure 60: PPL changes after steering LAPE neurons with **pmax** for SeaLLM3 1.5B.

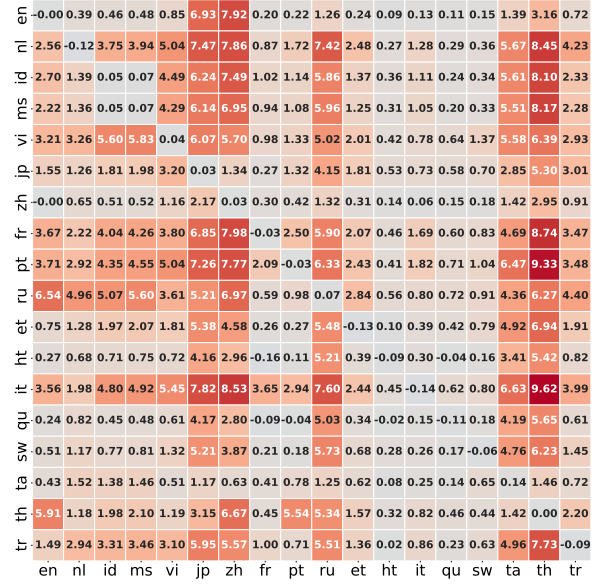
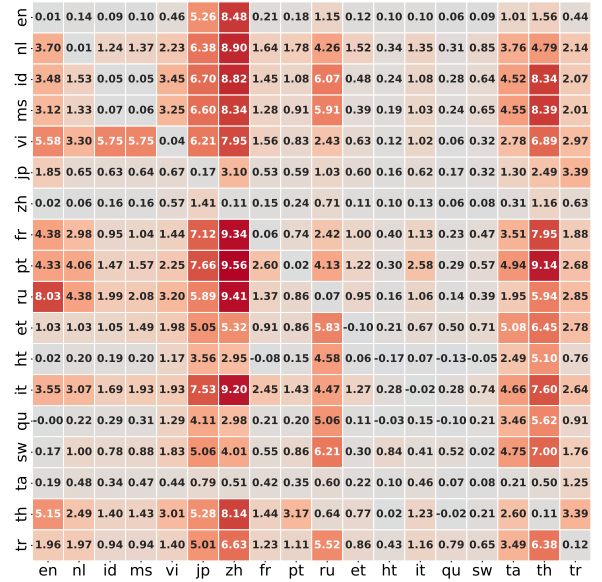


Figure 62: PPL changes after steering LAPE neurons with **pmedian** for Qwen2.5 0.5 B.



Figure 61: PPL changes after steering LAPE neurons with **pmax** for SeaLLM3 7B.



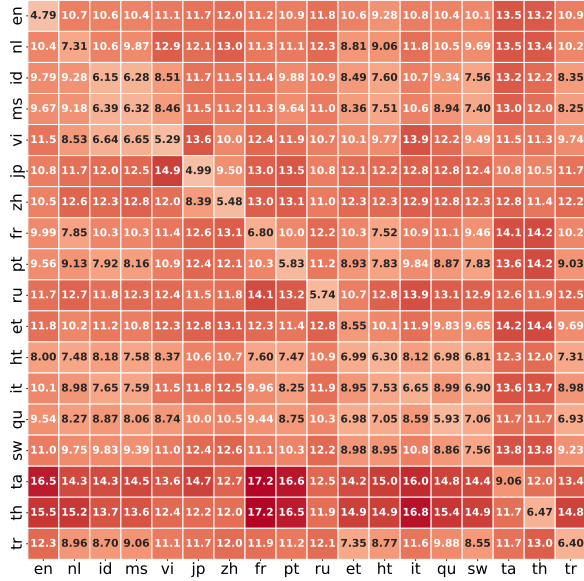


Figure 64: PPL changes after steering LAPE neurons with **pmedian** for Gemma2 2B.



Figure 65: PPL changes after steering Baseline neurons with **pmax** for Qwen2.5 0.5B.

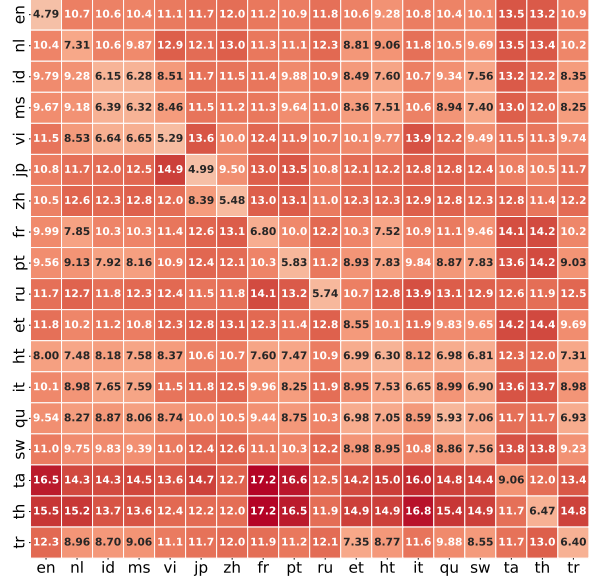
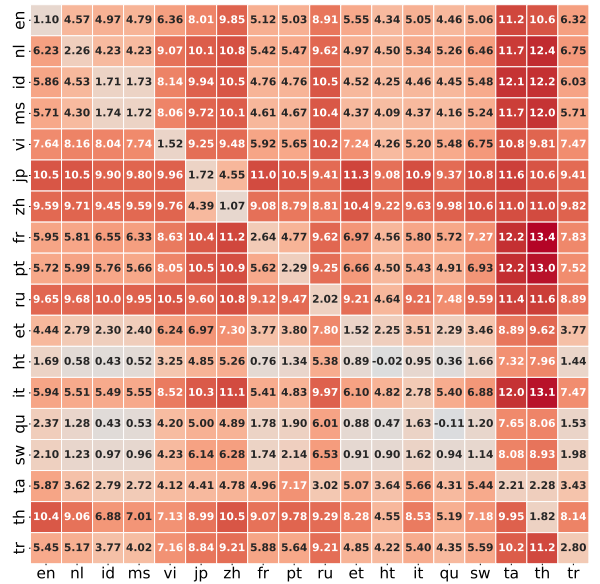


Figure 66: PPL changes after steering Baseline neurons with **pmax** for Gemma2 2B.



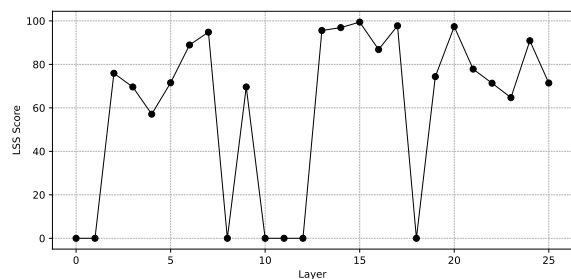


Figure 69: Layer-wise average LSS scores when intervened by en neurons for Gemma2 2B.

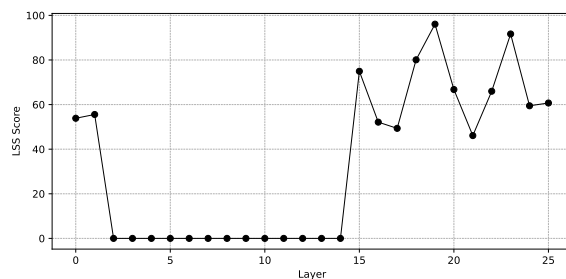


Figure 73: Layer-wise average LSS scores when intervened by id neurons for Gemma2 2B.

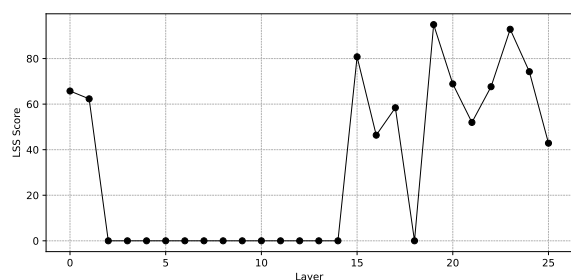


Figure 70: Layer-wise average LSS scores when intervened by ms neurons for Gemma2 2B.

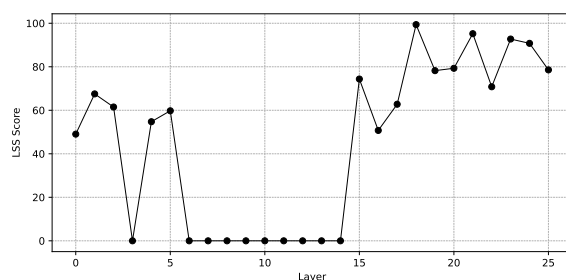


Figure 74: Layer-wise average LSS scores when intervened by vi neurons for Gemma2 2B.

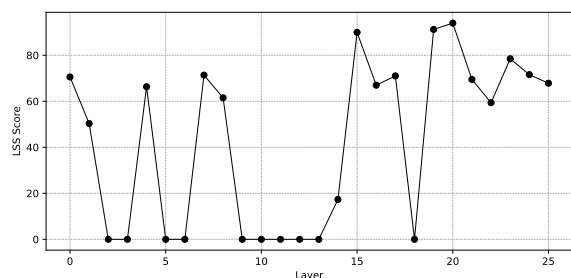


Figure 71: Layer-wise average LSS scores when intervened by et neurons for Gemma2 2B.

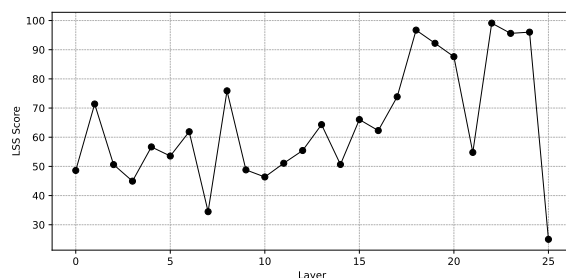


Figure 75: Layer-wise average LSS scores when intervened by jp neurons for Gemma2 2B.

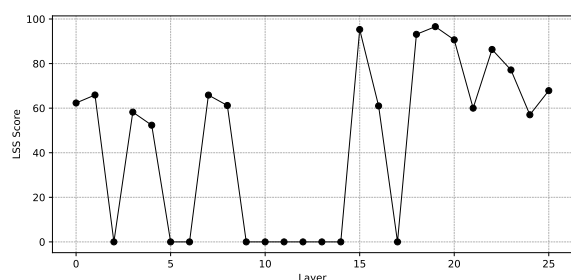


Figure 72: Layer-wise average LSS scores when intervened by n1 neurons for Gemma2 2B.

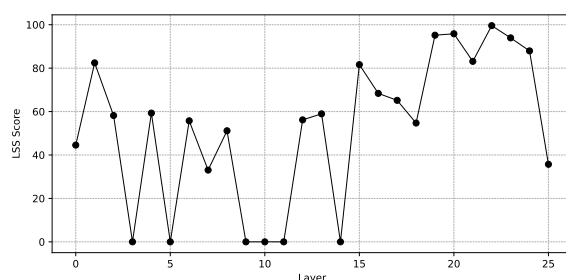


Figure 76: Layer-wise average LSS scores when intervened by zh neurons for Gemma2 2B.

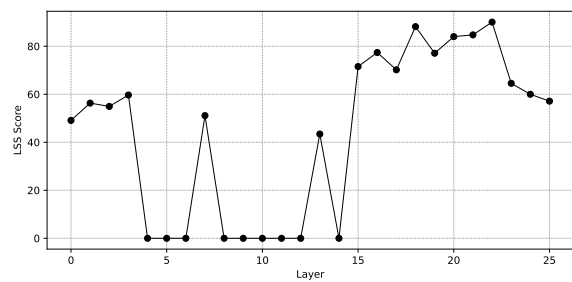


Figure 77: Layer-wise average LSS scores when intervened by fr neurons for Gemma2 2B.

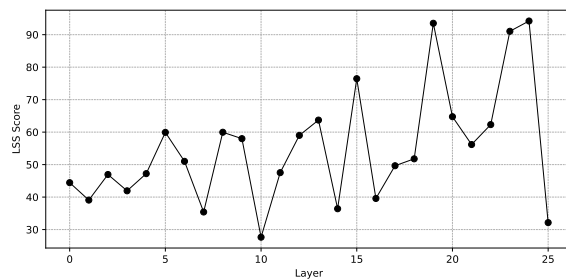


Figure 81: Layer-wise average LSS scores when intervened by ta neurons for Gemma2 2B.

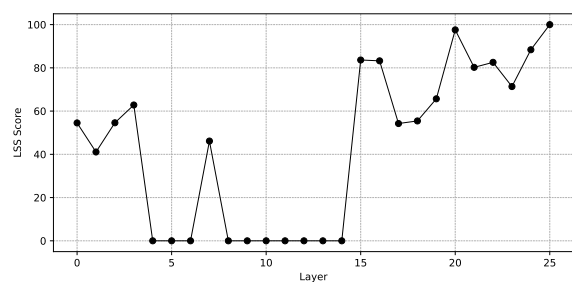


Figure 78: Layer-wise average LSS scores when intervened by pt neurons for Gemma2 2B.

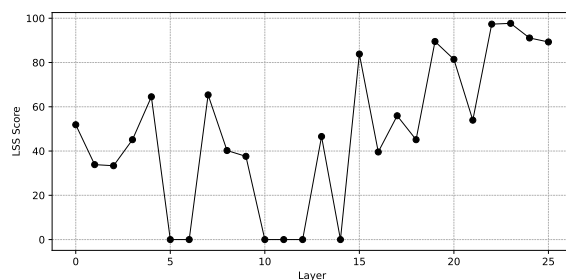


Figure 82: Layer-wise average LSS scores when intervened by th neurons for Gemma2 2B.

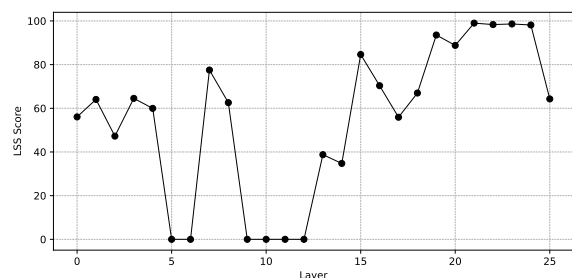


Figure 79: Layer-wise average LSS scores when intervened by ru neurons for Gemma2 2B.

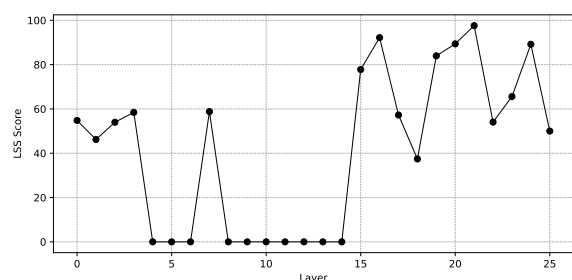


Figure 80: Layer-wise average LSS scores when intervened by it neurons for Gemma2 2B.

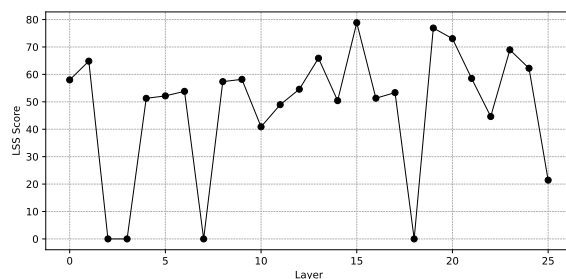


Figure 83: Layer-wise average LSS scores when intervened by tr neurons for Gemma2 2B.

Base language: en Question: "Madrid is the capital city of", ans_en: "Spain"			
target	ans_target	Non-intervened LLM	Intervened LLM
nl	"Spanje"	> Tokenized target: ['GSpan', 'je'] Rank target: 18 Logit: 12.40 Prob: 0.03% Token: GSpan Rank target: 1055 Logit: 6.90 Prob: 0.00% Token: je Ranks of the target tokens: [('GSpan', 18), ('je', 1055)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GSpan', 'je'] Rank target: 3 Logit: 15.27 Prob: 2.08% Token: GSpan Rank target: 7 Logit: 15.03 Prob: 3.99% Token: je Ranks of the target tokens: [('GSpan', 3), ('je', 7)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 18.65 Prob: 61.16% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]
id	"Spanyol"	> Tokenized target: ['GS', 'pany', 'ol'] Rank target: 335 Logit: 8.02 Prob: 0.00% Token: GS Rank target: 12 Logit: 11.65 Prob: 1.18% Token: pany Rank target: 1 Logit: 14.39 Prob: 2.17% Token: ol Ranks of the target tokens: [('GS', 335), ('pany', 12), ('ol', 1)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GS', 'pany', 'ol'] Rank target: 38 Logit: 9.78 Prob: 0.06% Token: GS Rank target: 0 Logit: 14.90 Prob: 55.03% Token: pany Rank target: 1 Logit: 15.41 Prob: 2.09% Token: ol Ranks of the target tokens: [('GS', 38), ('pany', 0), ('ol', 1)] > Tokenized base: ['GSpain'] Rank base: 4 Logit: 13.53 Prob: 2.47% Token: GSpain Ranks of the base tokens: [('GSpain', 4)]
ms	"Sepanyol"	> Tokenized target: ['GSep', 'any', 'ol'] Rank target: 1485 Logit: 5.47 Prob: 0.00% Token: GSep Rank target: 9 Logit: 11.88 Prob: 1.77% Token: any Rank target: 16 Logit: 11.99 Prob: 0.98% Token: ol Ranks of the target tokens: [('GSep', 1485), ('any', 9), ('ol', 16)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GSep', 'any', 'ol'] Rank target: 81 Logit: 8.51 Prob: 0.02% Token: GSep Rank target: 0 Logit: 16.48 Prob: 64.31% Token: any Rank target: 8 Logit: 12.26 Prob: 0.52% Token: ol Ranks of the target tokens: [('GSep', 81), ('any', 0), ('ol', 8)] > Tokenized base: ['GSpain'] Rank base: 4 Logit: 13.03 Prob: 1.43% Token: GSpain Ranks of the base tokens: [('GSpain', 4)]
vi	"Tây Ban Nha"	> Tokenized target: ['GTÃçy', 'GBan', 'GN', 'ha'] Rank target: 301 Logit: 8.19 Prob: 0.00% Token: GTÃçy Rank target: 2 Logit: 11.73 Prob: 5.06% Token: GBan Rank target: 0 Logit: 16.42 Prob: 97.35% Token: GN Rank target: 0 Logit: 21.45 Prob: 97.86% Token: ha Ranks of the target tokens: [('GTÃçy', 301), ('GBan', 2), ('GN', 0), ('ha', 0)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GTÃçy', 'GBan', 'GN', 'ha'] Rank target: 1 Logit: 12.50 Prob: 19.94% Token: GTÃçy Rank target: 0 Logit: 12.45 Prob: 16.44% Token: GBan Rank target: 0 Logit: 14.34 Prob: 74.70% Token: GN Rank target: 0 Logit: 19.49 Prob: 88.90% Token: ha Ranks of the target tokens: [('GTÃçy', 1), ('GBan', 0), ('GN', 0), ('ha', 0)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 12.70 Prob: 24.31% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]
jp	"スペイン"	> Tokenized target: ['GãH', 'ãh', 'ãH=ãh'] Rank target: 7964 Logit: 2.80 Prob: 0.00% Token: GãH Rank target: 1 Logit: 16.41 Prob: 22.12% Token: ãh Rank target: 0 Logit: 20.24 Prob: 96.04% Token: ãH=ãh Ranks of the target tokens: [('GãH', 7964), ('ãh', 1), ('ãH=ãh', 0)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GãH', 'ãh', 'ãH=ãh'] Rank target: 228 Logit: 7.78 Prob: 0.01% Token: GãH Rank target: 5 Logit: 14.13 Prob: 2.41% Token: ãh Rank target: 0 Logit: 18.39 Prob: 72.19% Token: ãH=ãh Ranks of the target tokens: [('GãH', 228), ('ãh', 5), ('ãH=ãh', 0)] > Tokenized base: ['GSpain'] Rank base: 4 Logit: 12.56 Prob: 1.64% Token: GSpain Ranks of the base tokens: [('GSpain', 4)]
zh	"西班牙"	> Tokenized target: ['Gè', '¥ç', 'çìN', 'çìL'] Rank target: 4052 Logit: 3.93 Prob: 0.00% Token: Gè Rank target: 0 Logit: 23.56 Prob: 90.92% Token: ¥ç Rank target: 0 Logit: 21.29 Prob: 99.77% Token: çìN Rank target: 0 Logit: 22.27 Prob: 99.90% Token: çìL Ranks of the target tokens: [('Gè', 4052), ('¥ç', 0), ('çìN', 0), ('çìL', 0)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['Gè', '¥ç', 'çìN', 'çìL'] Rank target: 148 Logit: 10.36 Prob: 0.00% Token: Gè Rank target: 0 Logit: 22.89 Prob: 84.35% Token: ¥ç Rank target: 0 Logit: 24.45 Prob: 99.93% Token: çìN Rank target: 0 Logit: 21.20 Prob: 99.67% Token: çìL Ranks of the target tokens: [('Gè', 148), ('¥ç', 0), ('çìN', 0), ('çìL', 0)] > Tokenized base: ['GSpain'] Rank base: 28 Logit: 13.28 Prob: 0.05% Token: GSpain Ranks of the base tokens: [('GSpain', 28)]
fr	"Espagne"	> Tokenized target: ['GEsp', 'agne'] Rank target: 158 Logit: 9.24 Prob: 0.00% Token: GEsp Rank target: 1 Logit: 16.14 Prob: 21.39% Token: agne Ranks of the target tokens: [('GEsp', 158), ('agne', 1)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GEsp', 'agne'] Rank target: 38 Logit: 11.45 Prob: 0.08% Token: GEsp Rank target: 0 Logit: 21.62 Prob: 94.81% Token: agne Ranks of the target tokens: [('GEsp', 38), ('agne', 0)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 17.36 Prob: 30.50% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]
pt	"Espanha"	> Tokenized target: ['GE', 'span', 'ha'] Rank target: 169 Logit: 9.15 Prob: 0.00% Token: GE Rank target: 0 Logit: 16.57 Prob: 59.42% Token: span Rank target: 6 Logit: 15.96 Prob: 0.20% Token: ha Ranks of the target tokens: [('GE', 169), ('span', 0), ('ha', 6)] > Tokenized base: ['GSpain'] Rank base: 0 Logit: 20.14 Prob: 76.20% Token: GSpain Ranks of the base tokens: [('GSpain', 0)]	> Tokenized target: ['GE', 'span', 'ha'] Rank target: 28 Logit: 11.59 Prob: 0.10% Token: GE Rank target: 0 Logit: 17.49 Prob: 84.41% Token: span Rank target: 0 Logit: 19.20 Prob: 69.07% Token: ha Ranks of the target tokens: [('GE', 28), ('span', 0), ('ha', 0)] > Tokenized prompt: ['Mad', 'rid', 'Gis', 'Gthe', 'Gcapital', 'Gcity', 'Gof'] Tokenized base: ['GSpain'] Rank base: 2 Logit: 16.27 Prob: 10.72% Token: GSpain Ranks of the base tokens: [('GSpain', 2)]

Figure 84: Visualization of token log-prob for *base* and *target* answers. Base answers are in the prompt language (en), target answers are in the language of intervention (nl,id,ms,vi,jp,zh,fr,pt).

[illegible]

Figure 86: A sample of Qwen2.5 0.5B output for targeted prompt translation of en to nl, id, ms, fr, qu, ht, sw.

Neuron lang	"Translate from English to target language. English: On Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool. Target language:"
None (not-intervened)	Russian\nRussian: 1. Студенты из Стэнфордского университета имени Лейла Маркеса объявили о создании нового диагностического инструмента.\n\n2. C'
nl	Dutch\nNiet te gebruiken:\nOn Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool. \n\nDutch:\n\nWe hebben op Monday de ontdekking van een nieuwe diagnoseertool uitgebracht in'
id	Bahasa Indonesia\nSaya ingin mengajak Anda untuk membuat teks yang sama dalam bahasa Indonesia. Berikut adalah contoh:\n\n"Senin, seorang ahli medis dari Universitas Stanford, Stanford University, menandatangani"
ms	Bahasa Indonesia\nIngatlah bahwa kata-kata yang Anda berikan adalah kata-kata Inggris. Jika ada yang tidak sesuai atau belum terbaca dalam bahasa Inggris, saya akan mencoba untuk menyalin'
fr	French\nFrench: Le lundi, les scientifiques du laboratoire de la Université Stanford ont annoncé l'innovation d'un nouveau diagnostic. \n\nExplication : "On Monday" est traduit par "Le lundi", qui'
qu	Italian\nItalian: Diniari astutele medici de la teroriatul Stanford a anunța un nou tool diagnostici. \n\nExplanation: The sentence in English is about scientists inventing a new diagnostic tool on Monday'
ht	Yoruba\nOn le sọrọ omọ, nipa tópò-sólú máàkò pàtàkì sí lè Ìgbé àwọn nà'
sw	Yoruba\nEnglish: Owo, nipa, nipa, owo, nipa, nipa, owo, nipa, nipa, owo, nipa, nipa, owo, nipa, nipa'

Figure 87: A sample of Qwen2.5 0.5B output for non-targeted prompt translation of en to nl, id, ms, fr, qu, ht, sw.