Distribution-aware Language Neuron Identification in Multilingual Large Language Models

Minjun Kim[1] **Inho Won**[1] **Junghun Yuk**[1] **Dongyeon Kim**[1]
Jihyo Kim[2†] **KyungTae Lim**[1†]
[1]KAIST [2] KAIST InnoCORE PRISM-AI Center
{mjkmain, ktlim}@kaist.ac.kr

Abstract

Multilingual large language models (mLLMs) contain a small fraction of feed-forward neurons that are sensitive to particular languages, commonly termed language-specific neurons. Existing work measures language specificity using the entropy of each neuron's language-wise probabilities of being active, where a neuron is considered active when its activation value is positive. However, this approach may not fully capture the multilingual nature of mLLMs, where language representations are distributional and mutually related. We propose *Distribution-aware Language Neuron* selection, which leverages pairwise relationships between per-language activation distributions over the full activation range, including negative values. Specifically, we quantify each neuron's language specificity by clustering languages using pairwise overlap coefficients between their activation distributions. Across two mLLMs and two held-out corpora, our identifier more effectively isolates language-specific causal effects, yielding up to 4.9× higher on-target language damage per neuron while preserving off-target language performance.

1 Introduction

Multilingual large language models (mLLMs) exhibit strong capabilities across dozens of languages within a single shared parameter space (Scao et al., 2022; Yang et al., 2025). However, how these models internally organize multilingual representations remains an open question.

One empirical approach to this question is neuron-level interpretability based on activation values. Previous studies identify language-specific neurons (LSNs) as neurons that exhibit preferential activation for particular languages, based on activation profiles that distinguish one language from the others (Tang et al., 2024; Kojima et al., 2024; Zhang et al., 2026). They consider neurons with positive activation values to be activated for each language and quantify language specificity using the entropy of per-language activation probabilities. In this view, neurons exhibiting language specificity can be conceptualized as "neurons that activate predominantly for a single language." These identified neurons provide a useful analytic tool for studying the mechanisms of cross-lingual transfer, intervening on language generation, and probing multilingual knowledge organization (Rahmanisa et al., 2025; Mondal et al., 2025; Deng et al., 2025).

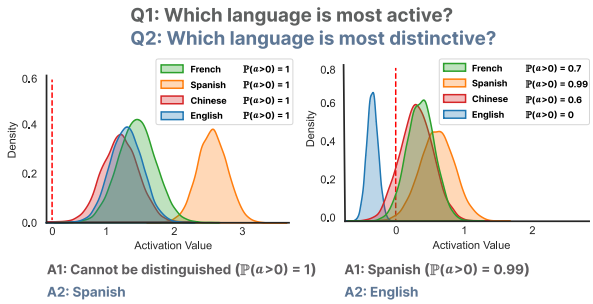


Figure 1: Identifying structurally distinctive language distributions matters more than identifying the most frequently active language. Here, a denotes the activation value. A distributional view reveals distinctions that activation probabilities alone miss.

However, previous LSN identification methods may not fully reflect the multilingual nature of mLLMs. Since mLLMs are trained on multiple languages under a shared parameter space, language representations are inherently distributional and mutually related. Binarizing continuous activations and treating them as separate per-language distributions introduces two critical limitations: first, it discards informative signals from negative activations since neurons with negative activation values are treated as inactive; second, because activation probabilities are computed from a binary state separately for each language, they do not preserve the unique shape of per-language activation distributions and the relational structure among languages.

[†]Corresponding Authors

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 26366–26393
October 24-29, 2026 ©2026 Association for Computational Linguistics

Figure 1 provides an example where the previous activation probability methods are insufficient to capture the language specificity. As shown in the left plot, even when all activation values are positive, previous methods may fail to identify this neuron as language-specific because the entropy remains high despite distinct activation distributions. Furthermore, the right plot shows a scenario in which focusing solely on activation probabilities may fail to capture the structural distinctiveness: in this example, English forms an isolated distribution in a region that previous methods would treat as inactive. These structural blind spots raise a fundamental question: given the multilingual nature of mLLMs, how can we identify language neurons by leveraging the activation distribution and its relational structure across languages while avoiding potential information loss?

To address this question, we propose *Distribution-aware Language Neuron* (DLN) selection, which estimates neuron-level language specificity from pairwise overlap coefficients between per-language activation distributions. By modeling the full range of activation values and clustering languages based on their pairwise overlap coefficient, DLN captures the relational structure among activation distributions. This formulation naturally identifies two types of language neurons, (1) *single-language neurons* (SLNs), specific to one language, and (2) *multi-language neurons* (MLNs), specific to multiple languages. Consequently, rather than focusing solely on neurons that activate predominantly for a single language, our method reframes the neurons to be identified as "neurons whose activation distributions differ across languages."

Through evaluations on Llama-3.1-8B (Grattafiori et al., 2024) and SmolLM3-3B-Base (Bakouch et al., 2025) over two held-out corpora, with additional validation on three further models in the appendix, we show that DLN more effectively isolates language-specific causal effects than activation-probability baselines. Ablating DLN-selected SLNs yields up to 4.9× higher on-target language damage per neuron while preserving off-target performance. DLN also captures MLNs, uncovering shared structures such as a Chinese–Japanese coalition that yields 7.12× greater within-cluster damage while using 2.1× fewer neurons. By modeling activation densities, DLN identifies neurons operating in the negative activation regime, including English SLNs that are invisible to positive-rate metrics yet act as functional triggers for target-language control. Finally, DLN remains effective beyond the standard gate-activation site, maintaining high selectivity at the gate-up product site where activation-probability baselines lose much of their effectiveness. Our contributions are as follows:

- **A distribution-aware language neuron identification framework.** By formulating neuron selection as a clustering problem using overlap coefficients of continuous activation densities, our method effectively leverages the distributional and inter-language relational structures reflecting the nature of mLLMs.
- **Effectiveness across SLNs and MLNs.** We evaluate the effectiveness of the neurons identified by our method for two types of language neurons. SLNs significantly outperform the baselines in isolating language-specific causal effects, while MLNs identify structurally coupled language coalitions.
- **What full activation distributions reveal.** Analysis results demonstrate that our method achieves its design objectives: (1) capturing the importance of neurons identified in the negative activation regime, and (2) enabling flexible intervention locations by considering the full range of activation values.

2 Related Work

mLLMs share parameters across many languages while preserving language identity in generation (Conneau et al., 2020; Grattafiori et al., 2024; Yang et al., 2025). Prior work suggests that languages occupy partially overlapping representational subspaces (Pires et al., 2019; Wu and Dredze, 2019) and that decoder-only mLLMs may internally rely on English-centric pivot representations (Wendler et al., 2024; Schut et al., 2025). These findings motivate a mechanistic question: how are language-specific and language-shared computations implemented inside a single mLLM?

Neuron-level analyses provide one way to study this question. Prior work shows that individual neurons can encode factual, syntactic, semantic, or task-relevant information in Transformer models (Geva et al., 2021; Dai et al., 2022; Stanczak et al., 2022). Extending this line to multilingual models, Kojima et al. (2024) identify language-specific neurons (LSNs) by measuring how pre-

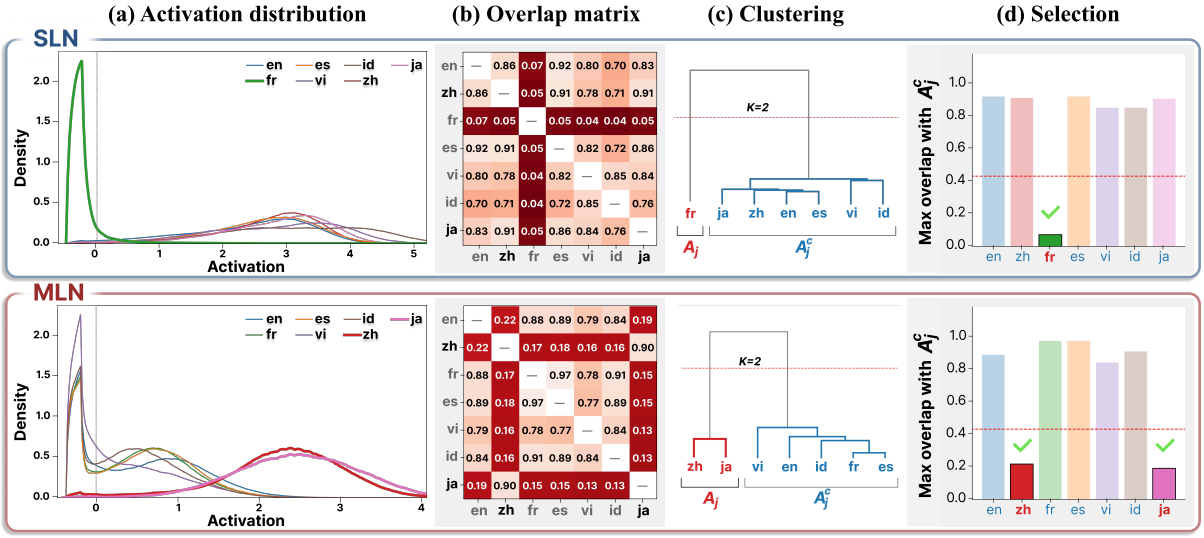


Figure 2: Overall framework of the proposed method.

cisely each neuron distinguishes a target language from others. Tang et al. (2024) introduce language activation probability entropy (LAPE), which measures language specificity from the entropy of per-language activation probabilities, treating neurons with positive activation values as active. Based on LAPE, Zhang et al. (2026) categorize neurons as LSNs, language-related neurons (LRNs), and general neurons based on how many languages exceed a predefined activation-probability threshold.[1]

However, these methods mainly rely on binary activation statistics, i.e., whether and how often a neuron is active for each language. This abstraction can miss distributional distinctions: languages with similar activation probabilities may still induce different activation-value distributions, and vice versa. In contrast, we analyze the full per-language activation distribution, enabling the identification of both single-language and multi-language neurons from distributional activation patterns.

3 Distribution-aware Language Neurons

In this section, we introduce *distribution-aware language neuron (DLN)* identification, which leverages the full range of per-language activation distributions and quantifies their relationships using pairwise overlap coefficients. Previous methods reduce each per-language activation distribution to a binary activation statistic based on the sign of activation values, although negative activations may contain useful information for determining language specificity. Moreover, multiple languages can have nearly identical binary activation statistics while occupying clearly separable activation regimes. Such neurons may still be language-specific, but entropy-based methods can fail to identify them. Figure 2 illustrates the overall framework of our method.

3.1 Distribution Overlap Coefficient Matrix

We study autoregressive Transformer language models with Gated Linear Unit (GLU) feed-forward blocks (Shazeer, 2020). We index each multilayer perceptron (MLP) neuron by the pair $j = (\ell, n)$, where ℓ is its layer and n is its position within the layer's intermediate dimension d_{ffn}. Let d denote the residual-stream hidden dimension. Given the gate projection $W^{\ell}_{\text{gate}} \in \mathbb{R}^{d_{\text{ffn}} \times d}$, the activation of neuron j on input $x \in \mathbb{R}^d$ is

$$a_j(x) \;=\; \left[\sigma(W^{\ell}_{\text{gate}} x)\right]_n,$$

where $\sigma(\cdot)$ is the activation function (e.g., SiLU).

We determine each neuron's language specificity by measuring the overlap coefficients, the shared areas, between per-language activation distributions. Let $\mathcal{L}$ denote the set of languages, and $p_j^k(v)$ denote the continuous probability density function of the activation value $v = a_j(x)$ for neuron j under language $k \in \mathcal{L}$. The pairwise overlap coefficient matrix $\mathbf{S}_j \in [0,1]^{|\mathcal{L}| \times |\mathcal{L}|}$ is calculated as:

$$\mathbf{S}_j[k, k'] = \int \min\left(p_j^k(v),\, p_j^{k'}(v)\right) dv.$$

This summarizes the relationship among $\{p_j^k(v)\}$.

[1] LSNs and LRNs are closely related to our notions of single-language neurons (SLNs) and multi-language neurons (MLNs), respectively. We use SLNs and MLNs to emphasize whether the identified neurons are associated with a single language or shared across multiple languages.

The overlap coefficient is one of several bounded, shape-sensitive statistics for comparing per-language distributions, and we adopt it as the most intuitive and computationally cheapest choice. Appendix E shows that alternatives such as Jensen-Shannon, squared Hellinger, and Kolmogorov-Smirnov yield near-identical neuron sets, whereas scale-bearing metrics (Wasserstein-1, energy distance (Székely and Rizzo, 2013)) diverge.

3.2 Cluster-based Neuron Classification

To distinguish languages with distinct activation patterns, we partition $\mathcal{L}$ into two clusters, (A_j, A_j^c), by minimizing the maximum between-cluster overlap coefficient, $\max_{k \in A_j,\, k' \in A_j^c} \mathbf{S}_j[k, k']$, over all non-trivial bipartitions. This procedure is equivalent to single-linkage clustering (Johnson, 1967) with the number of clusters $K = 2$. This is a selection decision rather than a bimodality assumption, as $\mathbf{S}_j$ retains all pairwise relations and a neuron with three or more separated groups is still detected through its most-separated bipartition. Appendix E validates this design together with alternative linkage methods.

We define A_j to be the smaller cluster and designate it as the neuron's candidate language set. This follows from the assumption of our method design that languages with distinct activation patterns are those most strongly associated with the neuron. The larger cluster A_j^c can be regarded as a reference group of languages with relatively similar activation behavior.

3.3 Neuron Selection Criterion

After obtaining A_j and A_j^c, we select neuron j as a language neuron only if all between-cluster overlap coefficients do not exceed a threshold τ:

$$\mathcal{N} = \left\{ j : \max_{k \in A_j, k' \in A_j^c} \mathbf{S}_j[k, k'] \leq \tau \right\}.$$

Otherwise, if any between-cluster pair has an overlap coefficient greater than τ, the neuron is not selected.

We denote each selected language neuron as $j^\star \in \mathcal{N}$ and categorize it according to the size of its language set $A_{j^\star}$ as follows:

- $|A_{j^\star}| = 1$: **single-language neuron (SLN)**,
- $|A_{j^\star}| > 1$: **multi-language neuron (MLN)**.

Because overlap coefficients vary across models, we set the threshold τ to the bottom P-th percentile of overlap coefficients computed over all neurons.

4 Experiments

4.1 Experimental Settings

Baselines. We employ two language neuron identification methods using the entropy of activation probabilities, LAPE (Tang et al., 2024) and Zhang et al. (2026), as our baselines. Specifically, LAPE identifies LSNs by their defined score computed as:

$$\mathrm{LAPE}_j \;=\; -\sum_{k \in \mathcal{L}} \tilde{r}_j^k \log \tilde{r}_j^k,$$

where $\tilde{r}_j^k$ is L1 normalized $r_j^k = \mathbb{P}\big(a_j(x) > 0 | k\big)$ across k.

Zhang et al. (2026) additionally incorporate the maximum activation probability, $\max_{k \in \mathcal{L}} r_j^k$, to quantify the strongest degree to which neuron j is active across languages to improve language neuron selection. Accordingly, they revise the LAPE score as follows:

$$\mathrm{score}_j \;=\; -\sum_{k \in \mathcal{L}} \tilde{r}_j^k \log \tilde{r}_j^k \;-\; \lambda \max_{k \in \mathcal{L}} r_j^k,$$

where λ is a balancing coefficient. The λ for each evaluated model can be found in Table 6 in Appendix A. In addition, they distinguish LRNs from LSNs as neurons involved in multiple languages.[2] Specifically, neuron j is considered an LRN if the number of languages for which r_j^k exceeds the predefined threshold is greater than 1.

Models. We study two pretrained base models with Gated Linear Unit feed-forward design (Shazeer, 2020) and SiLU activation: Llama-3.1-8B (Grattafiori et al., 2024) as our primary model, with SmolLM3-3B-Base (Bakouch et al., 2025).

Corpora and languages. For language neuron selection, we sample 100k tokens per language from `wikipedia-20231101`. For evaluation, we use FLORES+ devtest (Team et al., 2024; FLORES) and a held-out Wikipedia split (`wikipedia-20230601`; WIKI), up to 1M tokens per language, disjoint from the calibration source.

Following Tang et al. (2024), we use seven languages spanning four scripts (Latin, Han, Kanji, and Kana) and five language families (Indo-European, Sino-Tibetan, Austroasiatic, Austronesian, and Japonic): English (`en`), Chinese (`zh`), French (`fr`), Spanish (`es`), Vietnamese (`vi`), Indonesian (`id`), and Japanese (`ja`).

[2] Hereafter, for clarity, we refer to LSNs identified by Tang et al. (2024) as LAPE, and to LSNs and LRNs identified by Zhang et al. (2026) as LSN and LRN, respectively.

Density approximation. To compute the continuous overlap integral in practice, we approximate the density $p_j^k(v)$ using a histogram-based approach with 400 inner bins. The bins are defined on a per-neuron adaptive range $[\mu_j - 3\sigma_j,\ \mu_j + 3\sigma_j]$, where μ_j and σ_j are pooled estimates over the seven languages.

Neuron intervention. For neuron $j^\star$ identified for target k, we replace its $a_{j^\star}$ with the mean activation value over the non-target languages. This mean-patching intervention aligns with our method: we treat p_j^k as primitive and therefore call for replacing $a_{j^\star}$ with a non-target reference value estimated from the other languages, whereas the baseline methods operationalize language specificity through the binary event $a_{j^\star}(x) > 0$, for which zero serves as a natural counterfactual. Mean-patching realizes this intervention at the first-moment level while keeping the replacement within the neuron's typical activation range (Heimersheim and Nanda, 2024). We verified that mean-patching does not adversely affect the baselines compared with zero-ablation, as detailed in Figures 7, 8 in Appendix B.

Selectivity score. Let $\mathcal{A}$ denote the collection of all target language clusters under evaluation. For $A \in \mathcal{A}$, let A^c denote its complement $\mathcal{L} \setminus A$. Mean-patching its neuron subset $\mathcal{N}_A$ induces a negative log-likelihood (NLL) increase $\Delta\mathrm{NLL}_k(\mathcal{N}_A)$ on evaluation language k. The selectivity of neuron set $\mathcal{N}$ is:

$$\mathrm{Sel}(\mathcal{N}) := \frac{\sum_{A\in\mathcal{A}}\left[\frac{1}{|A|}\sum_{k\in A}\Delta\mathrm{NLL}_k(\mathcal{N}_A)\right]}{\sum_{A\in\mathcal{A}}\left[\frac{1}{|A^c|}\sum_{k'\in A^c}\Delta\mathrm{NLL}_{k'}(\mathcal{N}_A)\right]}.$$

For instance, in Figure 3, $\mathrm{Sel}(\mathcal{N})$ is computed as the ratio between the diagonal sum and the off-diagonal sum. Higher $\mathrm{Sel}(\mathcal{N})$ means the intervention concentrates damage on the target languages A and spares the off-target languages A^c.

4.2 Assessing the Neuron Selectivity

We evaluate the effectiveness of our method from two perspectives: (1) for SLNs, whether it yields more language-selective per-language interventions than the baselines; and (2) for MLNs, whether our clustering-based framework recovers cluster-internal intervention effects missed by single-language identifiers and concentrates these effects more strongly per neuron than LRN. To examine

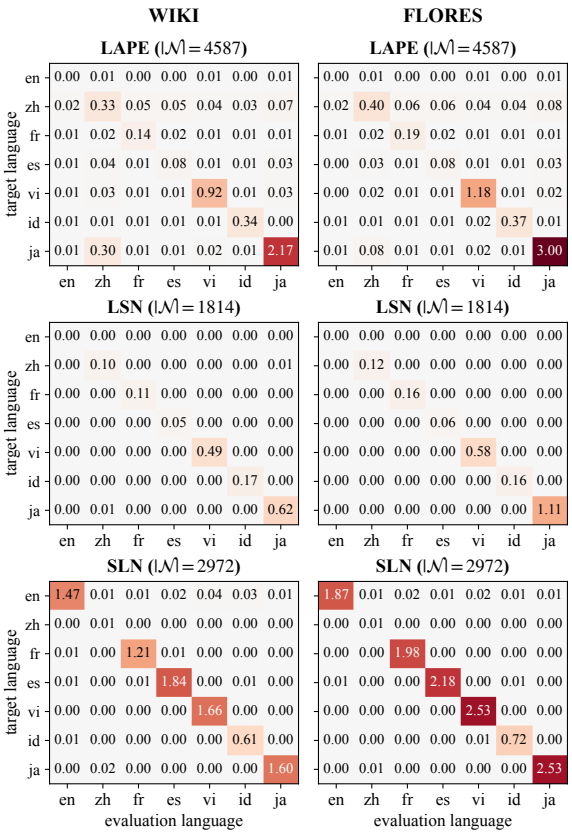


Figure 3: ΔNLL by language on Llama-3.1-8B for LAPE (top) and LSN (middle), and our SLN at $P = 1\%$ (bottom) on WIKI (left) and FLORES (right). $|\mathcal{N}|$ represents the number of ablated neurons.

these two perspectives, we analyze changes in NLL across languages after intervening on neurons associated with the target language.

Single-language neurons. Figure 3 compares the per-language ΔNLL matrices for LAPE, LSN, and SLN on Llama-3.1-8B. The three identifiers select different pool sizes (4,587 / 1,814 / 2,972); we therefore report on-target damage normalized by $|\mathcal{N}_{\{k\}}|$ alongside the raw ΔNLL. Three observations follow. (i) Averaged over the seven languages, SLN's on-target ΔNLL per neuron is $4.9\times$ (WIKI) and $5.4\times$ (FLORES) larger than that of LAPE, and 3.1 / $3.2\times$ larger than LSN. (ii) The advantage is most pronounced for English, which both baselines fail to identify as a language-specific subset: their English-assigned subsets produce no measurable damage when ablated, whereas our SLN delivers $\Delta\mathrm{NLL}_{\mathsf{en}}(\mathcal{N}_{\{\mathsf{en}\}}) = 1.47$ (WIKI) and 1.87 (FLORES). (iii) The exception is Chinese: the zh diagonal is weak under all three identifiers (0.33 / 0.10 / 0.01 on WIKI). Chinese activations are largely captured by the {zh, ja} multi-language cluster, and the damage is recovered under the MLN setting.

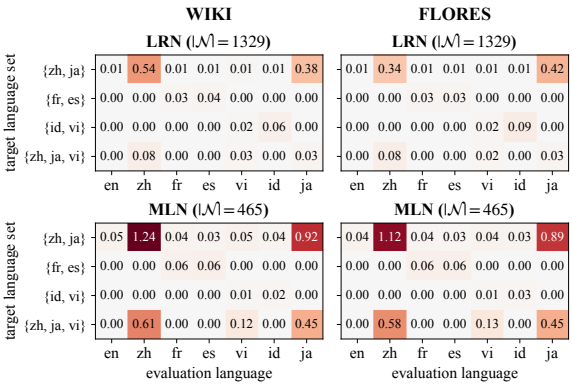


Figure 4: ΔNLL by cluster on Llama-3.1-8B for LRN (top) and MLN at P=1% (bottom) on WIKI (left) and FLORES (right). Each row corresponds to one of the top-4 most frequent MLN clusters under ablation, columns are the seven evaluation languages.

Method	{zh, ja}	{fr, es}	{id, vi}	{zh, ja, vi}
LRN	407	467	268	187
MLN	232	129	26	78

Table 1: Number of neurons that LRN and our MLN assign to each of the four most frequent multi-language sets on Llama-3.1-8B (P=1%). Summed over the four sets, LRN admits 1,329 neurons against our 465.

Multi-language neurons. Figure 4 contrasts LRN with our MLN on the four most-populated cluster compositions (Table 1). We make three observations. (i) Although our MLN uses about 2.86× fewer neurons than LRN (465 vs. 1329), the within-cluster damage is substantially stronger: averaged on-target ΔNLL per neuron over the four clusters, our MLN delivers 6.85× / 7.12× (WIKI / FLORES) the damage of LRN. (ii) The most notable case is Chinese, which all three single-language level identifiers miss: 93.3% of the Chinese-associated neurons identified by the proposed method are classified as MLNs (Table 22). It is the only language where a single MLN cluster accounts for the majority, with the {zh, ja} cluster comprising 57.6% of its language neurons. Consequently, for the {zh, ja} cluster, the Chinese damage registers $\Delta\text{NLL}_{\texttt{zh}}\left(\mathcal{N}_{\{\texttt{zh},\texttt{ja}\}}\right) = 1.24/1.12$ on WIKI / FLORES, with within-cluster Japanese damage 0.92/0.89 supporting a Chinese–Japanese coalition, reflecting their shared character-level structure. Chinese language neurons are thus not absent from the model but rather are organized as a multi-language cluster, not as a single-language set. (iii) The {fr, es} and {id, vi} clusters contribute only marginal MLN signal: for most languages, the proportion of SLNs is higher than that of MLNs. Specifically, the {fr, es} cluster accounts for only 13.2% of the fr-SLNs and 15.4% of the es-SLNs. The {id, vi} cluster also represents a small proportion compared to SLNs, comprising just 2.6% of the id-SLNs and 7.5% of the vi-SLNs. Therefore, these four languages are already captured at the single-language level, as their SLN diagonals reach ΔNLL ≥ 0.6 on WIKI and ≥ 0.7 on FLORES. Additional comparisons of SLN and MLN distributions are provided in Appendix I.1.

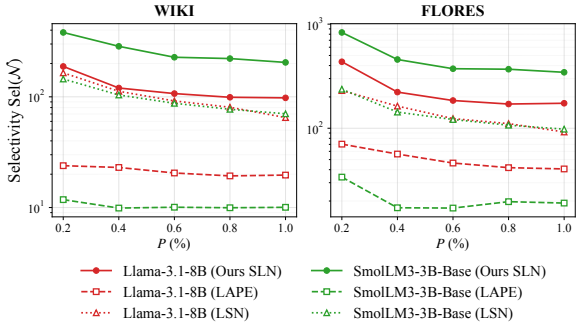


Figure 5: Selectivity of SLN across the seven languages over percentile P. Baselines are matched to our neuron count at every P.

Selectivity robustness. Figure 5 reports selectivity score across the seven languages depending on P, for two models on both corpora. The selected neurons at percentile P comprise the lowest P% of clustering candidates, jointly retaining single-language ($|A_j| = 1$) and multi-language ($|A_j| \in \{2, 3\}$) neurons. We match the baseline neuron budgets to ours at every P.

Two patterns hold consistently. First, our selectivity exceeds both LAPE and LSN at every P in the sweep on every (model, corpus) combination: the gap over LAPE reaches an order of magnitude on SmolLM3-3B-Base (20–32× on WIKI, 18–26× on FLORES) and stays at 4–8× on Llama-3.1-8B, while the gap over LSN is smaller but consistent (1.1–1.9× on Llama-3.1-8B and 2.6–3.5× on SmolLM3-3B-Base). Second, selectivity decreases monotonically with P. Since percentile thresholding admits top ranked candidates first, smaller P retains only the most concentrated subset, which is the expected behavior of a well-ordered ranking. Section 5.2 further verifies that the reported on-target damage is not a mechanical consequence of the mean-patch displacement magnitude. In an extended sweep to P=5%, selectivity remains an order of magnitude above the budget-matched LAPE throughout (Appendix D.6).

Method	Belebele			MGSM		
	$\lvert\mathcal{N}\rvert$	$k'{=}k$	$k'{\neq}k$	$\lvert\mathcal{N}\rvert$	$k'{=}k$	$k'{\neq}k$
Llama-3.1-8B						
LAPE	4,587	2.75	0.07	3,088	6.96	0.88
LSN	1,814	1.22	−0.02	1,169	4.40	0.30
SLN (ours)	2,972	**4.94**	−0.02	1,858	**9.44**	0.18
SmolLM3-3B-Base						
LAPE	3,962	2.46	0.14	2,499	3.60	0.06
LSN	1,443	1.25	0.01	867	0.72	−0.64
SLN (ours)	2,254	**5.56**	0.15	1,331	**7.76**	−0.40

Table 2: Accuracy drop under the mean-patch intervention, averaged over target languages. $k'{=}k$ is the on-target drop and $k'{\neq}k$ the mean off-target drop. Belebele covers all seven languages, MGSM only five (`en`, `zh`, `fr`, `es`, `ja`). **Bold** denotes the largest on-target drop.

Neuron Pool	$\lvert\mathcal{N}\rvert$	WIKI			FLORES		
		$k'=k$	$k'\neq k$	$\mathrm{Sel}(\mathcal{N})$	$k'=k$	$k'\neq k$	$\mathrm{Sel}(\mathcal{N})$
Llama-3.1-8B							
LAPE \ SLN	3598	0.103	0.021	4.95	0.111	0.018	6.30
LAPE ∩ SLN	989	0.265	0.001	**203.07**	0.395	0.001	**460.62**
SLN \ LAPE	1983	0.295	0.004	<u>75.17</u>	0.329	0.003	<u>108.83</u>
SmolLM3-3B-Base							
LAPE \ SLN	3213	0.150	0.049	3.06	0.176	0.038	4.62
LAPE ∩ SLN	749	0.331	0.001	**238.92**	0.422	0.001	**408.44**
SLN \ LAPE	1505	0.707	0.003	<u>204.26</u>	0.878	0.003	<u>340.75</u>

Table 3: Mean selectivity across the seven languages, obtained by mean-patching the intersection and set-difference neurons of LAPE and SLN. The $k' = k$ and $k' \neq k$ columns report the average $\Delta\mathrm{NLL}_k\left(\mathcal{N}_{\{k'\}}\right)$, and $\mathrm{Sel}(\mathcal{N})$ the resulting selectivity. **Bold** and <u>underline</u> denote the highest and second-highest, respectively.

4.3 Effect on Downstream Tasks

The evaluation so far measures the damage of an intervention as a change in token-level likelihood, ΔNLL. We now ask whether the same intervention also changes accuracy on multilingual downstream tasks. We evaluate on Belebele (Bandarkar et al., 2024), a multiple-choice reading comprehension task (900 questions per language), and MGSM (Shi et al., 2023), a multilingual grade-school math benchmark (250 problems per language), with prompts and scoring detailed in Appendix G.1. For each language neuron identification method and each target language k covered by the task, we repeat the mean-patch intervention, replacing the activations of $\mathcal{N}_{\{k\}}$ with their off-target means, and report the accuracy change on every evaluation language relative to the unintervened model.

Table 2 shows that the NLL result carries over to both tasks. Notably, our SLNs induce the largest on-target drop in every (model, task) combination, while off-target accuracy remains essentially unchanged under all three identifiers. The damage is concentrated rather than indiscriminate: on Belebele, whose chance accuracy is 25.0 and whose unintervened accuracy is 47.0 / 45.2 (Llama-3.1-8B / SmolLM3-3B-Base), our intervention removes 22% / 27% of the model's above-chance margin on the target language, yet costs no off-target language more than 2.40 points. Our larger on-target drop is not an artifact of the number of neurons patched. Although LAPE patches up to 1.9× as many neurons as we do, our SLNs still remove more on-target accuracy per neuron than both baselines, by up to 4.0× over LAPE and 7.0× over LSN. Distributional separation therefore identifies neurons whose causal effect extends beyond token-level likelihood to task behavior.

Appendix G.2 reports the results for each target language, where the Chinese entry is near zero, as in the ΔNLL evaluation. Our identifier classifies almost all of the Chinese language neurons as members of the {`zh`, `ja`} MLN cluster, leaving the Chinese SLN pool nearly empty, so the Chinese effect surfaces under the MLN setting of Section 4.2.

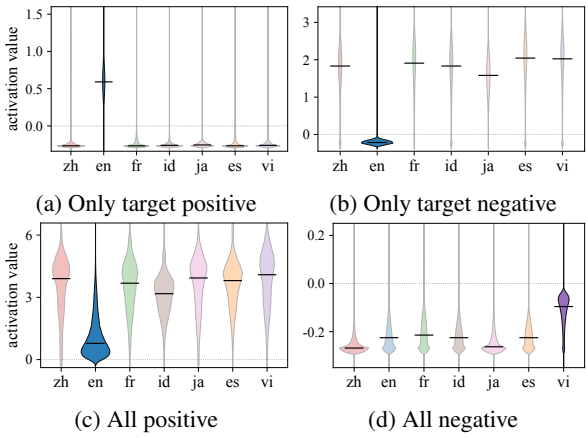
5 Analysis

5.1 A Comparison between LAPE and SLN

We first examine the following question: What are the key differences between our single-language neuron and the neurons identified by LAPE? To analyze these differences, we define and investigate the corresponding neuron pools and their subrelations: the intersection (LAPE ∩ SLN), and the set differences (LAPE \ SLN and SLN \ LAPE).

Effect on selectivity. We evaluate the functional importance of these distinct neuron pools by measuring their causal impact on language selectivity. As shown in Table 3, selectivity is highest for the intersection (LAPE∩SLN), followed by SLN\LAPE and LAPE \ SLN. This implies that neurons satisfying both methods are the most effective language neurons. Notably, SLN \ LAPE delivers a 2.86× (Llama-3.1-8B) and 4.71× (SmolLM3-3B-Base) higher on-target $\Delta\mathrm{NLL}_k(\mathcal{N}_{\{k\}})$ than LAPE \ SLN (0.295 vs. 0.103; 0.707 vs. 0.150), despite using roughly half as many neurons.

Distributional diversity of SLNs. We analyze the distributional characteristics of the SLNs identified by our method. As illustrated in Figure 6, the selected neurons can be categorized into four patterns: (a) only target positive, where only the



(a) Only target positive (b) Only target negative

(c) All positive (d) All negative

Figure 6: Four representative activation patterns of single language neurons across seven languages.

target language has positive activations; (b) only target negative, where only the target language has negative activations; (c) all positive, where all languages have positive activations but the target language shows a distinct magnitude shift; and (d) all negative, where all activations are negative but the target language remains distributionally separated. Unlike indicator-based methods, which are limited to the target-positive pattern, our distribution-aware approach successfully uncovers this broader spectrum of language neurons. For instance, Table 8 shows that 63–72% of English SLNs have a negative target-language mean, making them entirely undetectable to positive-rate metrics despite their functional relevance.

Functional role of negative-region neurons. To investigate what role these negative-region neurons serve in generation, we partition English neurons by the sign of μ_{tgt}, the mean activation on target-language inputs, and steer generation with mean-patching in both translation directions. Source-language neurons are patched to the off-source mean and target-language neurons are patched to the target mean, thereby specifying the direction of the language switch. We evaluate 200 samples per language from each of the WIKI and FLORES datasets, yielding 400 samples per language in total. To verify whether the steered output is emitted in the intended target language, we measure language accuracy, the fraction of generations whose top-1 fastText prediction (Joulin et al., 2016) matches the target, with LAPE included as a baseline.

The results in Table 4 reveal a clear functional split. When the target output is English, $\mathcal{N}^{+}_{\{\texttt{en}\}}$ and LAPE achieve near-perfect switching accuracies of 0.96 and 0.97, respectively, while the negative-region neurons are markedly weaker at 0.71. This pattern reverses when the target output is a non-English language k: only the negative subset is effective. Specifically, $\mathcal{N}^{-}_{\{\texttt{en}\}}$ drives the switch with average language accuracy 0.75, whereas $\mathcal{N}^{+}_{\{\texttt{en}\}}$ and LAPE are almost ineffective with accuracies of 0.15 and 0.01, respectively. This effect is not merely an artifact of set size, since $\mathcal{N}^{-}_{\{\texttt{en}\},\text{topk}}$, which is matched in size to the positive subset, retains strong performance of 0.76, indicating that the negative-region neurons drive the switch functionally rather than through the difference in set size.

Direction	k	LAPE	DLN		
			$\mathcal{N}^{+}_{\{\texttt{en}\}}$	$\mathcal{N}^{-}_{\{\texttt{en}\}}$	$\mathcal{N}^{-}_{\{\texttt{en}\},\text{topk}}$
en → k	zh	0.00	0.39	**0.74**	0.72
	fr	0.01	0.00	**0.90**	0.89
	ja	0.06	0.49	**0.71**	0.70
	es	0.01	0.00	0.74	**0.78**
	vi	0.00	0.00	0.82	**0.87**
	id	0.00	0.00	0.57	**0.60**
Average		0.01	0.15	0.75	**0.76**
k → en	zh	**0.96**	0.95	0.79	0.66
	fr	0.96	**0.98**	0.78	0.36
	ja	**0.97**	0.93	0.65	0.58
	es	0.98	**0.99**	0.90	0.54
	vi	**0.98**	**0.98**	0.59	0.21
	id	**0.95**	**0.95**	0.55	0.20
Average		**0.97**	0.96	0.71	0.43

Table 4: Language accuracy of steered output. LAPE uses its native 61 English neurons. DLN splits its identified English neurons by mean-activation sign on English inputs into $\mathcal{N}^{+}_{\{\texttt{en}\}}$ (68 neurons) and $\mathcal{N}^{-}_{\{\texttt{en}\}}$ (196 neurons), and $\mathcal{N}^{-}_{\{\texttt{en}\},\text{topk}}$ denotes the top-68 subset of $\mathcal{N}^{-}_{\{\texttt{en}\}}$ under the same selection ranking. **Bold** and underline denote the highest and second-highest accuracy, respectively.

In other words, the positive-region neurons are responsible for producing English output, whereas the negative-region neurons form a functionally distinct circuit that switches away from an English context into another language. Further details are provided in Appendix D.5. Whether such interventions can also improve task performance is a separate research problem, and Mondal et al. (2025) report negative results for language-specific neurons facilitating cross-lingual transfer. Our scope is the faithful identification of language neurons, validated causally and on downstream tasks, rather than performance enhancement.

5.2 Displacement-matched Control

Because mean-patching moves each neuron by the distributional separation it was selected for, one may ask whether the large on-target ΔNLL merely reflects a large displacement. To test this, we apply the intervention to neurons randomly sampled from outside the DLN-selected pools and count-matched on every layer. Each is shifted from its target-language mean by the same number of standard deviations as its matched SLN (1.6–2.2σ; 3 seeds). This matched displacement produces only $\sim$1% of the SLN effect (mean on-target ΔNLL of 0.013 vs. 1.198) and no language selectivity. Its on-target and off-target damage coincide (0.013 each), whereas the SLN intervention leaves off-target damage at only 0.006. The pattern replicates on SmolLM3-3B-Base (0.010 vs. 1.395). The reported ΔNLL thus reflects which neurons are patched, not how far they are moved. Per-language results are provided in Appendix F.

5.3 Identification across GLU Sites

A GLU feed-forward block exposes two natural signals: the gate activation $a_j(x) = [\sigma(W^{\ell}_{\text{gate}}x)]_n$ (the a-site) and the gate-up product $z_j(x) = a_j(x) \otimes u_j(x)$ (the z-site), where $u_j(x) = [W^{\ell}_{\text{up}}x]_n$. Prior neuron-level interpretability has focused on the a-site for its clean above-zero criterion (Tang et al., 2024; Dai et al., 2022; Geva et al., 2021), yet the z-site is the signal that is causally operative. We therefore ask whether our distribution-aware identifier extends to the z-site. The two sites differ sharply in sign: pooled across neurons and the seven languages, a is negative-dominant ($\mathbb{P}(a > 0) \approx 0.25$) while z is sign-balanced ($\mathbb{P}(z > 0) \approx 0.50$; Table 9). This matters because baselines rank neurons by the entropy of activation probabilities, which requires the activation rate to vary across languages. At the a-site this rate spans an approximately 3$\times$ gap between the most- and least-active languages, so the entropy is informative; at the z-site the sign-balanced product flattens every language to ≈ 0.5, collapsing this gap and saturating the entropy.

Our identifier demonstrates robustness when extending the selection from a to the z-site, retaining most of its efficacy with a decrease of 29–37%. Consequently, at the z-site, our method maintains a substantial advantage, outperforming LAPE by 25–44$\times$ and LSN by 15–21$\times$. In contrast, the baselines lose 74–90% of their selectivity under this transition, confirming that the performance drop is driven by the difficulty of selecting at the z-site, not by the change of intervention site.

	Selection Method		
(selection, patch)	LAPE	LSN	SLN
Llama-3.1-8B			
(a-site, a-site)	22.8	126.0	**273.7**
(z-site, a-site)	4.5	13.3	**193.9**
(z-site, z-site)	3.9	9.9	**97.2**
SmolLM-3-3B-Base			
(a-site, a-site)	28.2	89.5	**296.5**
(z-site, a-site)	7.4	8.8	**187.3**
(z-site, z-site)	5.3	9.2	**57.4**

Table 5: Mean selectivity across the seven languages on WIKI. Each row varies the neuron-selection site and the patching site. **Bold** denotes the highest selectivity.

In the (z-site, z-site) row, the cross-language mean spread at z-site is about 4$\times$ narrower than at a-site (per-neuron median 0.018 vs. 0.077 on Llama-3.1-8B), so mean-patching substitutes a smaller magnitude; this drop is uniform across methods and leaves the (z, a) ordering intact. That our criterion survives the move to z shows distribution-aware identification applies at the causally operative signal. This locus remains underexplored under the gate-only convention and is a natural target for mechanistic analysis.

6 Conclusion

In this paper, we propose a *distribution-aware language neuron* selection method, which considers the full range of activation values and the pairwise relationship among language activation distributions, to better reflect the multilingual nature of mLLMs. Experimental results demonstrate that our method effectively selects neurons that influence multilingual abilities. In addition, the identified MLNs form coherent language groups whose members exhibit relatively large performance degradation when intervened on. Further analysis shows that language neurons selected from the negative activation regime play functional roles in language switching during generation. Finally, we demonstrate that our method can also be applied to other feasible locations in the model, such as the gate-up product (z-site) in addition to the standard gate activation. Overall, our work demonstrates that comprehensively modeling activation distributions provides an effective framework for identifying language neurons in mLLMs.

Limitations

Limited language coverage. The experiments and language neuron selection sample only seven languages (English, Chinese, French, Spanish, Vietnamese, Indonesian, and Japanese). While these cover multiple scripts and language families, they represent only a tiny fraction of the languages that modern multilingual LLMs are trained on.

Architectural restriction to GLU. A key limitation of this work is its exclusive focus on autoregressive Transformer models equipped with Gated Linear Unit (GLU) feed-forward blocks with the SiLU activation function. Consequently, it remains unexplored whether this distribution-aware neuron identification generalizes effectively to models employing standard MLP architectures that lack this bifurcated signal structure.

Intervention penalty at the z-site. When shifting identification and intervention to the z-site (the gate-up product), we observe an intervention-side penalty. The cross-language mean spread at this site is approximately four times smaller than at the a-site. As a result, the mean-patching technique substitutes a proportionally smaller magnitude, causing a uniform drop in intervention effectiveness across all evaluated methods.

Use of AI Assistants

We used an AI assistant (Gemini) to polish the writing of this paper, limited to grammar correction and sentence-level wording suggestions. All suggestions were reviewed and revised by the authors.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS2024-00456709, A Development of Self-Evolving Deepfake Detection Technology to Prevent the Socially Malicious Use of Generative AI), and by the InnoCORE program of the Ministry of Science and ICT(N10260002). This work utilized GPU resources from the "Advanced GPU Utilization Support Program" funded by MSIT, Republic of Korea (awarded to KyungTae Lim).

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

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.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Inaya Rahmanisa, Lyzander Marciano Andrylie, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. Unveiling the influence of amplifying language-specific neurons. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 919–968, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Gábor J. Székely and Maria L. Rizzo. 2013. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. 2026. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34799–34807.

A Implementation Details for the Baselines

Zhang et al. (2026) use a balancing coefficient λ in their score (see Baselines in Section 4.1). It controls the trade-off between the entropy term, which favors neurons concentrated on a few languages, and the max-rate term, which favors neurons with at least one strongly-activating language. Table 6 provides the auto-determined λ_{auto} obtained by following their procedure.

Model	λ_{auto}
SmolLM3-3B-Base	0.297
Llama-3.1-8B	0.341
Qwen3-4B-Base	0.441

Table 6: λ_{auto} for each model.

		$\lvert\mathcal{N}_k\rvert$		zero ΔNLL		mean ΔNLL	
Model	Lang	LAPE	SLN	LAPE	SLN	LAPE	SLN
Llama-3.1-8B	zh	715	27	**0.280**	0.006	**0.332**	0.009
	en	61	188	0.004	**0.011**	0.004	**1.466**
	fr	822	780	0.121	**0.297**	0.141	**1.208**
	id	1057	915	0.239	**0.419**	0.336	**0.605**
	ja	871	247	**1.724**	0.834	**2.169**	1.596
	es	619	616	0.066	**0.165**	0.075	**1.843**
	vi	442	199	0.406	**0.881**	0.918	**1.662**
SmolLM3-3B	zh	657	44	**0.210**	0.012	**0.240**	0.020
	en	45	152	0.002	**0.009**	0.004	**0.506**
	fr	539	475	0.072	**0.136**	0.080	**1.096**
	id	916	566	0.642	**0.884**	0.741	**1.448**
	ja	827	235	**0.981**	0.409	0.928	**2.106**
	es	431	425	0.085	**0.119**	0.094	**1.922**
	vi	547	357	1.907	**2.451**	2.036	**2.671**
Qwen3-4B	zh	530	113	**0.150**	0.106	**0.146**	0.128
	en	90	120	0.004	**0.013**	0.004	**0.564**
	fr	667	651	1.458	**1.527**	**1.948**	1.918
	id	529	690	**1.253**	1.091	**1.566**	1.191
	ja	850	663	**3.185**	2.968	**3.402**	3.235
	es	595	534	**0.821**	0.595	0.997	**1.519**
	vi	241	287	2.790	**3.733**	3.240	**4.368**

Table 7: On-target ΔNLL on WIKI after intervening on language neurons identified by LAPE and SLN under each intervention setting. **Bold** indicates the best result within each intervention setting.

B Comparison under Different Interventions

We employ mean-patching for neuron intervention because it naturally aligns with the properties of our method, as discussed in Section 4.1. Since LAPE and Zhang et al. (2026) were originally evaluated with zero-out interventions, which align with their method designs, we additionally compare LAPE and SLN under zero-out intervention for a fair comparison. The SLN set is identical to that used in the main experiment in Figure 3, where $P = 1\%$. For LAPE, we use the bottom 1% of the entropy as the threshold, following Tang et al. (2024).

Table 7 shows the ΔNLL results under zero-out intervention. For the zero-out intervention, our method identifies more effective language neurons than LAPE in most cases, despite selecting fewer neurons. As discussed in MLNs results in Section 4.2, SLNs are less effective than LAPE under both intervention settings for languages with high coherence to other languages, such as zh and ja. However, as shown in Figures 7, 8, this is because our method separates SLNs and MLNs effectively. For example, zeroing out Chinese-specific neurons identified by LAPE also affects the model's Japanese ability, whereas zeroing out Chinese SLNs identified by our method has little effect on Japanese.

One notable observation is the difference in ΔNLL between the two interventions. For languages except en, both methods show relatively similar ΔNLL. However, for en, SLN yields a much larger increase in ΔNLL under mean-patching than under zero-out intervention, whereas LAPE shows little change. This suggests that English SLNs identified by our method encode language-specific information not only through positive activation but also through distributional structure in the negative activation regime. Mean-patching more directly disrupts this structure by replacing English-specific activations with non-English values, whereas zero-out may be insufficient when the relevant signal lies below zero.

Table 8 locates this difference in the activation distribution itself. For English SLNs, the per-neuron mean activation under target inputs sits within $\sim 0.5\sigma$ of zero, compared with 1.7–2.0σ for every other (model, language) cell. Indeed, 63–72% of SLN neurons have a negative target-language mean. Under this configuration, setting the activation to zero leaves it within its natural target-language manifold and produces no detectable damage on target inputs; setting it to the off-target mean, which is $\sim 1.5\sigma$ away, places it on a language-foreign value and yields the full on-target effect. This is the empirical signature of language-specificity that is encoded in distributional shift rather than in the binary event $a_j(x) > 0$, and it is the structure that the indicator-based identifier family is structurally blind to.

Taken together, the zero-ablation results function as a direct empirical confirmation of the

Model	Lang	μ_{tgt} LAPE	μ_{tgt} SLN	μ_{off} SLN	$/\sigma$ (SLN) $\lvert\mu_{\text{tgt}}\rvert$	$/\sigma$ (SLN) $\lvert\Delta\mu\rvert$	%neg SLN
Llama-3.1-8B	zh	+0.45	+0.94	+0.05	1.76	1.85	4%
	en	+0.44	<u>+0.05</u>	+0.69	<u>0.50</u>	1.61	**70%**
	fr	+0.65	+1.13	−0.03	2.03	2.12	1%
	id	+0.55	+1.05	−0.04	1.94	2.04	0%
	ja	+0.52	+1.07	−0.04	2.00	2.17	2%
	es	+0.62	+1.17	−0.01	1.98	2.05	2%
	vi	+0.52	+1.08	−0.02	1.79	1.91	2%
SmolLM3-3B	zh	+0.37	+0.71	+0.01	1.72	1.98	5%
	en	+0.41	<u>−0.01</u>	+0.52	<u>0.48</u>	1.57	**72%**
	fr	+0.47	+0.83	−0.01	2.03	2.12	1%
	id	+0.29	+0.58	−0.02	1.77	1.89	2%
	ja	+0.48	+0.73	−0.04	1.81	2.06	2%
	es	+0.52	+0.80	+0.01	1.94	2.04	4%
	vi	+0.43	+0.73	+0.00	1.70	1.85	6%
Qwen3-4B	zh	+1.16	+1.59	+0.15	1.52	1.70	9%
	en	+1.00	<u>+0.62</u>	+0.97	<u>0.69</u>	1.38	**63%**
	fr	+2.88	+3.28	−0.04	2.05	2.11	2%
	id	+3.18	+3.35	−0.05	1.98	2.03	3%
	ja	+1.93	+2.36	−0.03	1.99	2.07	8%
	es	+2.88	+3.23	−0.04	2.02	2.09	2%
	vi	+3.12	+2.93	−0.05	1.90	1.96	7%

Table 8: Per-language activation profile of the LAPE and SLN sets (wiki calibration). For each set, μ_{tgt} is the per-neuron mean activation under target-language inputs and μ_{off} is the mean over off-target inputs. The $/\sigma$ columns normalize $|\mu_{\text{tgt}}|$ and the language gap $|\mu_{\text{off}} - \mu_{\text{tgt}}|$ by the neuron's pooled standard deviation; %neg is the fraction of SLN neurons with $\mu_{\text{tgt}} < 0$. For SLN-en (underlined), the target distribution lies within $\sim 0.5\sigma$ of zero with the majority of neurons having a negative target mean, while the language gap remains $\sim 1.5\sigma$. This is precisely the configuration under which zero is approximately no operation on target inputs but the off-target mean is not.

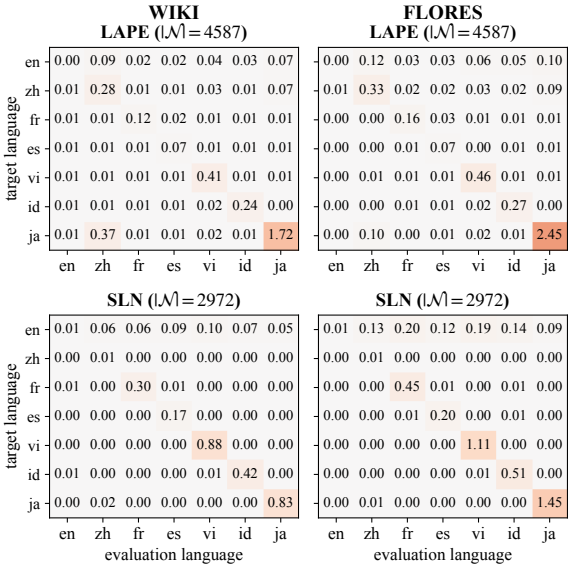


Figure 7: Per-language ΔNLL under zero-ablation on Llama-3.1-8B: LAPE (top) vs SLN (bottom); WIKI (left), FLORES (right).

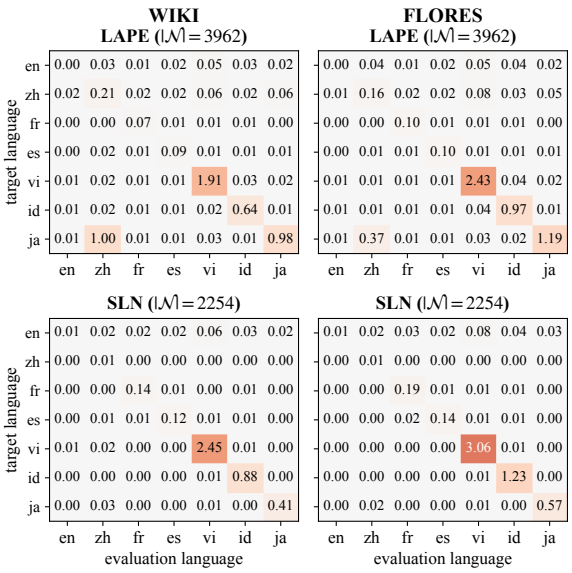


Figure 8: Per-language ΔNLL under zero-ablation on SmolLM3-3B-Base: LAPE (top) vs SLN (bottom); WIKI (left), FLORES (right).

identifier–intervention coupling of Section 4.1: when our SLN set is evaluated with the LAPE-native counterfactual, the activation-rate-encoded subset of language-specificity is preserved, while the distribution-encoded subset becomes invisible. The full advantage of the distributional identifier is realized only under the matched intervention, as a consequence of two distinct kinds of language-specificity coexisting in the model. Figures 7–8 provide the full per-language ΔNLL heatmaps under zero-ablation.

C Site-comparison: Supplementary Results

This collects the empirical material supporting Section 5.3: per-layer sign rates that quantify the structural sign-balance of z (Table 9), and the full selectivity grid across both held-out corpora (Table 10).

Per-layer sign rates. Table 9 lists the mean activation-positive rate $P(\cdot > 0)$ per layer, pooled across the seven calibration languages, for the post-gate signal a and the gate-up product z. At the z-site every layer in both models satisfies $|P(z > 0) - 0.5| \leq 0.002$, so the sign statistic carries no cross-layer or cross-language signal. At the a-site the rate is consistently below 0.5 and varies widely with depth (range $[0.16, 0.48]$ on Llama-3.1-8B, $[0.18, 0.46]$ on SmolLM3-3B-Base), reflecting SiLU's negative-side asymmetry.

Full selectivity grid. Table 10 extends Table 5 with FLORES values for all three (select-site,

Layer	Llama-3.1-8B		SmolLM3-3B-Base	
	$P(a{>}0)$	$P(z{>}0)$	$P(a{>}0)$	$P(z{>}0)$
0	0.379	0.500	0.456	0.500
1	0.366	0.500	0.268	0.500
2	0.343	0.500	0.247	0.500
3	0.290	0.500	0.186	0.500
4	0.214	0.500	0.214	0.501
5	0.197	0.501	0.288	0.500
6	0.192	0.499	0.268	0.500
7	0.210	0.500	0.233	0.500
8	0.211	0.499	0.184	0.501
9	0.211	0.500	0.214	0.500
10	0.211	0.500	0.217	0.501
11	0.212	0.501	0.246	0.501
12	0.230	0.499	0.260	0.499
13	0.208	0.500	0.279	0.499
14	0.196	0.501	0.272	0.500
15	0.168	0.500	0.271	0.499
16	0.169	0.500	0.293	0.500
17	0.164	0.499	0.289	0.502
18	0.176	0.501	0.260	0.500
19	0.184	0.500	0.246	0.501
20	0.192	0.499	0.284	0.501
21	0.192	0.500	0.248	0.501
22	0.204	0.500	0.250	0.499
23	0.227	0.500	0.237	0.500
24	0.249	0.500	0.222	0.500
25	0.262	0.500	0.254	0.499
26	0.280	0.500	0.281	0.500
27	0.292	0.500	0.330	0.500
28	0.308	0.500	0.333	0.500
29	0.341	0.500	0.345	0.500
30	0.351	0.499	0.386	0.500
31	0.479	0.499	0.401	0.500
32	-	-	0.403	0.500
33	-	-	0.387	0.500
34	-	-	0.389	0.499
35	-	-	0.451	0.499
mean	0.247	0.500	0.289	0.500
min	0.163	0.499	0.184	0.499
max	0.479	0.501	0.456	0.502

Table 9: Mean activation-positive rate $P(\cdot > 0)$ per layer, pooled across the seven calibration languages, on 50,000 tokens per language sampled from `WIKI-20231101`. The z-site rate stays within 0.002 of 0.5 across all layers in both models, while the a-site rate spans an order of magnitude in dynamic range.

patch-site) configurations and all three identifiers, on both held-out corpora. Moving the selection from a to z, indicator baselines lose 74–94% of their selectivity across (model, corpus) combinations (LAPE 74–86%; LSN 89–94%), while our identifier retains 50–71%. At (z, a) our identifier is 25–55× stronger than LAPE and 15–29× stronger than LSN across all (model, corpus) combinations, of the same order as the main-result a-site gap. The drop from (z, a) to (z, z) within each method follows the intervention-side mechanical penalty discussed in Section 5.3 and applies uniformly across identifiers; it does not change the ordering established by the (z, a) comparison.

D Further Analysis of DLN

D.1 Detailed Comparison with LAPE and SLN

Figures 9 and 10 present heatmap results that decompose the three subsets (LAPE $\setminus$ SLN, LAPE $\cap$ SLN, SLN $\setminus$ LAPE), which were summarized by mean selectivity in Table 3, into 7×7 cells of (target language, evaluation language).

Three patterns are consistently observed across the two models. First, LAPE $\cap$ SLN (middle) exhibits the non-zero diagonal cells and near-zero off-diagonal cells, indicating that the core neurons jointly recognized by both identifiers have the most concentrated on-target effect. Second, al-

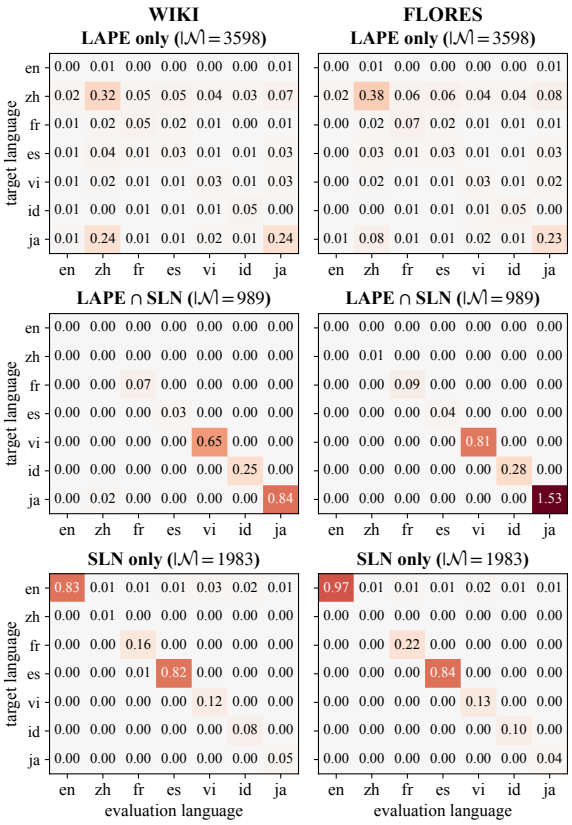


Figure 9: ΔNLL by language on Llama-3.1-8B for LAPE $\setminus$ SLN (top) and LAPE $\cap$ SLN (middle), and SLN $\setminus$ LAPE single-language neuron at $P = 1\%$ (bottom) on WIKI and FLORES.

Model	Corpus	(selection-site, patch-site)	LAPE	LSN	SLN
Llama-3.1-8B	WIKI	(a-site, a-site)	22.8	126.0	**273.7**
		(z-site, a-site)	4.5	13.3	**193.9**
		(z-site, z-site)	3.9	9.9	**97.2**
	FLORES	(a-site, a-site)	40.5	307.0	**641.2**
		(z-site, a-site)	5.8	17.5	**318.4**
		(z-site, z-site)	4.9	12.6	**176.1**
SmolLM3-3B-Base	WIKI	(a-site, a-site)	28.2	89.5	**296.5**
		(z-site, a-site)	7.4	8.8	**187.3**
		(z-site, z-site)	5.3	9.2	**57.4**
	FLORES	(a-site, a-site)	43.3	189.7	**632.1**
		(z-site, a-site)	9.8	12.9	**372.3**
		(z-site, z-site)	7.0	13.3	**87.6**

Table 10: Full selectivity grid for the three identifiers across (select-site, patch-site) configurations and both held-out corpora. (a, a) is the main-result protocol; (z, a) varies only the selection site (the controlled comparison); (z, z) varies both. Each method is evaluated at its native $P{=}1\%$ operating point. Bold entries are highest within each row.

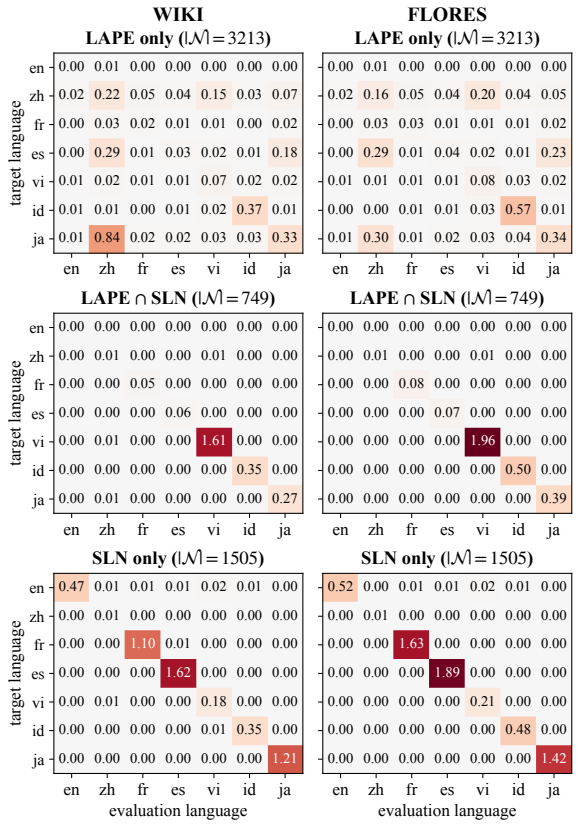


Figure 10: ΔNLL by language on SmolLM3-3B-Base for LAPE \ SLN (top) and LAPE ∩ SLN (middle), and SLN \ LAPE single-language neuron at $P = 1\%$ (bottom) on WIKI and FLORES.

though the neurons in SLN \ LAPE (bottom) lie outside the LAPE pool, the magnitude of their diagonal cells exceeds that of the diagonal cells in LAPE \ SLN (top) in nearly all languages. The gap is most pronounced in the English (en) cell, which is consistent with the distributional observation in Section 5.1 and Appendix B that the target-off separation of English SLNs occurs in the negative region. Third, LAPE \ SLN (top) contains the largest number of neurons among the three subsets, yet exhibits the weakest diagonal cells and a relatively broader off-diagonal spread, showing that the peripheral neurons of the LAPE pool, which are selected solely on the basis of activation-rate statistics, possess on average both a weaker signal and a more diffuse off-target influence.

D.2 Impact of MLNs on Language-specific Performance

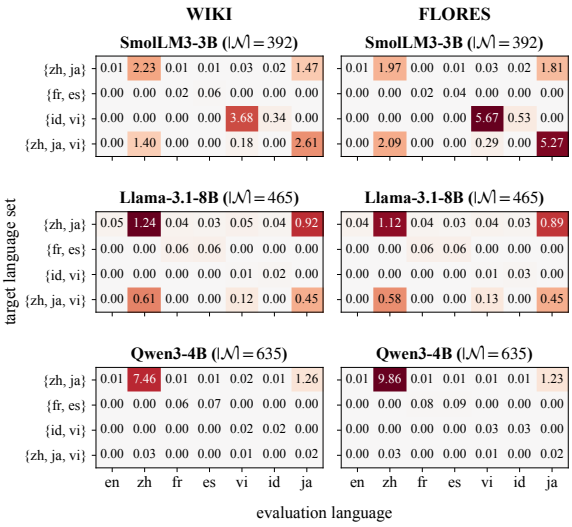


Figure 11: Comparison of ΔNLL scores for the four most frequently occurring MLN clusters across Llama-3.1-8B, SmolLM3-3B-Base, and Qwen3-4B-Base. The results are evaluated on WIKI and FLORES.

Figure 11 presents the ΔNLL heatmaps for four of the most frequent MLN clusters across the three

models. The performance of the languages included in these MLN clusters generally deteriorated in all three models.

Furthermore, a larger performance drop tends to occur when an MLN cluster accounts for a higher proportion of the neurons associated with a specific language. For instance, the {zh, ja} cluster accounts for a substantial proportion of the neurons for both zh and ja, which leads to significant performance degradation. Conversely, for the {fr, es} and {id, vi} clusters, the performance degradation is marginal because the proportion of SLNs for each language remains higher than that of the corresponding MLNs.

However, the SmolLM3-3B-Base model presents a notable exception. Despite accounting for a relatively low proportion of the language-specific neurons, the {id, vi} MLN cluster still caused a significant performance degradation. This phenomenon can be attributed to the fact that these languages are not among the model's officially supported languages (English, French, Spanish, German, Italian, and Portuguese), resulting in relatively unstable signals compared to the target languages. In other words, for unsupported languages, crucial information may heavily rely on a small subset of neurons, suggesting that MLNs likely encompass the neurons essential for processing these languages. Consequently, this demonstrates that the nature and importance of the information encoded within MLNs can vary depending on the model and its training data. Therefore, to fully comprehend language-specific mechanisms, an in-depth analysis of MLNs, alongside SLNs, is essential.

We scaled up the above experiment to a larger model, revealing a consistent trend. Figure 13 presents the ΔNLL heatmap for the four most frequently observed MLN clusters in Llama-3.1-70B. For the {zh, ja} cluster, which accounts for a high proportion of MLNs in its respective languages, a substantial performance drop is observed. Conversely, the {fr, es} cluster, characterized by a lower MLN and higher SLN proportion, exhibits a relatively minor degradation. A similar pattern is evident in the {en, fr, es} cluster, where the MLN share within each language is too minimal to cause performance drop. These findings demonstrate that our proposed methodology and the observed MLN dynamics are not constrained by model scale, highlighting their broad applicability.

D.3 Scaling to Llama-3.1-70B

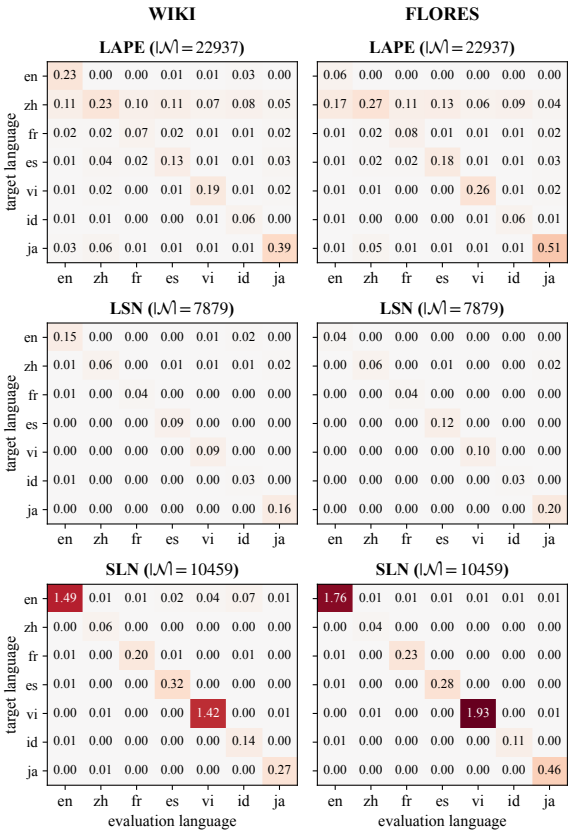


Figure 12: ΔNLL by language on Llama-3.1-70B for LAPE, LSN, and SLN at $P = 1\%$ on WIKI and FLORES.

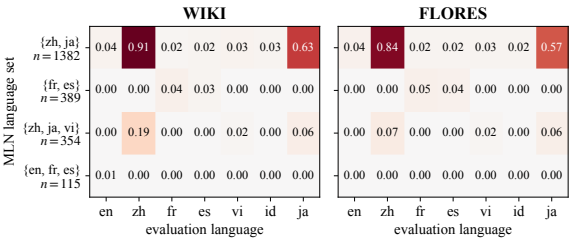


Figure 13: Comparison of ΔNLL scores for the four most frequently occurring MLN clusters of Llama-3.1-70B. The results are evaluated on WIKI and FLORES.

Figure 12 reports the SLN comparison on Llama-3.1-70B at $P{=}1\%$. The findings from the smaller models carry over. Our SLN concentrates far larger on-target damage than LAPE and LSN while using less than half the neurons of LAPE, and its off-target damage stays near zero. English, where both baselines remain weak (0.23 and 0.15 on WIKI), is again recovered only by SLN (1.49 on WIKI, 1.76 on FLORES). The one weak diagonal, Chinese, mirrors the 8B models: its signal is organized in the {zh, ja} MLN cluster (Figure 13).

lang	WIKI			FLORES		
	LAPE	LSN	SLN	LAPE	LSN	SLN
en	0.004	0.004	**1.835**	0.006	0.005	**2.569**
zh	0.166	0.057	**0.248**	0.129	0.055	**0.207**
fr	0.191	0.127	**1.986**	0.297	0.185	**3.204**
es	0.121	0.092	**1.736**	0.143	0.114	**2.169**
vi	0.623	0.385	**1.213**	0.802	0.451	**1.583**
id	0.481	0.240	**1.472**	0.727	0.331	**2.356**
ja	0.351	0.173	**2.528**	0.382	0.201	**3.394**
mean	0.277	0.154	**1.574**	0.355	0.192	**2.212**

Table 11: Per-language on-target ΔNLL on Aya-23-8B at P=1%. **Bold** denotes the largest damage per (language, corpus).

D.4 Generalization to Aya-23-8B

We run the full pipeline on Aya-23-8B (Aryabumi et al., 2024), an instruction-tuned model trained explicitly for multilinguality, under the identical protocol (P=1%; LAPE 4,587 / LSN 1,618 / SLN 3,800 neurons). Table 11 reports per-language on-target ΔNLL. All findings replicate: our SLN produces the largest on-target damage on all seven languages and both corpora, while its mean off-target damage remains the lowest (0.006 vs. LAPE's 0.012 on WIKI), yielding mean selectivity 254 against 133 (LSN) and 24 (LAPE) on WIKI, and 691 against 350 and 38 on FLORES.

The multi-language structure is likewise stable. The {zh, ja} coalition is again the largest MLN cluster (355 neurons), followed by {fr, es} and {id, vi}, and Chinese is the one language whose SLN diagonal is weak (0.25); the union pool recovers it (2.36 on WIKI), with the extra damage concentrated on its cluster-mate Japanese, exactly as the framework predicts. Since no base checkpoint of Aya-23 is released, these results show that the method transfers to post-trained models.

D.5 Generation Validation via Cross-lingual Steering

To analyze the causal impact of the identified language-specific neurons on text generation, we employ activation steering during inference. We apply this steering to the neurons extracted by each respective methodology, utilizing a paired intervention technique that simultaneously manipulates two distinct sets of neurons: source SLNs and target SLNs. For the source SLNs, we suppress their language-specific signals by applying mean-patching (Section 4.1), which replaces their activations with a language-neutral baseline. However, to ensure a fair comparison with LAPE, zero-patching, which deactivates the neurons by setting their values to zero, is applied to LAPE instead of mean-patching. Concurrently, the target language neurons are activated by injecting their pre-computed average activation values specific to the target language. When applying this technique to MLNs, any overlap between source and target neurons is resolved by prioritizing the target neuron activation. Ultimately, this dual intervention effectively spoofs the model's internal state, mimicking the natural activation patterns in the target language.

lang	LAPE		SLN		SLN+MLN	
	ACC	p_{tgt}	ACC	p_{tgt}	ACC	p_{tgt}
SmolLM3-3B						
zh	**1.000**	**0.963**	**1.000**	0.935	**1.000**	0.951
ja	**1.000**	**0.955**	0.970	0.895	0.980	0.867
vi	**1.000**	**0.965**	**1.000**	0.962	**1.000**	0.950
id	**1.000**	0.946	**1.000**	0.955	**1.000**	**0.958**
es	**1.000**	**0.948**	**1.000**	0.922	**1.000**	0.934
fr	**1.000**	**0.964**	**1.000**	0.953	**1.000**	0.956
avg	**1.000**	**0.957**	0.995	0.937	0.997	0.936
Llama-3.1-8B						
zh	**1.000**	**0.958**	0.990	0.920	**1.000**	0.874
ja	**1.000**	**0.925**	**1.000**	0.865	**1.000**	0.899
vi	**1.000**	**0.970**	**1.000**	0.968	**1.000**	0.966
id	**1.000**	0.964	**1.000**	0.964	**1.000**	**0.965**
es	**1.000**	0.961	**1.000**	0.962	**1.000**	**0.964**
fr	0.990	0.948	**1.000**	0.962	**1.000**	**0.966**
avg	0.998	**0.954**	0.998	0.940	**1.000**	0.939
Qwen3-4B						
zh	0.968	0.869	0.990	0.918	**1.000**	**0.944**
ja	0.959	0.886	0.934	0.800	**1.000**	**0.937**
vi	0.940	0.870	**1.000**	0.955	**1.000**	**0.968**
id	**1.000**	0.942	**1.000**	**0.974**	**1.000**	0.952
es	0.970	0.941	**1.000**	**0.965**	0.990	0.961
fr	0.990	0.953	**1.000**	0.963	**1.000**	**0.972**
avg	0.971	0.910	0.987	0.929	**0.998**	**0.956**

Table 12: Performance comparison of LAPE, SLN, and SLN+MLN on the X-to-English (X2en) transfer task. The table reports accuracy and target language probability (p_{tgt}) across three models. Zero-patching is applied to LAPE, whereas mean-patching is utilized for our proposed SLN and SLN+MLN interventions. **Bold** indicates the highest accuracy.

For the experimental setup, we extract 200 samples per language from the FLORES dataset. To ensure rigorous language verification, we utilize the fastText language identification model (Joulin et al., 2016), selecting only those samples where the predicted probability for the target language is 0.98 or higher.

Table 12 presents the results of transferring text

lang	LAPE ACC	LAPE p_{tgt}	SLN ACC	SLN p_{tgt}	SLN+MLN ACC	SLN+MLN p_{tgt}
SmolLM3-3B						
en	0.000	0.004	0.000	0.001	**0.845**	**0.802**
ja	0.236	0.211	**0.990**	**0.941**	**0.990**	0.922
vi	0.970	0.941	0.178	0.136	**1.000**	**0.968**
id	0.000	0.000	0.000	0.000	**1.000**	**0.985**
es	**0.969**	**0.937**	0.000	0.000	0.929	0.904
fr	0.863	0.779	0.000	0.000	**0.917**	**0.861**
avg	0.506	0.479	0.195	0.180	**0.947**	**0.907**
Llama-3.1-8B						
en	0.000	0.000	0.000	0.000	**0.891**	**0.791**
ja	0.768	0.694	0.811	0.755	**0.955**	**0.795**
vi	**0.910**	**0.764**	0.011	0.014	0.880	0.693
id	**0.875**	**0.806**	0.000	0.000	0.375	0.417
es	0.705	0.599	0.000	0.000	**0.863**	**0.801**
fr	**0.652**	0.535	0.000	0.000	0.606	**0.582**
avg	0.652	0.567	0.137	0.128	**0.762**	**0.680**
Qwen3-4B						
en	0.900	0.845	**1.000**	0.980	**1.000**	**0.987**
ja	0.979	**0.906**	**1.000**	0.850	0.989	0.848
vi	**1.000**	0.940	0.920	0.910	**1.000**	**0.993**
id	**1.000**	**0.995**	0.500	0.521	**1.000**	0.987
es	0.979	0.935	0.660	0.653	**1.000**	**0.977**
fr	0.907	0.868	0.320	0.319	**1.000**	**0.979**
avg	0.961	0.915	0.733	0.705	**0.998**	**0.962**

Table 13: Performance comparison of LAPE, SLN, and SLN+MLN on the X-to-Chinese (X2zh) transfer task. The table reports accuracy and target language probability (p_{tgt}) across three models. Zero-patching is applied to LAPE, whereas mean-patching is utilized for our proposed SLN and SLN+MLN interventions. **Bold** indicates the highest accuracy.

lang	LAPE ACC	LAPE p_{tgt}	SLN ACC	SLN p_{tgt}	SLN+MLN ACC	SLN+MLN p_{tgt}
SmolLM3-3B						
zh	0.000	0.004	0.000	0.001	**0.845**	**0.802**
ja	0.000	0.001	0.000	0.001	**0.854**	**0.814**
vi	0.000	0.000	0.010	0.011	**0.414**	**0.406**
id	0.000	0.000	0.051	0.038	**0.131**	**0.096**
es	0.000	0.001	**0.717**	**0.671**	0.653	0.615
fr	0.000	0.001	**0.788**	**0.765**	0.697	0.674
avg	0.000	0.001	0.261	0.248	**0.599**	**0.568**
Llama-3.1-8B						
zh	0.000	0.000	0.000	0.000	**0.891**	**0.791**
ja	0.020	0.023	0.060	0.061	**0.956**	**0.860**
vi	0.000	0.000	0.927	0.889	**0.947**	**0.892**
id	0.000	0.000	**0.735**	**0.610**	0.714	0.543
es	0.000	0.000	**0.840**	**0.803**	0.780	0.734
fr	0.010	0.010	**0.970**	**0.941**	0.900	0.871
avg	0.005	0.006	0.589	0.551	**0.865**	**0.782**
Qwen3-4B						
zh	0.900	0.845	**1.000**	0.980	**1.000**	**0.987**
ja	0.990	0.970	**1.000**	**0.999**	**1.000**	0.996
vi	0.040	0.046	0.660	0.657	**0.990**	**0.990**
id	0.010	0.004	0.450	0.364	**0.800**	**0.669**
es	0.010	0.015	0.420	0.390	**0.930**	**0.866**
fr	0.250	0.222	0.520	0.498	**0.990**	**0.960**
avg	0.367	0.350	0.675	0.648	**0.952**	**0.911**

Table 14: Performance comparison of LAPE, SLN, and SLN+MLN on the English-to-X (en2X) transfer task. The table reports accuracy and target language probability (p_{tgt}) across three models. Zero-patching is applied to LAPE, whereas mean-patching is utilized for our proposed SLN and SLN+MLN interventions. **Bold** indicates the highest accuracy.

from a non-English source language X to English as the target language. The performance discrepancies are marginal, with an average accuracy difference of 0.027 and a maximum per-language difference of 0.066. This robust performance stems from the models being predominantly pre-trained on English. The models frequently default to generating English responses even when provided with non-English inputs, naturally leading to consistently high baseline scores.

To evaluate cross-lingual transfer capabilities toward a non-primary language, we conduct experiments transferring from a source language X to Chinese (`zh`) as the target language. As demonstrated in Table 13, DLN consistently outperforms LAPE across all scenarios. Notably, a substantial performance gap is observed between SLNs and MLNs. As analyzed in Appendix I.1 and D.2, this gap arises because Chinese-specific textual features are predominantly processed within MLNs.

Table 14 demonstrates the performance when transferring from English as the source language to all other target languages. In stark contrast to the X-to-English transfer task, the performance of LAPE approaches zero. Conversely, DLN maintains solid performance, with accuracy ranging from 0.599 to 0.952 depending on the architecture. This phenomenon is primarily driven by the negative regions analyzed in Section 5.1. However, our method exhibits lower performance for Vietnamese (`vi`) and Indonesian (`id`) within the SmolLM3-3B architecture, which stems from unstable internal signals provided by the model, as discussed in Appendix D.2. This explanation is supported by the results of Llama-3.1-8B, which underwent extensive multilingual pre-training and consistently delivers stable, high-tier performance.

		ours			LAPE
P (%)	$\lvert\mathcal{N}\rvert$	on-tgt	off-tgt	Sel	Sel
1.5	3,885	0.403	0.003	145.8	10.3
3.0	7,336	0.322	0.004	88.7	9.8
5.0	12,444	0.294	0.004	69.4	8.1

Table 15: Extended percentile sweep of the DLN pool on SmolLM3-3B-Base (WIKI). LAPE is matched to our neuron budget $|\mathcal{N}|$ at every P.

D.6 Extended Percentile Sweep

Figure 5 sweeps the percentile threshold over $P \in [0.2, 1.0]$. Here we extend the sweep to $P{=}5\%$ on SmolLM3-3B-Base (WIKI) under the identical protocol, evaluating the DLN pool with the LAPE budget matched to ours at every P. As Table 15 shows, our selectivity remains an order of magnitude above LAPE (9–14$\times$) at every extended operating point.

The mild decline of the raw on-target damage ($0.40 \rightarrow 0.29$) is a composition effect rather than a weakening of the method: higher P admits additional three-language MLN clusters with small per-cluster damage ($13 \rightarrow 33$ clusters), which dilutes the unweighted cluster mean in the selectivity numerator. Within the SLN stratum, whose cluster count is fixed at seven, on-target damage instead rises monotonically ($1.47 \rightarrow 1.64$) while off-target damage stays below 0.02 nats. The percentile choice is therefore not load-bearing anywhere in the extended range.

D.7 Histogram Bin-count Ablation

#bins	τ	$\lvert\mathcal{N}_{\text{SLN}}\rvert$	Jaccard	time (s)
10	0.4231	3,116	0.919	0.8
50	0.4015	2,983	0.996	1.7
100	0.4006	2,975	0.998	3.4
200	0.4003	2,974	0.999	6.1
400	0.4001	2,972	1.000	11.6
800	0.3998	2,973	0.999	23.5

Table 16: Bin-count ablation on Llama-3.1-8B ($P{=}1\%$). Jaccard is measured against the paper's 400-bin selection; time is the CPU cost of the overlap computation and selection. Target-language assignments agree at 100% for every bin count.

Our identifier approximates each per-language density with a 400-bin adaptive histogram (Section 4.1). To probe this choice, we rebuild the histogram cache at 800 bins on Llama-3.1-8B and derive every coarser resolution by aggregating adjacent bins, which is exact because the per-neuron bin edges nest; each resolution is then passed through the identical selection at $P{=}1\%$. As Table 16 shows, the selection is essentially invariant from 50 to 800 bins, a 16$\times$ range. The threshold τ moves by less than 0.002, the SLN pool stays within Jaccard 0.996 of the paper's, and every selected neuron keeps the same target-language assignment at every resolution, including 10 bins. Only at 10 bins does the pool itself drift (Jaccard 0.92), as coarse bins begin to blur the near-zero negative-region structure that distinguishes English neurons. Since the bin count affects only the post-forward overlap computation, which takes seconds on CPU, the resolution can be reduced substantially at no cost to the selection.

E Robustness of the Identifier Design

Variant	$\lvert\mathcal{N}\rvert$	Jac.	ΔNLL (W / F)	Sel. (W / F)
Linkage (overlap statistic)				
overlap, single (ours)	2,972	1.00	1.20 / 1.69	198 / 408
overlap, average	2,939	0.99	1.14 / 1.60	198 / 403
overlap, complete	2,848	0.96	0.97 / 1.37	179 / 374
Bounded statistic (single linkage)				
Jensen–Shannon	2,999	0.95	1.27 / 1.83	202 / 424
squared Hellinger	3,066	0.92	1.29 / 1.90	207 / 445
Kolmogorov–Smirnov	2,970	0.99	1.17 / 1.64	201 / 400
Scale-bearing statistic (single linkage)				
Wasserstein-1	5,603	0.47	0.96 / 1.31	81 / 152
energy distance	4,797	0.58	0.99 / 1.31	120 / 234
W1, top-$\lvert\mathcal{N}\rvert$	2,972	—	0.62 / 0.85	117 / 212
energy, top-$\lvert\mathcal{N}\rvert$	2,972	—	0.77 / 1.05	156 / 302

Table 17: Effect of the pairwise statistic and linkage rule on the selected pool (Llama-3.1-8B, $P = 1\%$). Jac. is the Jaccard overlap with our overlap+single set; ΔNLL is the on-target damage and Sel. the selectivity, each on WIKI (W) and FLORES (F). The last two rows truncate the scale-bearing pools to our pool size by top-ranked score (neuron-matched control).

We examine three design choices of our identifier on Llama-3.1-8B at $P{=}1\%$. These are the pairwise overlap coefficient $\mathbf{S}_j$ (Section 3.2), the single-linkage bipartitioning, and the fixed cluster count $K{=}2$. For a general pairwise distance, we take the most-separated bipartition and threshold at the top-P percentile. Table 17 reports the pool size $|\mathcal{N}|$, its Jaccard overlap with our set, and the on-target ΔNLL and selectivity on WIKI / FLORES.

Bounded statistics and linkage agree. Every bounded, shape-sensitive statistic selects essentially the same neuron set. Jensen–Shannon, squared Hellinger, and Kolmogorov–Smirnov stay within 3% of our pool size, with Jaccard 0.92–0.99

and selectivity within $\sim 10\%$ on both corpora. They also preserve the multi-language structure, with Jensen–Shannon recovering 650 MLNs against our 654 under the same leading coalitions. The linkage rule is equally immaterial (Jaccard 0.96–0.99 across single/average/complete), and we keep single linkage because it is exactly the stated max-overlap criterion.

Scale-bearing metrics reward magnitude, not discriminability. Wasserstein-1 and the energy distance instead inflate the pool (1.6–1.9×), fall to Jaccard 0.47–0.58, roughly halve selectivity, and distort the MLN structure. This is not merely over-selection but a property of the ranking itself. Truncating each pool to our size by top-ranked score makes them worse rather than better (0.62 vs. 0.96 on-target ΔNLL for W1 on WIKI). Their ranking places the largest-variance neurons first.

Why the overlap coefficient. Among these interchangeable statistics, we adopt the overlap coefficient for interpretability and cost. It equals one minus the total variation distance, $\mathbf{S}_j[k, k'] = 1 - \mathrm{TV}(p_j^k, p_j^{k'})$, so τ reads directly as a bound on shared probability mass, and it is the cheapest to compute over all 458k neurons (10 s on CPU, against 96 s for Jensen–Shannon).

The $K{=}2$ bipartition is a selection decision, not a bimodality assumption. For every neuron we recompute the selection statistic inside the larger reference cluster A_j^c of its most-separated bipartition, so a genuinely three-group neuron clears τ on this residual split as well. The empirical structure is overwhelmingly two-group. The median residual-split overlap is 0.78 among selected neurons and 0.92 among rejected ones, nearly twice the threshold $\tau{=}0.40$, and only 50 of the 3,626 selected neurons (1.4%) admit a threshold-clearing residual split, with linguistically coherent nestings such as {zh, ja} then {en}. Moreover, none of the 455,126 rejected neurons admits a three-group structure, which follows from the criterion itself, since three mutually separated groups already provide a bipartition whose cross-cluster overlap clears τ. Such neurons are selected through their most-separated split and labeled by its minority side. A natural refinement is to choose the number of clusters by within-cluster cohesion at additional computational cost, which we leave to future work.

		on-target $\Delta\mathrm{NLL}_k$			
k	d (σ)	SLN	Rand	Rand $\times\frac{1}{2}$	Ratio
en	1.61	1.466	0.010	0.005	148×
zh	1.85	0.009	0.001	0.000	15×
fr	2.12	1.208	0.021	0.017	58×
es	2.05	1.843	0.014	0.006	131×
vi	1.91	1.662	0.007	0.006	242×
id	2.04	0.605	0.028	0.011	21×
ja	2.17	1.596	0.010	0.001	155×
mean	—	1.198	0.013	0.006	92×

Table 18: Displacement-matched random control on Llama-3.1-8B (WIKI, $P{=}1\%$). d is the normalized displacement of the matched SLN set; Rand patches random layer- and count-matched neurons by the same d, Rand $\times\frac{1}{2}$ by half. Ratio is SLN over Rand on-target damage. The mean ratio is the ratio of the mean damages.

F Displacement-matched Random Control

For each SLN identified for target language k, we measure its normalized displacement under mean-patching, i.e., the gap between its mean activation on the target language and its mean over the non-target languages, in units of its pooled standard deviation. We then draw random neurons outside every selected pool and the LAPE pool, matched to the SLN set in layer and count, and shift each away from its own target-language mean by the same normalized amount and direction (3 seeds; we report the mean). A *half-displacement* condition repeats this at half the magnitude as a dose–response probe. Table 18 reports per-language results on Llama-3.1-8B (WIKI, $P{=}1\%$).

Three observations follow. (i) The same displacement applied at random locations yields $\sim$1% of the SLN effect, so displacement magnitude does not mechanically produce the reported damage. (ii) The control has no language selectivity: its on-target and off-target damage are identical (0.013 vs. 0.013), i.e., random displacement acts as diffuse noise, while the SLN intervention concentrates its effect on the target language with off-target damage (0.006) below the control's noise floor. (iii) Halving the displacement leaves the control equally inert (0.006), ruling out a dose–response account. The smallest ratio, Chinese, is small for both interventions (only 27 Chinese SLNs at $P{=}1\%$); as in Section 4.2, the Chinese signal resides in the {zh, ja} MLN cluster. All three observations replicate on SmolLM3-3B-Base: control 0.010 vs. SLN

1.395 on-target, per-language ratios 23–378×, and no dose response (0.005 at half displacement).

G Downstream Evaluation

G.1 Experimental Settings

Belebele. Each question is scored zero-shot. We build one prompt per answer option,

```
{passage}
Q: {question}
A: {option}
```

and take the summed log-probability of the `{option}` tokens as its score, predicting the highest-scoring option.

MGSM. Each problem is prompted with the 8 chain-of-thought exemplars from the dataset's own training split for that language, so the exemplars, their reasoning, and the question prefixes are all in the evaluation language. The prompt is

```
{q1}
{a1}
⋮
{q8}
{a8}
{test question}
{answer prefix}
```

where each question sits on the line above its answer and the blocks are separated by a blank line, so the trailing answer prefix leaves the model to produce the reasoning. Table 19 lists the prefixes per language. We decode greedily for at most 512 new tokens, stopping at a blank line or at a new question prefix, and take the last number in the generation as the predicted answer, matched exactly against the gold value.

	Question prefix	Answer prefix
en	Question:	Step-by-Step Answer:
zh	问题:	逐步解答:
fr	Question :	Réponse étape par étape :
es	Pregunta:	Respuesta paso a paso:
ja	問題:	ステップごとの答え:

Table 19: MGSM prompt prefixes for the five languages MGSM covers.

G.2 Per-language Downstream Results

Table 20 breaks Table 2 down by target language. English separates the methods most sharply. Both baselines assign English a small subset, 61 and 28 neurons on Llama-3.1-8B and 45 and 16 on SmolLM3-3B-Base, and patching it costs at most 0.33 points on Belebele and 3.60 on MGSM, whereas our English SLNs remove 6.56 and 12.00 points on Llama-3.1-8B and 3.22 and 10.00 on SmolLM3-3B-Base. This reproduces on task accuracy the English gap reported for ΔNLL in Section 4.2, and it is the downstream counterpart of the negative-region analysis in Section 5.1: 63–72% of the English SLNs patched here have a negative target-language mean and therefore lie in the region that positive-rate metrics discard by construction.

	Method	en	zh	fr	es	vi	id	ja	Avg
Llama-3.1-8B									
$\lvert\mathcal{N}_{\{k\}}\rvert$	LAPE	61	715	822	619	442	1057	871	–
	LSN	28	107	481	318	105	540	235	–
	SLN	188	27	780	616	199	915	247	–
Belebele	LAPE	0.33	**3.33**	0.67	1.44	4.56	**2.67**	6.22	2.75
	LSN	0.00	1.56	0.56	0.56	3.44	0.89	1.56	1.22
	SLN	**6.56**	0.11	**5.00**	**7.11**	**6.78**	2.44	**6.56**	**4.94**
MGSM	LAPE	0.40	**8.00**	0.80	2.00	–	–	**23.60**	6.96
	LSN	−0.80	3.20	0.80	4.80	–	–	14.00	4.40
	SLN	**12.00**	−0.40	**11.60**	**8.80**	–	–	15.20	**9.44**
SmolLM3-3B-Base									
$\lvert\mathcal{N}_{\{k\}}\rvert$	LAPE	45	657	539	431	547	916	827	–
	LSN	16	119	296	207	201	375	229	–
	SLN	152	44	475	425	357	566	235	–
Belebele	LAPE	0.11	**1.78**	1.56	1.00	6.89	2.00	3.89	2.46
	LSN	0.00	−0.33	0.78	0.56	5.44	1.33	1.00	1.25
	SLN	**3.22**	−0.22	**7.89**	**9.00**	**8.33**	**4.78**	**5.89**	**5.56**
MGSM	LAPE	3.60	0.00	0.00	1.20	–	–	13.20	3.60
	LSN	0.40	**1.60**	−0.40	3.60	–	–	−1.60	0.72
	SLN	**10.00**	0.40	**5.20**	**9.20**	–	–	**14.00**	**7.76**

Table 20: Per-language on-target accuracy drop under the mean-patch intervention, expanding Table 2. The $\lvert\mathcal{N}_{\{k\}}\rvert$ block gives the number of neurons each identifier assigns to target language k; these pools are shared between the two tasks, so the MGSM totals in Table 2 sum only over the five languages MGSM covers. **Bold** denotes the largest drop per column.

Chinese is the one target where our SLNs fall below LAPE, at 0.11 versus 3.33 on Belebele and −0.40 versus 8.00 on MGSM for Llama-3.1-8B. This mirrors the ΔNLL exception in Section 4.2 and follows from the same property of our identifier: almost all Chinese language neurons are assigned to the {zh, ja} multi-language cluster, leaving 27 neurons on Llama-3.1-8B and 44 on SmolLM3-3B-Base as single-language, so MLNs account for 93.3% and 88.1% of the Chinese language neurons (Table 22). The Chinese SLN column therefore patches these 27 and 44 neurons, not the pool that carries the Chinese effect; that pool is the MLN cluster, where mean-patching $\mathcal{N}_{\{\texttt{zh},\texttt{ja}\}}$ yields $\Delta\mathrm{NLL}_{\texttt{zh}} = 1.24$ on WIKI and 1.12 on FLORES.

	Method	on-target ($k'=k$)			off-target ($k'\neq k$)		
		b	c	p	b	c	p
Llama-3.1-8B							
Belebele	LAPE	392	219	2×10^{-12}	409	381	0.34
	LSN	282	205	6×10^{-4}	109	118	0.60
	SLN	612	301	4×10^{-25}	153	161	0.69
MGSM	LAPE	162	75	2×10^{-8}	297	253	0.07
	LSN	113	58	3×10^{-5}	90	75	0.28
	SLN	212	94	1×10^{-11}	124	115	0.61
SmolLM3-3B-Base							
Belebele	LAPE	402	247	1×10^{-9}	510	458	0.10
	LSN	282	203	4×10^{-4}	113	110	0.89
	SLN	685	335	3×10^{-28}	228	173	<u>0.007</u>
MGSM	LAPE	127	82	2×10^{-3}	263	260	0.93
	LSN	79	70	<u>0.51</u>	70	102	<u>0.02</u>
	SLN	176	79	1×10^{-9}	111	131	0.22

Table 21: Two-sided exact McNemar tests on the item-level outcomes behind Table 2, pooled over target languages. b counts items answered correctly by the unintervened model and incorrectly after the intervention, and c counts the reverse, so $b > c$ indicates a drop. <u>Underline</u> marks the three entries discussed in the text.

Outside Chinese, a baseline exceeds our raw on-target drop in 2 of the 20 remaining (model, task, language) cells: Indonesian on Belebele with Llama-3.1-8B (2.67 versus 2.44) and Japanese on MGSM with Llama-3.1-8B (23.60 versus 15.20). In both, LAPE patches a larger pool, 1057 against 915 neurons and 871 against 247, and our per-neuron drop is still larger, by $1.06\times$ and $2.27\times$. The averages in Table 2 therefore do not rest on a few target languages.

Because the accuracy differences in Table 2 are small in absolute terms, we also test them at the item level. Table 21 reports two-sided exact McNemar tests on the discordant pairs between the unintervened and intervened runs, pooled over target languages. Every on-target drop reaches $p < 0.01$ except LSN on MGSM with SmolLM3-3B-Base ($p = 0.51$), whose 0.72-point entry in Table 2 is therefore not distinguishable from no effect. Our SLNs give the smallest p-values in all four settings, at $p \leq 4\times10^{-25}$ on Belebele and $p \leq 1\times10^{-9}$ on MGSM.

Ten of the twelve off-target entries are not significant at the 0.05 level. The two exceptions are our SLNs on Belebele with SmolLM3-3B-Base ($p = 0.007$), whose magnitude is 0.15 accuracy points, and LSN on MGSM with SmolLM3-3B-Base ($p = 0.02$ with $c > b$, that is, off-target accuracy rises). Both are likely overstated: unlike the on-target test, where each target language contributes a disjoint item set, the off-target test reuses the same evaluation items under the six or four other target-language interventions, so the pooled discordant pairs are not independent and the test is anti-conservative.

H Concrete Examples of Activation Distribution

In this section, we illustrate the four representative activation patterns presented in Figure 6. Figures 14–17 display four SLN samples corresponding to each case as violin plots across the seven calibration languages, with each inset indicating the neuron's location (L, N), the target language, and whether it is identified as SLN and/or LAPE.

Case 1 (Figure 14) is the canonical form in which only the target language activates in the positive region while the remaining languages are distributed in the negative region, and both identifiers identify the same neuron as a target-language SLN. Cases 2–4(Figure 15–17) are all SLN-only samples that LAPE fails to identify: case 2 is a sign-inverted form in which the target language lies in the negative region and the remaining languages lie in the positive region; case 3 is a form in which all languages reside in the positive region but the target language is separated only by magnitude; and case 4 is a form in which all languages are distributed in the negative region but only the target language is separated by magnitude.

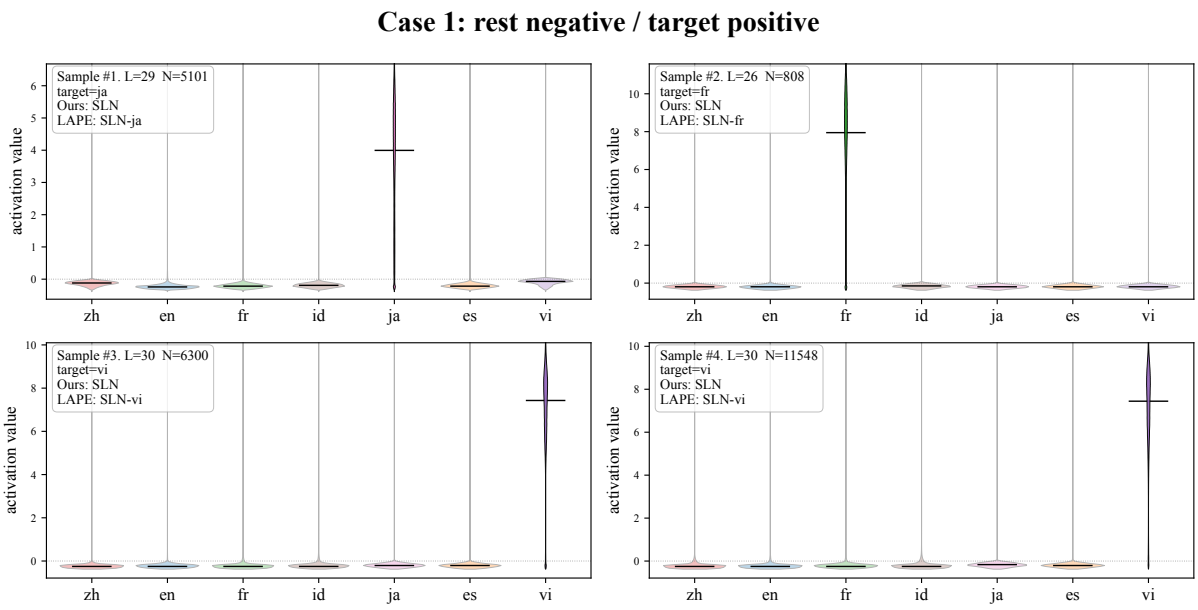


Figure 14: Four SLN samples for case 1 (rest negative / target positive). Both identifiers agree on the target-language SLN; inset box reports $(L, N,$ target language, Ours / LAPE verdict$)$.

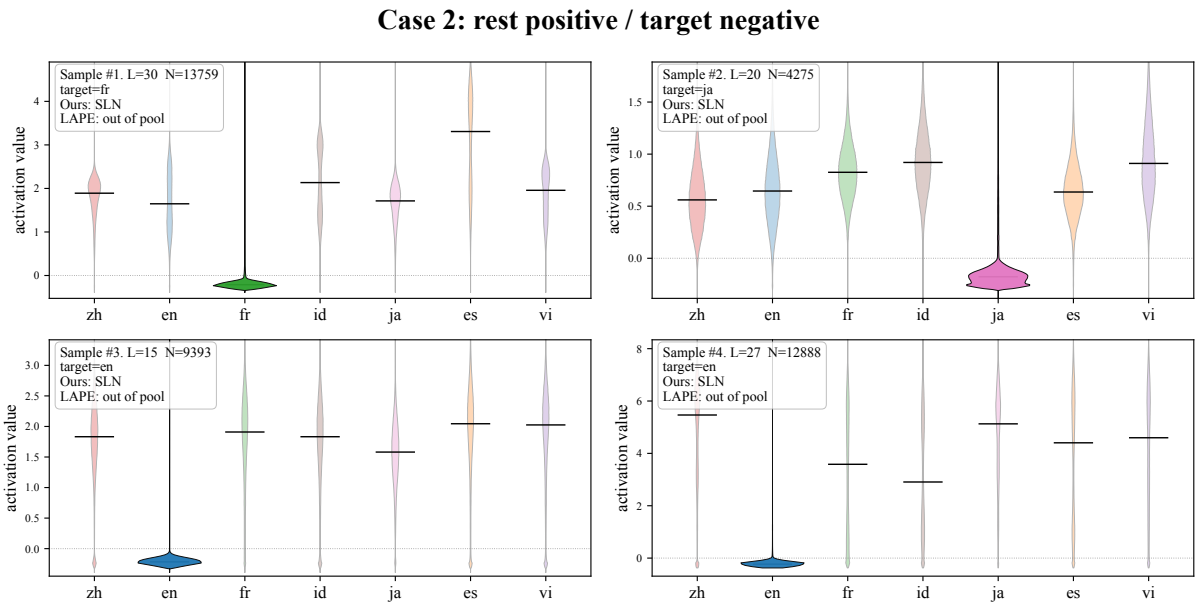


Figure 15: Four SLN samples for case 2 (rest positive / target negative). SLN assigns the target-language SLN while LAPE misses — the sign reversal leaves the normalized activation-rate entropy near-uniform; inset box reports $(L, N,$ target language, Ours / LAPE verdict$)$

Case 3: rest positive / target positive, different shape

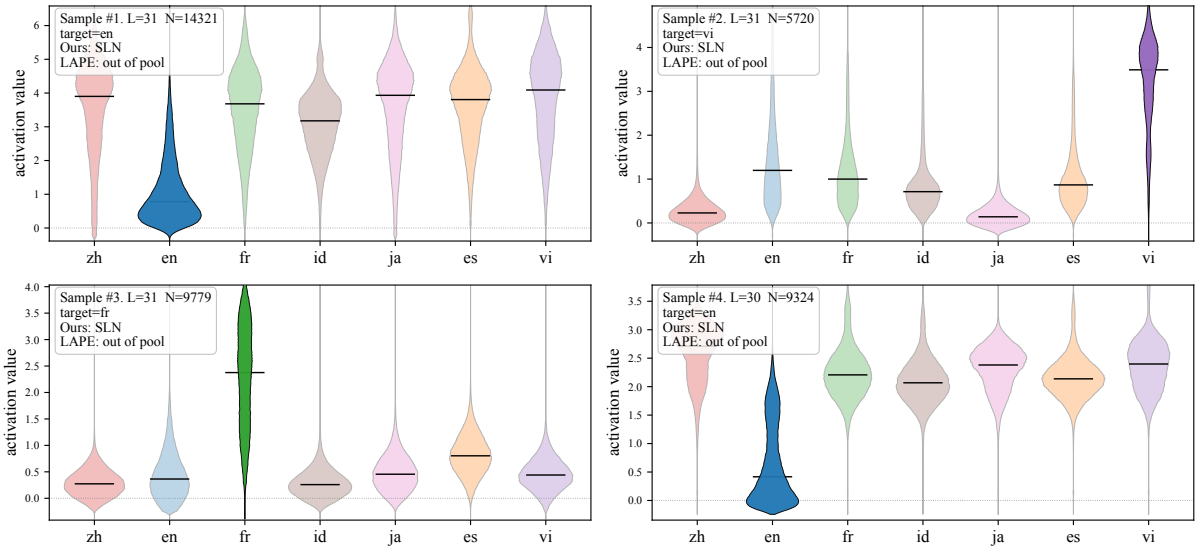


Figure 16: Four SLN samples for case 3 (rest positive / target positive, different shape). SLN assigns the target-language SLN while LAPE misses — all languages clear $a > 0$, so the activation-rate statistic carries no separating signal; inset box reports $(L, N,$ target language, Ours / LAPE verdict$)$.

Case 4: rest negative / target negative, different shape

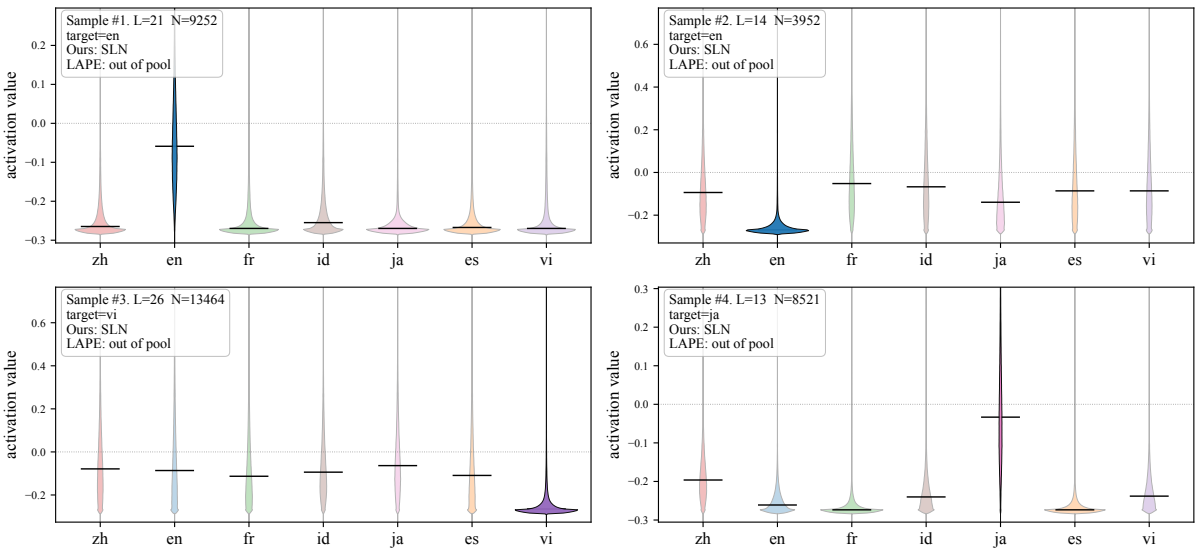


Figure 17: Four SLN samples for case 4 (rest negative / target negative, different shape). SLN assigns the target-language SLN while LAPE excludes these neurons from its pool — no language clears the max-rate threshold τ; inset box reports $(L, N,$ target language, Ours / LAPE verdict$)$

I Detailed Neuron Population

I.1 Language Population in SLN, MLN

Model	Language	|SLN|	|MLN|	Ratio
Llama-3.1-8B	zh	27	376	**93.3%**
	ja	247	371	**60.0%**
	vi	199	148	42.7%
	en	188	76	28.8%
	es	616	220	26.3%
	fr	780	198	20.2%
	id	915	84	8.4%
SmolLM3-3B	zh	44	325	**88.1%**
	ja	235	337	**58.9%**
	en	152	75	33.0%
	vi	357	161	31.1%
	fr	475	133	21.9%
	es	425	119	21.9%
	id	566	76	11.8%
Qwen3-4B	zh	113	567	**83.4%**
	en	120	214	**64.1%**
	ja	663	546	45.2%
	es	534	341	39.0%
	vi	287	181	38.7%
	fr	651	331	33.7%
	id	690	196	22.1%

Table 22: Per-language counts of SLNs and MLNs, on Llama-3.1-8B, SmolLM3-3B-Base, and Qwen3-4B-Base. $|\cdot|$ denote the number of neurons associated with each language. Ratio $= \frac{|\text{MLN}|}{|\text{SLN}|+|\text{MLN}|}$ gives the share of MLNs among all language neurons.

Figure 18 illustrates the number of SLNs and MLNs for each language across three models: Llama-3.1-8B, SmolLM3-3B-Base, and Qwen3-4B-Base. The "top-n MLN" denotes the n-th most frequent cluster among the MLNs associated with a given language. The overall distributions of SLNs and MLNs exhibit similar patterns across all three models. For SLNs, `id` accounts for the highest proportion, whereas `zh` exhibits the lowest distribution in all models. Notably, for Chinese (`zh`), the proportion of MLNs is higher than that of SLNs across all models, with the {`zh`, `ja`} cluster accounting for the largest share among the MLN populations. This indicates that Chinese and Japanese share the most information within these models.

The distributions of the most frequent MLN clusters, such as {`fr`, `es`}, {`vi`, `zh`, `ja`}, and {`id`, `vi`}, are also consistent across the models. These results demonstrate that the proposed methodology is model-agnostic. Furthermore, this demonstrates that our proposed methodology enables the analysis of linguistic similarities and relational dynamics, such as those between Chinese and Japanese.

While the {`en`, `fr`, `es`} cluster is the most frequent among English MLNs in the other two models, the {`en`, `zh`} cluster is the most frequent in the Qwen3-4B-Base model. This discrepancy can be attributed to the Qwen3-4B-Base model being heavily specialized in Chinese compared to the others, treating it as a major language with a representation hierarchy comparable to English.

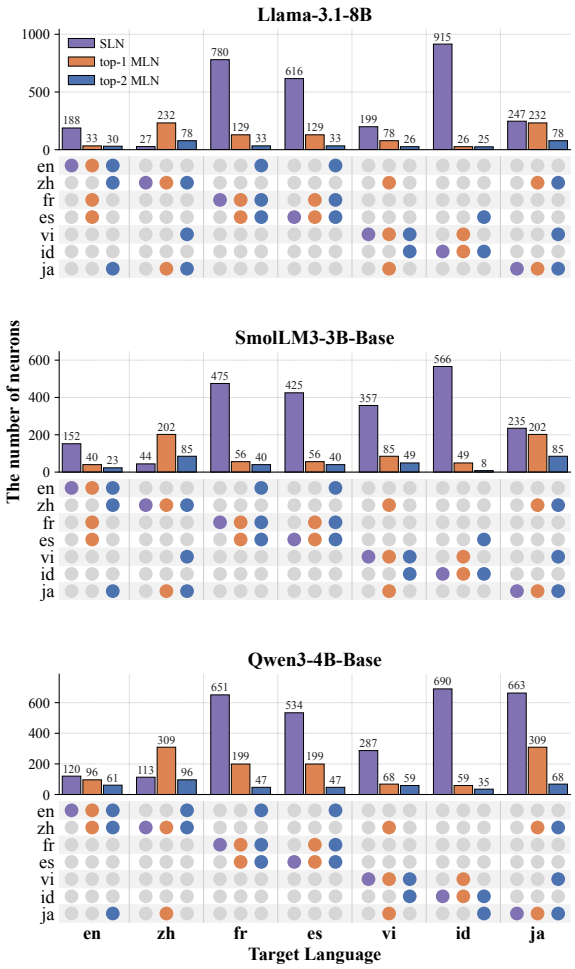


Figure 18: Language distribution of SLN and MLN neurons across different languages in three language models. The top-n MLN denotes the MLN cluster with the n-th highest frequency among all MLN clusters that include the corresponding language.

I.2 Layer-wise Neuron Population

In this section, we examine the layer-wise distributions and differences of the neurons identified by the previous works and our proposed approach. Similar to LAPE, both methodologies exhibit a U-shaped distribution, where the identified neurons are concentrated in the early and late layers. A key difference is that the neurons selected by our proposed methodology are more densely clustered in the extreme early and late layers compared to LAPE. Furthermore, MLNs are rarely observed in the initial layers but become increasingly prevalent in deeper layers. Detailed layer-wise statistics are provided in Tables 23 and 24.

Layer	LAPE	LSN	SLN	LRN	MLN	LAPE\SLN	SLN\LAPE	LAPE∩SLN
0	43	6	22	15	2	42	21	1
1	2	3	11	5	2	2	11	0
2	11	6	5	2	0	8	2	3
3	90	40	91	106	17	56	57	34
4	34	7	7	19	6	30	3	4
5	19	4	7	14	3	16	4	3
6	12	2	6	4	1	10	4	2
7	5	0	4	3	0	5	4	0
8	23	9	3	3	1	23	3	0
9	36	8	8	23	11	34	6	2
10	34	5	9	27	9	30	5	4
11	21	5	5	12	2	20	4	1
12	15	1	4	10	3	14	3	1
13	8	3	11	13	1	7	10	1
14	29	7	12	20	13	24	7	5
15	38	13	14	23	12	34	10	4
16	75	16	29	47	20	66	20	9
17	152	28	36	92	22	132	16	20
18	146	41	37	69	20	125	16	21
19	138	36	39	47	12	111	12	27
20	126	33	40	37	10	105	19	21
21	162	46	56	37	12	132	26	30
22	179	46	60	29	9	144	25	35
23	112	35	64	39	9	85	37	27
24	223	83	116	65	12	163	56	60
25	212	69	111	72	18	165	64	47
26	416	143	200	136	32	332	116	84
27	604	262	342	195	37	458	196	146
28	691	314	504	297	43	515	328	176
29	406	233	473	344	61	304	371	102
30	348	238	423	407	88	254	329	94
31	177	72	223	504	166	152	198	25
Total	4587	1814	2972	2716	654	3598	1983	989

Table 23: Per-layer neuron-population counts for Llama-3.1-8B at $P = 1\%$. Each cell is the number of neurons of the given population in that layer; set-difference and intersection columns are over (layer, neuron) pairs.

Layer	LAPE	LSN	SLN	LRN	MLN	LAPE\SLN	SLN\LAPE	LAPE∩SLN
0	44	81	131	20	0	23	110	21
1	8	5	5	0	0	5	2	3
2	145	65	80	15	4	80	15	65
3	51	19	14	8	1	44	7	7
4	243	59	41	58	3	216	14	27
5	9	1	17	3	1	7	15	2
6	3	2	2	2	1	3	2	0
7	19	5	4	4	3	19	4	0
8	18	5	2	4	2	17	1	1
9	14	2	0	3	0	14	0	0
10	4	0	1	2	1	4	1	0
11	6	1	1	1	2	6	1	0
12	6	1	0	3	1	6	0	0
13	2	0	3	5	0	2	3	0
14	5	1	3	3	1	5	3	0
15	16	0	6	13	3	13	3	3
16	6	2	2	7	1	4	0	2
17	34	9	8	18	6	29	3	5
18	15	4	14	19	3	15	14	0
19	56	15	14	43	25	49	7	7
20	46	10	8	28	16	44	6	2
21	39	11	18	41	8	33	12	6
22	82	31	28	56	13	60	6	22
23	103	33	36	74	26	82	15	21
24	161	52	47	59	18	131	17	30
25	150	43	37	63	18	128	15	22
26	147	30	43	75	16	128	24	19
27	208	64	67	80	24	173	32	35
28	192	42	58	65	12	164	30	28
29	225	56	78	65	11	198	51	27
30	281	103	180	164	20	237	136	44
31	253	106	225	166	18	210	182	43
32	291	134	221	145	24	246	176	45
33	296	144	249	171	24	228	181	68
34	390	143	297	345	74	297	204	93
35	394	164	314	520	150	293	213	101
Total	3962	1443	2254	2348	530	3213	1505	749

Table 24: Per-layer neuron-population counts for SmolLM3-3B-Base at $P = 1\%$. Each cell is the number of neurons of the given population in that layer; set-difference and intersection columns are over (layer, neuron) pairs.