The Transfer Neurons Hypothesis: An Underlying Mechanism for Language Latent Space Transitions in Multilingual LLMs

Hinata Tezuka[1] **Naoya Inoue**[1,2]
[1]Japan Advanced Institute of Science and Technology [2]RIKEN
{hinata-t, naoya-i}@jaist.ac.jp

Abstract

Recent studies have suggested a processing framework for multilingual inputs in decoder-based LLMs: early layers convert inputs into English-centric and language-agnostic representations; middle layers perform reasoning within an English-centric latent space; and final layers generate outputs by transforming these representations back into language-specific latent spaces. However, the internal dynamics of such transformation and the underlying mechanism remain underexplored. Towards a deeper understanding of this framework, we propose and empirically validate **The Transfer Neurons Hypothesis**: certain neurons in the MLP module are responsible for transferring representations between language-specific latent spaces and a shared semantic latent space. Furthermore, we show that one function of language-specific neurons, as identified in recent studies, is to facilitate movement between latent spaces. Finally, we show that transfer neurons are critical for reasoning in multilingual LLMs[1].

1 Introduction

Multilingual **L**arge **L**anguage **M**odels (LLMs), pre-trained on multilingual corpora, can process multiple languages within a single model. Recent studies suggest that decoder-based multilingual LLMs operate on a semantic latent space that is independent of the input language, e.g., language-agnostic latent space (Zhang et al., 2024; Zeng et al., 2025; Bandarkar et al., 2025) and English latent space (Zhao et al., 2024b; Wendler et al., 2024; Schut et al., 2025).

The previous findings imply that multilingual LLMs potentially process inputs by (i) transferring input into the shared latent space, (ii) performing semantic processing, and (iii) mapping results back to the input language, as shown in Fig. 1 (a). However, the dynamics of internal representation transformation and the underlying mechanism of the representation transfer remain underexplored, preventing a comprehensive understanding of semantic processing in multilingual LLMs.

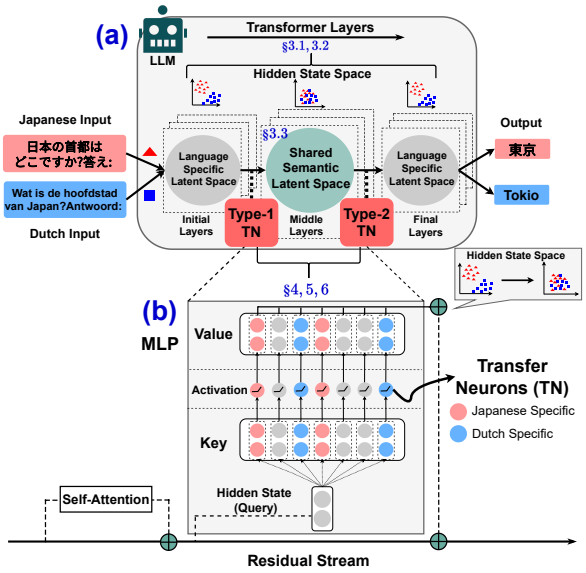


Figure 1: **An Overview of the Transfer Neurons Hypothesis**. We hypothesize that specific neurons in MLP module move representations from language-specific latent spaces to a shared semantic latent space, and vice versa.

To address these issues, first, we study how internal representations evolve across layers and examine the existence of a shared semantic latent space in the middle layers. Second, as a plausible explanation for the representation transfer mechanism, we propose the *Transfer Neurons Hypothesis*: specific neurons in the **M**ulti **L**ayer **P**erceptron (MLP) module are responsible for this transfer. As shown in Fig. 1 (b), we hypothesize that certain neurons, named *transfer neurons*, are sensitive to the input language, and that their firing results in adding corresponding value vectors to the residual stream,

[1]Our codes are available at https://github.com/HinaTezuka/emnlp2025-transfer-neurons.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31731–31781
November 4-9, 2025

thereby facilitating the representation transfer.

Our Contributions are summarized as follows:

- We uncover the dynamics of internal representations in multilingual LLMs: initial representations are language-specific, subsequently converge into a shared latent space across languages, and finally diverge again (§3).
- We empirically demonstrate the existence of transfer neurons that mediate transitions between language-specific latent spaces and the shared semantic latent space. We further show their necessity for the representation transfer through intervention experiments (§4, §5).
- We suggest that one of the key role of language-specific neurons identified in recent studies is to facilitate movement between latent spaces, and show that transfer neurons are essential for downstream tasks (§6).

2 Background

2.1 Related Works

The Framework for Multilingual Processing. By detecting and controlling language-specific neurons, Zhao et al. (2024b) introduced the `MWork` framework, arguing that LLMs process multilingual queries in the first few layers, reason in English in the middle layers, and subsequently generate responses in input language. Although they do not explicitly define it as English-centric, Zeng et al. (2025) proposed a similar framework by introducing the concept of a `Lingua Franca` — a common semantic latent space that facilitates language-agnostic processing. Assuming language-agnostic reasoning occurs in the middle layers, Bandarkar et al. (2025) improved multilingual reasoning performance using a layer swapping technique. They trained language-expert and reasoning-expert models separately, then merged them by assigning the language-expert to the first and last few layers, and the reasoning-expert to the middle layers. Wendler et al. (2024); Schut et al. (2025) observed that when non-English queries are presented, the models often identify an answer in English internally, and then translate it into the target output language.

Language-Agnostic Representations. It has been shown that knowledge neurons cross-lingually encoding specific concepts exist (Chen et al., 2024; Zhao et al., 2024a; Zhang et al., 2025). Building on findings that encoder-based LLMs exhibit a degree of language-agnosticism (Pires et al., 2019; Libovický et al., 2020), Yoon et al. (2024) developed the LangBridge model, which connects a multilingual encoder to LLM, and improved performance on reasoning tasks even in low-resource languages.

Language-Specific Regions. Recent studies (Kojima et al., 2024; Tang et al., 2024; Zhao et al., 2024b; Zeng et al., 2025; Duan et al., 2025) have revealed the existence of language-specific regions or neurons in decoder-based LLMs, which respond strongly to inputs in a particular language. These regions are primarily distributed in the initial and final few layers of the models, a finding that is also supported by our experiment in Appendix H.3. Mondal et al. (2025) demonstrated that fine-tuning only these neurons is not sufficient for cross-lingual improvements on downstream tasks. The role of these regions or neurons, however, has not yet been sufficiently revealed.

Although these studies offer valuable yet fragmentary evidence, the overall process of the latent space transitions still remains unclear and lack quantitative demonstration.

2.2 Neuronic View of MLPs in Transformers

Given an input vector $\boldsymbol{x} \in \mathbb{R}^d$, the MLP module in Transformer at layer l operates as follows:

$$\mathrm{MLP}^l(\boldsymbol{x}) = a(\boldsymbol{x}M^l_{\mathrm{up}})M^l_{\mathrm{down}} \tag{1}$$

where a represents a non-linear activation function, and $M^l_{\mathrm{up}} \in \mathbb{R}^{d\times d_m}, M^l_{\mathrm{down}} \in \mathbb{R}^{d_m\times d}$ refer to the up and down projection matrices, respectively. Geva et al. (2021) proposed a vector-based key-value store view of the MLP module. Here, the keys are d_m column vectors in M^l_{up}, the values are d_m row vectors in M_{down}, and the query is $\boldsymbol{x}$. Viewing $\boldsymbol{\alpha}^l = a(\boldsymbol{x}M^l_{\mathrm{up}})$ as the relevance scores between the keys and query, the MLP module can be rewritten as follows:

$$\mathrm{MLP}^l(\boldsymbol{x}) = \boldsymbol{\alpha}^l M^l_{\mathrm{down}} = \sum_{i=1}^{d_m} \alpha^l_i \boldsymbol{v}^l_i \tag{2}$$

where $\boldsymbol{v}^l_i$ represents i-th row vector in M^l_{down}. In our experiments, we adopt the MLP module with a gating mechanism (Shazeer, 2020; Liu et al., 2021), namely $\boldsymbol{\alpha}^l = a(\boldsymbol{x}M^l_{\mathrm{gate}}) \odot \boldsymbol{x}M^l_{\mathrm{up}}$, where $\odot, M^l_{\mathrm{gate}} \in \mathbb{R}^{d\times d_m}$ represents element-wise multiplication and gated projection, respectively. Following Dai et al. (2022), we call each α^l_i *neuron*.

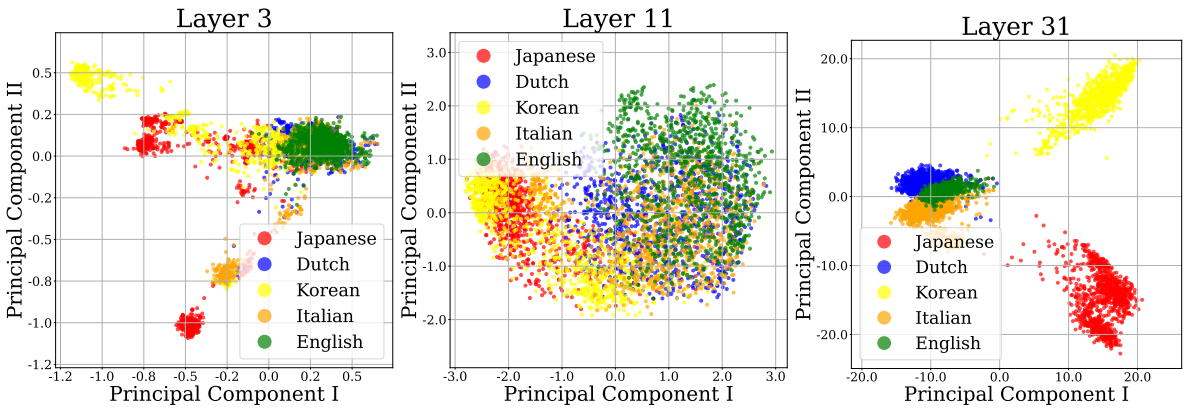


Figure 2: **The results of PCA applied to hidden language representations for capturing spatial transitions (LLaMA3-8B).** Japanese, Italian, Duch, English and Korean hidden representations, respectively. Each input sentence from each language has the same meaning. The results for other layers and models can be found in Appendix F.

3 The Dynamics of Hidden States

Based on the findings of prior work (Zhao et al., 2024b; Wendler et al., 2024; Schut et al., 2025), we assume that a shared semantic latent space is predominantly English-centric. In other words, the English latent space effectively serves as the shared semantic space across languages. Accordingly, we adopt English as the core language within this shared latent space throughout the experiments in this study.

In this section, we provide several evidence supporting the latent space transitions across layers and the existence of the shared semantic latent space in middle layers, as illustrated in Fig. 1(a).

3.1 Spatial Transition Phenomenon

Representations Form Language-Specific Latent Spaces and a Shared Latent Space in Hidden State Space. We encode 5k sentences from MKQA dataset (Longpre et al., 2021), a multilingual dataset of parallel QA sentences in five languages, using LLaMA3-8B (1k sentences per language). We extract the hidden states corresponding to the final token of each sentence from each layer, apply PCA, and plot the results along the top two principal components in Fig. 2. It shows that in the initial layers, each language forms its own latent space, which gradually converges into a shared latent space towards the middle layers. In the final layers, each language re-establishes its distinct latent space, which confirms our hypothesis. An unexpected finding, however, is that two latent spaces specific to Dutch and Italian, which have linguistic proximity to English, remain close to the English latent space even in the initial and final layers, in contrast to Japanese and Korean. Additionally, Japanese and Korean, which are linguistically more distant from English, do not fully converge into the shared latent space even in the middle layers. These trends are consistent across models (see Appendix F). These results suggests that while representations tend to converge into the English-centric shared semantic latent space, the degree of convergence varies across languages.

3.2 Geometry of Language-Specific Latent Spaces

Hidden States Reside in Language Latent Spaces. To investigate whether hidden states for each language can be captured by low dimensional latent space, we conduct a **S**ingular **V**alue **D**ecomposition (SVD) to a matrix formed by hidden states of each language and compute the cumulative explained variance ratio, to estimate the number of basis vectors necessary to form each language latent space. Let $M^l_{L_i} \in \mathbb{R}^{n \times d}$ be a matrix containing hidden state vectors at layer l (each of dimension d) for n sentences in language L_i. Then, $M^l_{L_i}$ is decomposed as follows:

$$M^l_{L_i} = U \Sigma V^T \tag{3}$$

where U and V are orthogonal matrices that can be interpreted as a set of orthonormal basis for row space and column space, respectively. Σ is a diagonal matrix containing the singular values of $M^l_{L_i}$, which represent the magnitude of each corresponding component in the lower dimensional latent space in hidden state space. These singular values are ordered in descending order. Finally, we

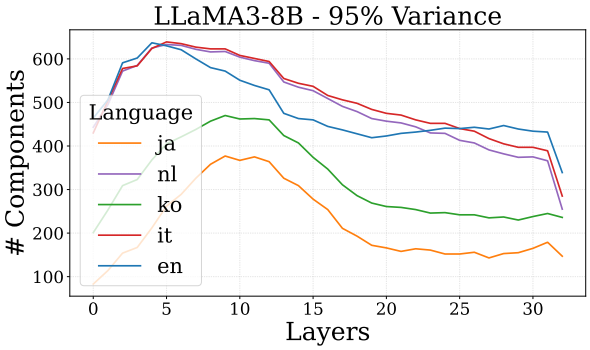


Figure 3: **Dimensionality of language latent spaces across layers estimated via SVD and CEVR (LLaMA3-8B).**

compute the cumulative explained variance ratio as follows:

$$\text{CEVR}^l_{L_i} = \sum_{i=1}^{k} \frac{\sigma_i^2}{\Sigma_{j=1}^{r} \sigma_j^2} \tag{4}$$

where σ_i^2 denotes the i-th singular value, k is the number of components retained, and r is the total number of non-zero singular values. The $\text{CEVR}^l_{L_i}$ thus represents the proportion of the total variance explained by the top-k singular values.

The results are shown in Fig. 3. As indicated, in initial and final few layers where hidden states specific to each language form their own latent space, most (95%) of representations lie in a relatively low-dimensional latent space, a tendency that is expecially pronounced for languages that are more distant from English (i.e., Japanese and Korean). On the other hand, representations in the middle layers, where semantic processing and inference mainly occur, form a relatively high-dimensional latent space. Additionally, as quantitatively demonstrated in Appendix D.2, the transitions of the distance among the centroids of language-specific latent spaces across layers are well aligned with the observation of spatial transition in §3.1 and Fig. 2. The results for other variance thresholds and models can be found in Appendix D.1.

These results indicate that the hidden states of each language shown in Fig. 2 lie in a relatively low dimensional latent space of the hidden state space.

3.3 The Existence of a Shared Semantic Latent Space in Middle Layers

Sentence Representations with Similar Meanings across Languages Converge to Similar Locations in Middle Layers. To numerically

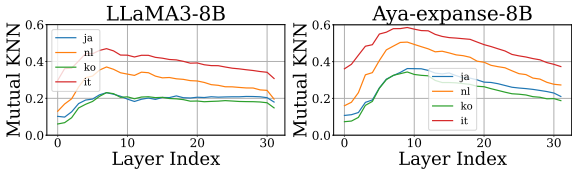


Figure 4: **Kernel-based similarity between language latent spaces across layers (LLaMA3-8B and Aya expanse-8B, k=5).** English-Italian, English-Dutch, English-Japanese, English-Korean pairs, respectively.

demonstrate whether language-agnostic semantic processing indeed occurs in the shared semantic latent space shown in Fig. 1(a), we measure *Mutual k-Nearest Neighbor Alignment Metric* proposed in (Huh et al., 2024), between a set of parallel sentence representations across language pairs. If similarities peak in middle layers, it suggests that semantically similar or equivalent sentence representations converge to similar positions across languages in hidden state space. This increases shared nearest neighbors, indicating aligned latent spaces and language-agnostic semantic processing. The computation is formalized in Appendix E.1.

Fig. 4 shows that latent space similarity between English and other languages peaks in the middle layers, suggesting shared, language-agnostic semantic processing. As noted in Appendix E.2, this pattern holds across models.

Additionally, in Appendix C, we measured the similarity of hidden states and activation patterns for parallel sentences across languages, as well as the linear separability of representations for parallel and non-parallel sentences, further supporting the existence of this processing.

4 Identifying Transfer Neurons

4.1 Hypothesis: Specific Activations and Value Vectors Facilitate a Parallel Shift of Representations to the Target Latent Space

Based on the Eqs. 1 and 2, we hypothesize that certain activations and their corresponding value vectors $\alpha_i^l \boldsymbol{v}_i^l$ are responsible for shifting internal representations between language-specific latent spaces and the shared semantic latent space, as described in Fig. 1(b). These activation units are what we define as the *Transfer Neurons*.

4.2 Preparation: Two Types of Transfer Neurons

As discussed in §1, we consider two types of neurons in the context of transfer neurons:

1. Neurons that transfer input representations from the language-specific latent space to the shared semantic latent space, located in the initial layers (**Type-1 Transfer Neurons**).

2. Neurons that transfer reasoned representations from the shared semantic latent space to the language-specific latent space for output generation, located in the final layers (**Type-2 Transfer Neurons**).

Based on the observation that the difference in representation similarity between parallel and non-parallel sentence pairs widens in specific layers (Fig. 10 in Appendix C.1), we target layers 1–20 for detecting Type-1 neurons and layers 21–32 (up to the final layer) for detecting Type-2 neurons[2].

4.3 Scoring the Candidate Neurons

Our goal is to assign higher scores to candidate neurons whose activations and value vectors move the representation closer to the desired latent space, when the model is given an input sentence in the target language.

Centroids Estimation for Latent Spaces. We begin by estimating the centroids of each representational latent space, which allows us to compute the distances to these latent spaces. Let $\boldsymbol{h}^l_{\text{L2},k}$ denote the hidden state from the l-th layer corresponding to the k-th sample sentence in the particular non-English language "L2". The centroids of the respective latent spaces are estimated as follows:

$$\boldsymbol{C}^l_{\text{L2}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{h}^l_{\text{L2},k} \tag{5}$$

$$\boldsymbol{C}^l_{\text{shared}} = \frac{1}{n}\sum_{k=1}^{n} \text{mean}(\boldsymbol{h}^l_{\text{en},k}, \boldsymbol{h}^l_{\text{L2},k}) \tag{6}$$

where $\boldsymbol{C}^l_{\text{L2}}$ denotes the centroid of the L2-specific latent space in the l-th layer, whereas $\boldsymbol{C}^l_{\text{shared}}$ is the centroid of the shared semantic latent space in the l-th layer, and n is the total number of sample sentences. To compute Eq. 6, we use parallel sentence pairs from an English–L2 pair. This is because, as mentioned in §3, we assume that the English latent space serves as the shared semantic latent space[3].

Scoring Methodology. We score each candidate neuron in the target layers with the following computations:

$$L^l_k = \text{dist}\left((\boldsymbol{h}^{l-1}_k + \boldsymbol{A}^l_k), \boldsymbol{C}^l\right) \tag{7}$$

$$N^l_{i,k} = \text{dist}\left((\boldsymbol{h}^{l-1}_k + \boldsymbol{A}^l_k + \alpha^l_{i,k}\boldsymbol{v}^l_{i,k}), \boldsymbol{C}^l\right) \tag{8}$$

$$\text{Score}^l_i = \frac{1}{n}\sum_{k=1}^{n}(N^l_{i,k} - L^l_k) \tag{9}$$

where $\boldsymbol{h}^{l-1}_k$ is the hidden state from the previous layer, and $\boldsymbol{A}^l_k$ is the output of the self-attention in the l-th (current) layer. Here, k refers to the index of the sample sentence. Therefore, $(\boldsymbol{h}^{l-1}_k + \boldsymbol{A}^l_k)$ represents the hidden state at the l-th layer immediately before the MLP module for the k-th input sample. The *dist* function measures how close two vectors are[4]. L^l_k denotes the layer score, which indicates how close the hidden state before the MLP is to the l-th layer centroid of the target latent space. $N^l_{i,k}$ denotes the neuron score, which expresses how effectively the neuron and its corresponding value vector, on their own, bring the representation closer to the centroid of the target latent space, in comparison to L^l_k.

If the score of the i-th neuron in the l-th layer, $\text{Score}^l_i > 0$, it indicates that the neuron and value vector brings representations closer to the centroid of the target latent space for most of the samples. Conversely, if $\text{Score}^l_i < 0$, it suggests that the neuron pushes the hidden states further apart from the centroid across various samples. We set $\boldsymbol{C}^l$ to $\boldsymbol{C}^l_{\text{shared}}$ for detecting Type-1 neurons, and to $\boldsymbol{C}^l_{\text{L2}}$ for detecting Type-2 neurons. The larger the Score^l_i is, the more effectively the neuron and its corresponding value vector bring the representations closer to the target latent space. Finally, we sort all candidate neurons in descending order based on Score^l_i, and extract the top-n neurons.

[2]All LLMs adopted in this study have 32 decoder layers.

[3]The rationale for using the centroid for every candidate layer is explained in Appendix B.

[4]We adopt Cosine similarity as a *dist* function in this study; however, we also experimented with Euclidean distance, which yielded similar results.

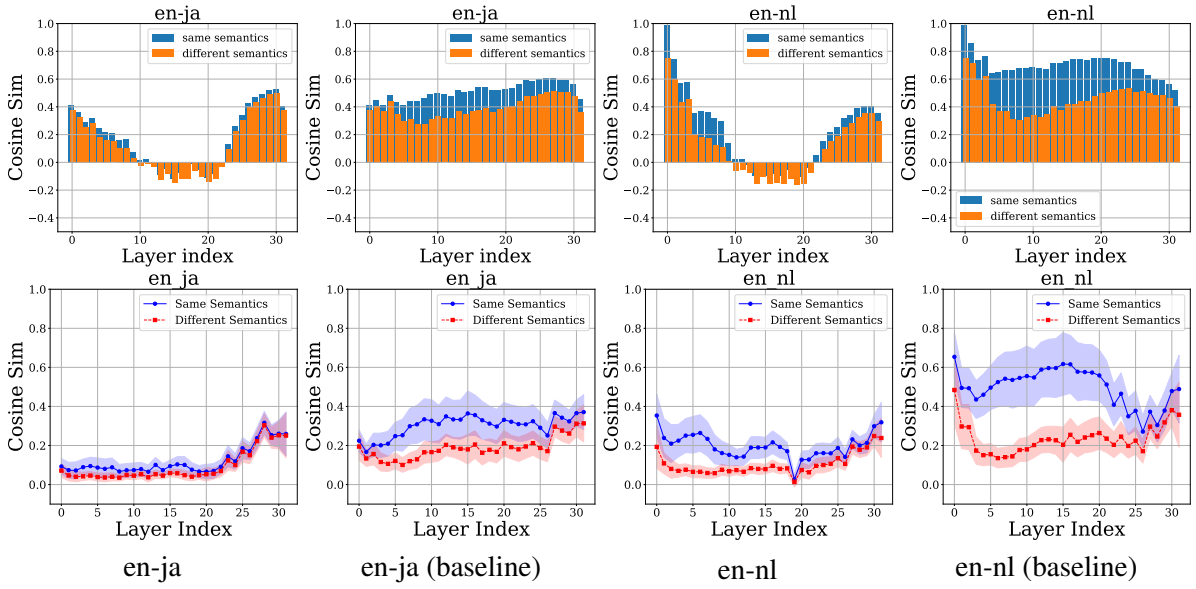


Figure 5: **Similarity of hidden states and activation patterns across layers while deactivating top-1k Type-1 Transfer Neurons (LLaMA3-8B).** The upper figures present the similarity of hidden states, and the lower ones present the similarity of activation values.

5 Detecting and Controlling Transfer Neurons

5.1 Distribution

Fig. 6 and Appendix H.1 shows the distributions of two types of transfer neurons across layers. For Type-1 neurons, the first layers and the middle layers contain a noticeable amount. In contrast, Type-2 neurons are predominantly found in the final layer, which aligns with the observation that representations exhibit the largest shift in the final layers, as shown in Appendix F. Additionally, these distributions closely resemble those of language-specific neurons, as shown in Appendix H.3. These tendencies are consistent across languages and models.

5.2 Representation Similarity Measurement while Deactivating Transfer Neurons

Type-1 Neurons Facilitate the Mapping of Representations to the Shared Semantic Latent Space. Fig. 5 presents the similarity of hidden states and MLP activation patterns for parallel and non-parallel English–L2 sentence pairs. The measurements are taken while deactivating the top-1k Type-1 neurons, which account for only **0.2%** of the total neuron population[5]. Although only a very small proportion of neurons were deactivated, we observe a sharp reduction in the difference in similarity between parallel and non-parallel sentence pairs for both activation patterns and hidden states. By contrast, the baseline — defined as deactivating 1k randomly sampled neurons from the same layers as the Type-1 neurons — has almost no effect on the similarity compared to the original state (Fig. 10 and 11 in Appendix C.1). This tendency is consistent across other language pairs and models.

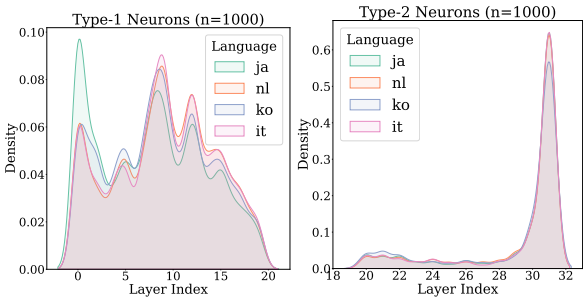


Figure 6: **Distribution of top-1k Transfer Neurons (LLaMA3-8B).** Left shows Type-1 neurons, whereas right shows Type-2 neurons.

These results indicate that the models deactivated Type-1 neurons are unable to map input representations to the correct positions in the shared semantic latent space in a way that reflects sentence meaning, since the similarity remains nearly unchanged regardless of whether the pairs are parallel or non-parallel. This indicate that Type-1 neurons we detected play a critical role in this mapping process.

Also, we surmise that the relatively remaining difference of similarity between parallel and non-

[5]Here, "deactivating" refers to setting the activation values of the target neurons to zero, thereby nullifying their effect.

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.16	0.03	0.05	0.03	**Type-1 TN**	0.16	0.02	0.05	0.03
Type-2 TN	0.25	0.22	0.17	0.16	**Type-2 TN**	0.40	0.27	0.33	0.19

Table 1: **Correlation ratio of top-100 and top-1k Transfer Neurons for language specificity (LLaMA3-8B).** Typically, a correlation ratio above **0.1** suggests a correlation, above **0.25** suggests a moderately strong correlation, and above **0.5** indicates a strong correlation.

Top-1000	ja-nl	ja-ko	nl-it	ja-it
Type-1 TN	0.39	0.51	0.75	0.41
Type-2 TN	0.23	0.37	0.51	0.26

Table 2: **Jaccard index for top-1k Transfer Neurons across language pairs. (LLaMA3-8B).** A score closer to 1 indicates a greater overlap between the neurons of each language pair.

parallel sentence pairs in the en-nl (Dutch) and en-it (Italian) settings across models (Fig. 5 and Appendix G.2.2) — even after the deactivation of the top-1k Type-1 neurons — is due to the linguistic and spatial proximity of these languages to English. That is, even if we disturb representations from shifting into the shared semantic latent space by deactivating Type-1 neurons, they are already very close to the english-centric shared semantic latent space (see §3.1 and Appendix F).

Appendix G.2.1 shows that deactivating Type-1 neurons significantly reduces kernel-based similarity, underscoring their role in aligning representations to a shared semantic latent space.

Ablation studies, including experiments with different numbers of deactivated neurons, are provided in Appendix G.2.2.

The results of deactivating Type-2 neurons are presented in Appendix G.3, including evidence of a significant inhibition in latent space transitions when the neurons are deactivated.

6 The Nature of Transfer Neurons

6.1 Language and Language-Family Specificity

Language Specificity. To investigate the language specificity of transfer neurons, we measure the correlation ratio between neuron activations and sentence labels[6], assigning `label1` to sentences in the target language and `label0` to all others. This allows us to measure the strength of the correlation between the activations of the transfer neurons and inputs in a specific language.

[6]Detailed explanation about the correlation ratio is given in Appendix H.6

As shown in Tab. 1, Type-1 neurons generally do not exhibit correlation, except for those involved in shifting representations from the Japanese latent space to the shared semantic latent space. We surmise that this is because the Japanese-specific latent space in first few layers is highly distant from the English-centric shared latent space (see PCA results in Appendix F), which is why certain Type-1 neurons must undergo a considerable shift to align with the shared latent space.

On the other hand, it turns out that **for Type-2 neurons, the more their activations and corresponding value vectors move representations towards the language-specific latent spaces, the stronger the correlation with the input sentences of the target language**. In other words, the neurons that strongly shift representations from the shared semantic latent space to each language-specific latent space for output generation can be considered language-specific. This is further supported by the high similarity in the distributions between language-specific neurons and Type-2 neurons, as demonstrated in Appendix H.1 and H.3.

For both Type-1 and Type-2 neurons, the strength of the correlation ratio score correlates with the linguistic distance from English, suggesting that the greater the distance neurons must shift to the target latent space, the more language-specific those neurons are likely to be (As shown in Tab. 1, §3.1, Appendices F, and H.4).

Additionally, when comparing the distributions of the language-specific neurons (detected in Appendix H.3) with PCA results of language-specific latent spaces (Appendix F) and the correlation ratio scores, we observe that **languages whose original latent spaces are closer to the English-specific latent space in the initial layers (such as Dutch and Italian) tend to exhibit significantly sparser distributions of language-specific neurons in those layers, along with relatively low correlation ratio scores**. We surmise that this is because these languages require fewer language-specific neurons, as they are already closely aligned with the English-centric shared semantic latent space.

These results suggest one of the key roles of language-specific neurons, as identified in recent studies: **facilitating the movement of internal representations between language-specific latent spaces and a shared semantic latent space**.

The results for other values of n (i.e., top-n neurons) and models can be found in Appendix H.4.

Language-Family Specificity. Tab. 2 presents the Jaccard Index measurements for transfer neurons specific to each language, providing insight into the extent of overlap among transfer neurons. As shown, language pairs that are linguistically similar tend to exhibit relatively greater overlap, with this effect being particularly pronounced in Type-2 neurons. The results suggest that, from the perspective of the shared semantic latent space, the locations of each language-specific latent space within the hidden state space of the model are likely to be similar, which explains why there are many overlapping neurons used to shift to the target latent space. This is further supported by the visualization of each language latent space (Fig. 2, Appendix F). In Appendix H.5.2, we show the results of other models and correlation ratio measurement to further verify language-family specificity of these neurons.

The results of the hypothesis testing conducted to verify the reliability of the correlation scores reported above are provided in Appendix H.7.

6.2 Assessing the Importance of Transfer Neurons in Reasoning

In this section, we examine the role of Type-1 neurons in enabling reasoning. Specifically, we test whether disrupting the shift of input representations to the shared semantic latent space during inference degrades performance. A notable drop when deactivating these neurons — thereby disrupting the inference framework of the hypothesis in Fig. 1 — would suggest that the model struggles to generate appropriate responses when inputs are misaligned with the shared semantic latent space.

Task and Evaluation Metric. We use a simple multilingual knowledge QA task called MKQA (Longpre et al., 2021), which consists of 10k knowledge question-answer pairs aligned across 26 languages. Following the original work, we adopt the token-based F1 score as the evaluation metric.

We conduct the following three experimental setups for the same questions to investigate the influence of Type-1 neurons: (a) Performance without any intervention. (b) Performance with deactivation in the top-1k Type-1 neurons. (c) Performance with deactivation in 1k randomly sampled neurons from the same layers as Type-1 neurons (baseline neurons). Answer generation was performed under a zero-shot setting.

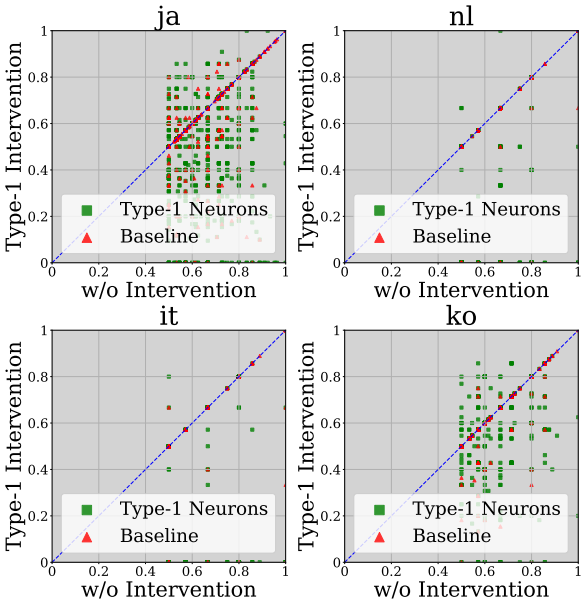


Figure 7: **Token-based F1 score when deactivating top-1k Type-1 transfer neurons (LLaMA3-8B).** Green denotes the score with intervention in the Type-1 neurons, while red represents the score with intervention in the randomly sampled neurons (i.e., baseline). Points below the $y = x$ line indicate a decrease in the score due to the intervention, whereas points above the $y = x$ line indicate an increase in the score. Points on the $y = x$ line denote no change in the score before and after the intervention.

MKQA (F1)	ja	nl	ko	it
(a) w/o Intervention	0.66	0.64	0.64	0.64
(b) Type-1 Δ	**-0.15**	**-0.41**	**-0.06**	**-0.49**
(c) Baseline Δ	-0.01	-0.03	-0.01	-0.04

Table 3: **Changes in token-based F1 scores for questions with original scores above 0.5 (LLaMA3-8B).**

Type-1 Transfer Neurons are Critical for Reasoning. Fig. 7 and Tab. 3 shows the intervention results for questions of each language where the model's generated answer, under setup (a), exceeds an F1 score of 0.5. This allows us to examine how interventions in neurons affect questions that the model is originally able to answer to some extent. As indicated, although the number of deactivated Type-1 neurons is very small (only 0.2% of all neurons), deactivating them causes a significant degradation, while deactivating baseline neurons

has almost no effect. These results suggest that Type-1 neurons play an essential role in eliciting an answer. This tendency was consistent across models. The results for other models and F1 threshold can be found in Appendix J.1.

To further validate the influence of deactivating Type-1 neurons, we conducted the same experiments on MMLU-ProX (Xuan et al., 2025), a multilingual benchmark designed to evaluate comprehensive reasoning ability. The results are highly consistent with the MKQA results reported above and are provided in Appendix J.2.

7 Discussion, Future Studies and Conclusion

Discussion 1: The Direction of Representational Transfer Suggests Language-Specificity of Transfer Neurons. We surmise that the unique direction of the representational shift is highly correlate with the language-specificity of the transfer neurons. That is, if the direction of the transfer (i.e., the direction of the target latent space) is language-specific, the neurons which strongly facilitate the representational shift towards the direction tend to be language-specific. This explains why most Type-1 neurons are less language-specific, while Japanese Type-1 neurons and Type-2 neurons are more likely to be language-specific, as discussed in §6.1 and Appendix H.4. In our hypothesis illustrated in Fig. 1, Type-1 neurons share the same target latent space (i.e., the English latent space). Consequently, as representations move closer to the English latent space towards the middle layers, languages exhibit greater overlap in Type-1 neurons as indicated in Figs. 8 and 34, revealing their language-agnostic nature.

Discussion 2: The Difference between Transfer Neurons and Language-Specific Neurons. As mentioned in §2.1, although language-specific neurons and regions were discovered in previous studies, their functional role was not clearly identified. In contrast, our work empirically proves (i) certain neurons facilitate the movement of hidden states between language latent spaces, and (ii) some of the transfer neurons we identify tend to be language-specific (e.g., most Type-2 neurons), while others tend not to be (e.g., most Type-1 neurons). Therefore, it is highly likely that some of the language-specific neurons discovered in previous studies correspond to the Type-2 neurons identified in our work. However, our analysis goes further by also identifying Type-1 neurons and non-language-specific Type-2 neurons, which were not captured in the previous studies. The reason why not all transfer neurons are fully language-specific, is that, as discussed in Discussion 1 above, the trajectories for representational transfer can be partly shared by languages especially those with linguistic and spacial proximity. This might be the rationale for language-family specificity of transfer neurons.

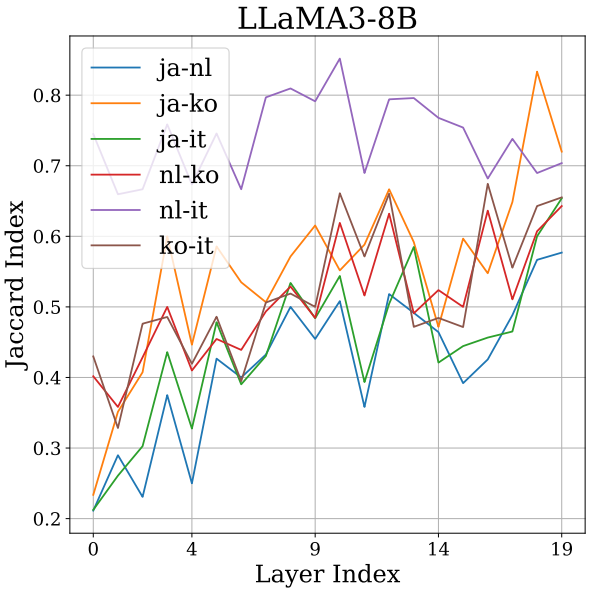


Figure 8: **Overlap ratio of Type-1 transfer neurons across language pairs and decoder layers (LLaMA3-8B)**. The higher the value on the vertical axis, the greater the overlap of neurons across languages.

Future Studies. We argue that transfer neurons play a central role in reasoning and cross-lingual transfer between high- and low-resource languages in multilingual LLMs. A practical implication is to exploit these neurons to enhance multilingual ability: after multilingual pretraining, selectively fine-tuning transfer neurons could better align hidden states of low-resource languages with the English-centric latent space where reasoning occurs. Such alignment may allow low-resource languages to leverage conceptual and structural representations learned from high-resource languages, thereby improving multilingual performance.

Conclusion. In this study, we proposed and empirically validated the Transfer Neurons Hypothesis. We provided evidence for language-specific latent spaces and a shared semantic latent space, identified transfer neurons highlighting their language and language family specificity, and showed that transfer neurons are essential for reasoning.

Limitations

Linear Transformation as Representational Transfer. Throughout the series of experiments, we hypothesize that neurons in the MLP module are responsible for transferring internal representations to and from a shared semantic latent space. As described in Eqs. 1 and 2, the MLP computation ultimately reduces to a weighted sum of activations and their corresponding value vectors in down projection matrix. Therefore, when considering only the MLP module, the transformation between latent spaces can be interpreted as a parallel shift (i.e., vector summation). This implies that our analysis is limited to linear transformations in the representational space. We did not consider other modules that might induce alternative types of vector movements, which may represent a limitation of this study.

Other Languages and Datasets. In this study, we used three LLMs and validated four languages for each model, including both those with linguistic proximity to English (i.e., Dutch and Italian) and those without such proximity (i.e., Japanese and Korean). As sentence datasets, we primarily employed tatoeba (Tiedemann, 2020) and MKQA (Longpre et al., 2021) for neuron detection and validation. Although we believe the settings demonstrate the generalizability of our findings well, other types of languages and datasets (e.g., those involving longer input sentences) warrant further investigation.

Acknowledgments

This work was supported by The Nakajima Foundation. The authors would like to express their gratitude to Mr. Hakaze Cho, Mr. Kenshiro Tanaka, and Mr. Koga Kobayashi of the Japan Advanced Institute of Science and Technology, for providing valuable comments for this work.

References

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *AAAI Conference on Artificial Intelligence*.

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.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv Preprint*, arXiv:2412.04261.

Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhenguang Cai. 2025. Unveiling language competence neurons: A psycholinguistic approach to model interpretability. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10148–10157, Abu Dhabi, UAE. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. 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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv Preprint*, arXiv:2407.21783.

Tatsuya Hiraoka and Kentaro Inui. 2025. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 483–495, Albuquerque, New Mexico. Association for Computational Linguistics.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hy-

pothesis. In *International Conference on Machine Learning*.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2025. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv Preprint*, arXiv:2409.17892.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv Preprint*, arXiv:2310.06825.

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

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. 2021. Pay attention to mlps. In *Advances in Neural Information Processing Systems*, volume 34, pages 9204–9215. Curran Associates, Inc.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

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.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv Preprint*, arXiv:2502.15603.

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

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2022. Self-conditioning pre-trained language models. *International Conference on Machine Learning*.

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

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, Wei Peng, and Alexandra Birch. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv Preprint*, arXiv:2406.09265.

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.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv Preprint*, arXiv:2503.10497.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st*

International Conference on Computational Linguistics, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024a. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102, St. Julian's, Malta. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Experimental Settings

Models. Throughout the experiments in this study, we used LLaMA3-8B(Grattafiori et al., 2024), Mistral-7B(Jiang et al., 2023), and Aya expanse-8B(Dang et al., 2024).

Datasets. For the similarity measurements in §5, we used the Tatoeba parallel corpus (Tiedemann, 2020). This dataset was also used to investigate the language and language-family specificity of transfer neurons in §6, by randomly sampling sentences for each language.

For the QA task and the dimensionality reduction presented in §3.1 and Appendix F, we used the MKQA dataset (Longpre et al., 2021).

For similarity measurements of hidden states and activation patterns, computation of centroids of each latent spaces, identifying transfer neurons, applying PCA to hidden representations, and investigating language specificity of transfer neurons, we sampled 1k sentences per language. Finally, for SVD experiments described in §3.2, we sampled 1k hidden states per layer and per language using MKQA sentences.

Train-Test Split for Identifying and Controlling Transfer Neurons. We split the sentence data into a 50:50 ratio (train:test), using 1k samples for each split (2k samples in total) to identify and deactivate transfer neurons. In other words, the similarity re-measurements described in §5 were conducted using sentences different from those used to identify the transfer neurons, in order to prove their generalization performance. We believe that this setting further strengthens the reliability of our findings.

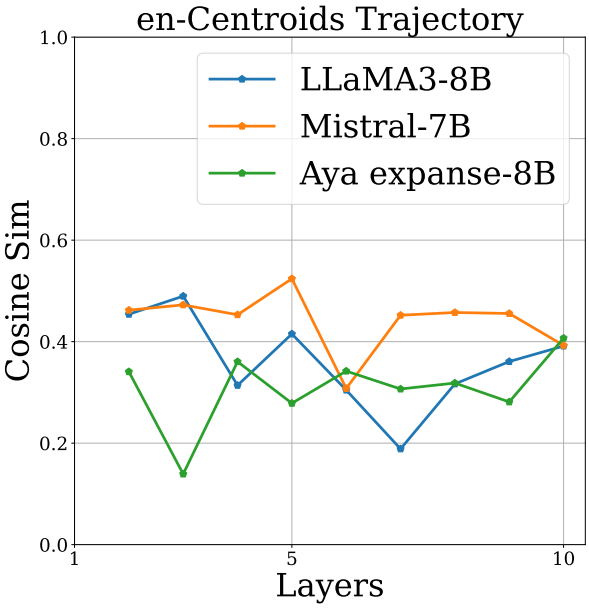


Figure 9: **Similarity between layer-wise trajectories and linear path from decoder layer 1 to the middle layers for the centroid of the English latent space**.

Representations. Throughout the experiments, we extracted representations corresponding to the final token of each input, reflecting our focus on sentence-level rather than word-level semantic processing.

Computational Resources. We conducted the experiments using NVIDIA A40 (48 GB), A100 (40 GB), and H200 (141 GB) GPUs. However, all experiments are reproducible on a single A100, and the majority can be executed with a single A40.

B Trajectory of the English Latent Space Moving towards Middle Layers

Computational Settings. Does English latent space move towards the middle layers in a linear or non-linear manner? To answer this question, we computed the similarity between a linear trajectory and actual layer-wise trajectory of centroid of the English latent space across layers.

If the similarity is high, it indicates that the layer-wise trajectory closely follows a linear path, suggesting that the representations move almost

linearly towards a target convergence point in the middle layers, where semantic processing and reasoning primarily occur. On the other hand, if the similarity is relatively low, it indicates that the representations move non-linearly.

Computations are as follows:

$$\boldsymbol{P}_{\text{en}} = \boldsymbol{C}^{m}_{\text{en}} - \boldsymbol{C}^{1}_{\text{en}} \tag{10}$$

$$\boldsymbol{T}^{l}_{\text{en}} = \boldsymbol{C}^{l}_{\text{en}} - \boldsymbol{C}^{l-1}_{\text{en}}, \quad l \in \{2, \ldots, m\} \tag{11}$$

$$\text{Sim\_score}^{l} = \cos(\boldsymbol{T}^{l}_{\text{en}}, \boldsymbol{P}_{\text{en}}) \tag{12}$$

where $\boldsymbol{P}_{\text{en}}$ denotes a vector expressing the linear path of the centroid of the English latent space from decoder layer 1 to layer m (the middle layer). It is defined as the difference vector between the centroid at layer m, $\boldsymbol{C}^{m}_{\text{en}}$, and the centroid at layer 1, $\boldsymbol{C}^{1}_{\text{en}}$. A vector $\boldsymbol{T}^{l}_{\text{en}}$ represents the actual trajectory step of the centroid of the English latent space between layer $l-1$ and layer l. As we mentioned, the higher the similarity score (Sim_score^{l}), the more **linearly** the centroid moves in hidden state space from layer $l-1$ to layer l, indicating that the internal representations move linearly towards the target convergence point in the middle layers. To the contrary, lower scores across layers indicates that the representations move **non-linearly**.

Based on the observations from hidden states and activation patterns similarity (Figs. 10 and 11), kernel-based similarity (Fig. 4 in §3.3 and Fig. 19 in Appendix E), and PCA visualization of language latent spaces (Figs. 16 and 17), we set m to 10 representing middle layer where each language latent spaces aligns well.

English Latent Space Transitions Non-Linearly towards Middle Layers. Fig. 9 present the results. As indicated, English latent space move non-linearly, and this tendency is highly consistent with all the models adopted in this paper.

Based on this observation, we adopt the layer-wise centroid, $\boldsymbol{C}^{l}$, as the target centroid to detect transfer neurons with Eqs. 7 and 8 in §4.3. This is because using the centroid of a specific layer instead of layer-wise centroids could significantly hinder the effective identification of neurons that contribute to the shift of representations towards the moving English latent space, given that English latent space moves non-linearly.

C Similarity and Linear Separability of Internal Representations

C.1 Similarity of Hidden States and Activation Patterns for Parallel Sentences across Languages

Similar Hidden States for Similar Semantics in Middle Layers. To quantitatively investigate the similarity of hidden states across languages, given $\{(\boldsymbol{s}^{l}_{i}, \boldsymbol{t}^{l}_{i})\}^{N}_{i=1}$, a set of l-th layer hidden states for parallel sentences, and $\{(\boldsymbol{u}^{l}_{i}, \boldsymbol{v}^{l}_{i})\}^{N}_{i=1}$, the ones for *non*-parallel sentences, we show the difference between $\frac{1}{N}\sum_{i}\cos(\boldsymbol{s}^{l}_{i}, \boldsymbol{t}^{l}_{i})$ and $\frac{1}{N}\sum_{i}\cos(\boldsymbol{u}^{l}_{i}, \boldsymbol{v}^{l}_{i})$ at each layer in Fig. 10. The $\boldsymbol{u}_{i}$ are fixed to the set of hidden states for English sentences. Clearly, models process similarly for parallel sentence pairs compared to non-parallel ones, especially in the middle layers. The relatively pronounced divergence in similarity between parallel and non-parallel ones observed in the middle layers suggests that, at these stages, the inputs are processed within a shared semantic latent space, wherein semantically similar sentences are more likely to be mapped to proximate locations. On the other hand, the relatively small divergence observed in the initial and final layers might be attributed to the model operating within language-specific representational latent spaces at these stages, where the primary focus is on processing the input and generating the output in the specific language. We found that this tendencies are consistent across other pair patterns and models.

Similar Activations for Similar Semantics in Middle Layers. Interpreting the MLP as key-value memory, activation values reflect how strongly the model accesses value vectors, which encode concepts across languages (Geva et al., 2021, 2022; Dai et al., 2022; Chen et al., 2024). From this perspective, relatively high similarity in activation patterns for parallel sentences compared to non-parallel ones suggests the model encodes similar concepts across languages, supporting the existence of a shared semantic latent space.

Fig. 11 show the similarity of the activation patterns across layers with the same inputs as similarity measurement of hidden states, which is calculated as $\frac{1}{N}\sum_{i}\cos(\boldsymbol{\alpha}^{l}_{\text{en},i}, \boldsymbol{\alpha}^{l}_{\text{L2},i})$ where $\boldsymbol{\alpha}^{l}$ denotes the activation values vector in l-th layer described in Eq. 2, and n represents the number of sample sentence pairs. L2 denotes the languages other than English. Similar to the results for hidden states sim-

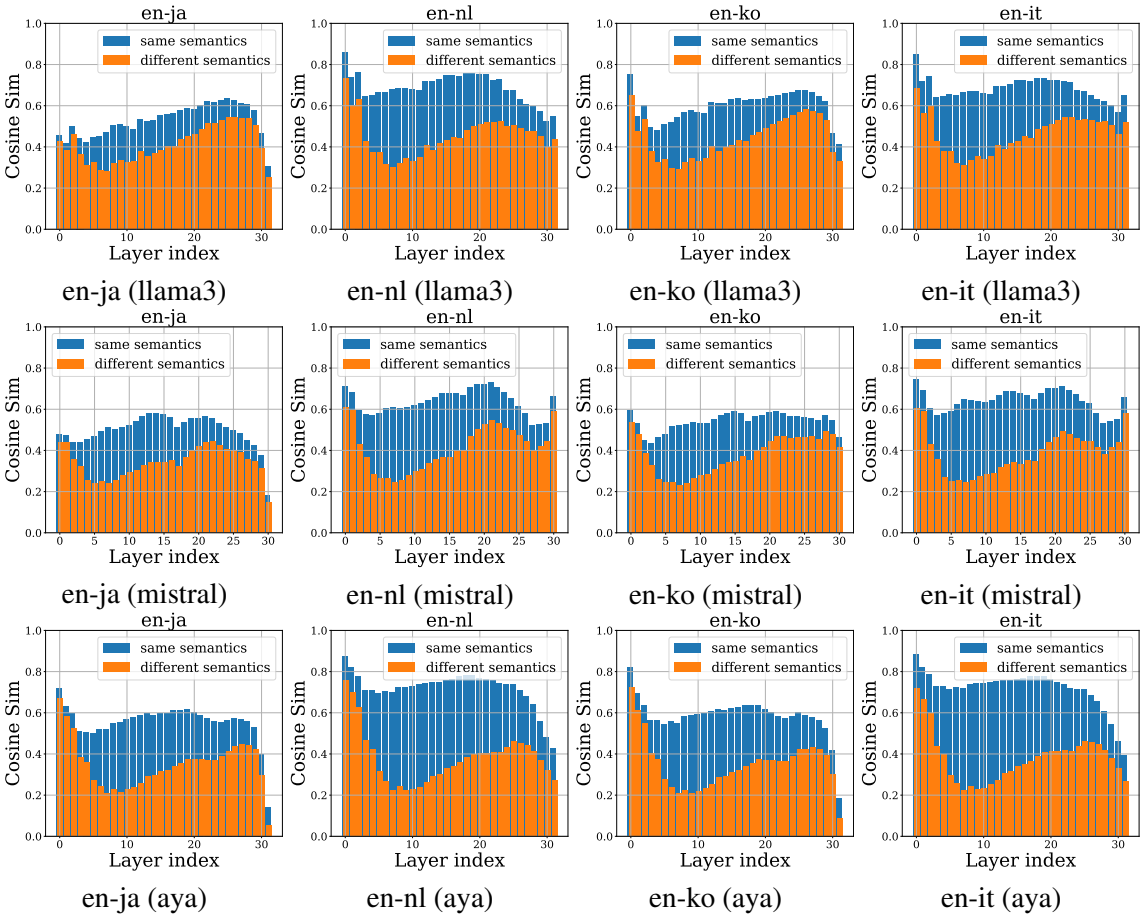


Figure 10: **Similarity of hidden states across layers.**

ilarity, the difference in similarity between parallel and non-parallel inputs is particularly prominent in the middle layers, supporting the existence of a shared semantic latent space. This tendency is consistent across other language pairs and models, and it also aligns with the findings of Zeng et al. (2025), despite differences in experimental settings.

C.2 Investigating Linear Separability between Parallel and Non-Parallel Inner Representations with Logistic Regression Model

To investigate linear separability between parallel and non-parallel pairs of representations, we trained a logistic regression model on the hidden representations from each layer. For each sentence pair (parallel or non-parallel), the input features were constructed by concatenating their hidden states. Parallel pairs were labeled as `1`, and non-parallel pairs as `0` (1000 samples for `label 1` and `label 0`, respectively). We employed stratified 10-fold cross-validation to ensure balanced evaluation across classes. As shown in Fig. 12, except for the initial few layers, most layers achieve an test accuracy of approximately over 70%. Moreover, we observe that layers with higher similarity between hidden language latent spaces — where the shared semantic latent space appears to exist, as shown in Figs. 4, 18, and 19 — tend to yield higher classification accuracy. This result suggests that the model can effectively classify whether a pair of input sentences in two languages is parallel or non-parallel in the middle layers. This results also support the existence of a shared semantic latent space in middle layers.

D Latent Space Property of Hidden States

D.1 Dimensionality of Latent Spaces Across Layers

Fig. 13 shows the estimated dimensionality (i.e, the number of orthonormal basis) of the latent space across languages and layers. As shown, each language latent spaces has lower dimensionality than that of hidden states, indicating they are latent spaces in the hidden state space.

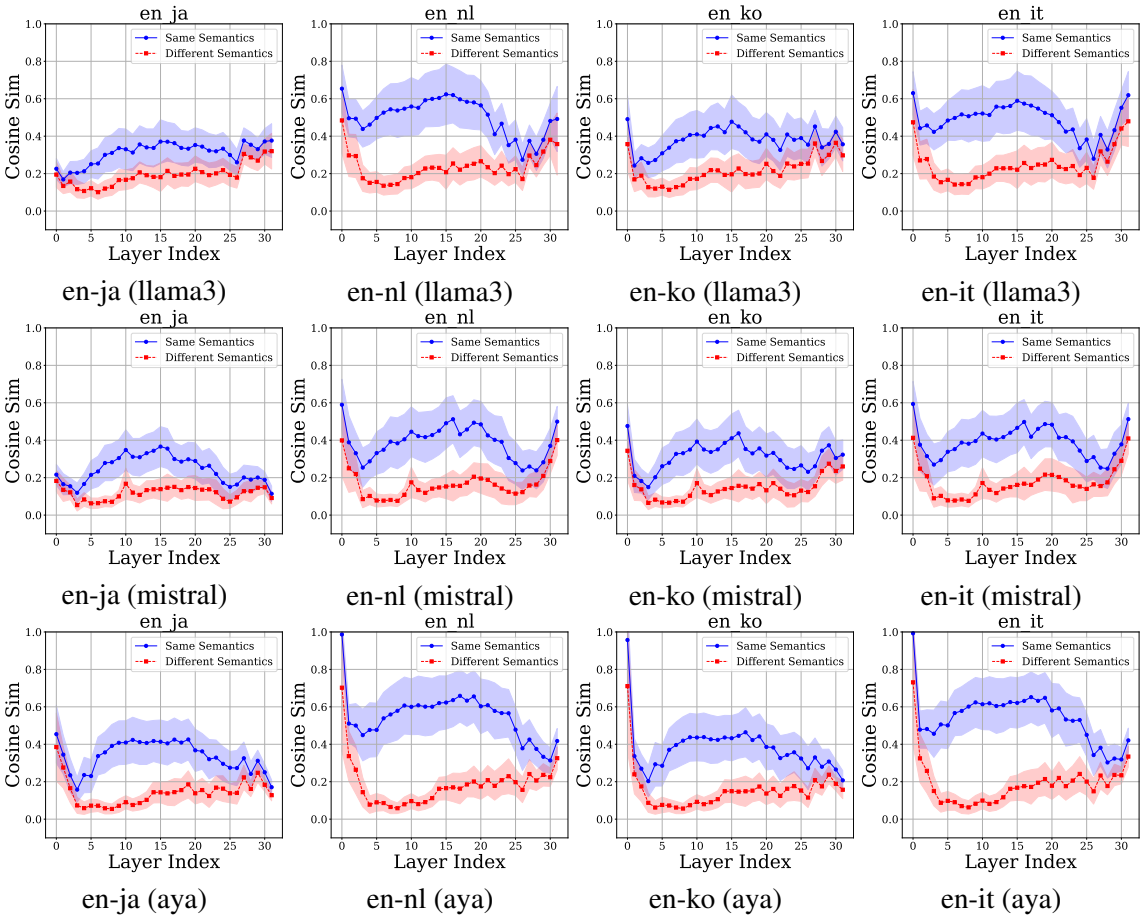


Figure 11: **Similarity of activation patterns in MLP module across layers.**

D.2 Distance among Language Latent Spaces

To quantitatively demonstrate the spacial transition phenomena and measure distance among language latent spaces, we compute $\cos(\boldsymbol{C}^l_{\mathrm{L}_1}, \boldsymbol{C}^l_{\mathrm{L}_2})$, where both $\boldsymbol{C}^l_{\mathrm{L}_1}$ and $\boldsymbol{C}^l_{\mathrm{L}_2}$ are the centroid of each latent space computed by Eq. 5.

Figs. 14 and 15 show the results. As indicated, these results are well align with the hypothesis shown in Fig. 1 (a): in the initial and final few layers, language latent spaces remain relatively distant from each other, while they become closer in the middle layers where language-agnostic semantic processing and reasoning are mainly occurred.

E Mutual *k*-Nearest Neighbor Alignment Metric

E.1 Formalization of the Computation for *Mutual k-NN Alignment Metric* across Layers and Input Language Pairs

This metric was originally for computing the similarity between representations formed by semantically equivalent inputs from different modalities, each residing in separate model spaces. In our setting, however, we adapt it to compare hidden latent spaces formed by different input languages within a single model, allowing us to assess kernel-based layer-wise similarity between language-specific latent spaces (English–L2). Building on Huh et al. (2024), we adjust the computation as follows:

Let f be a single model (LLM), and let x_i and y_i denote an English sentence and its corresponding sentence in another language (L2), respectively, both expressing the same meaning.

$$\phi^l_i = f(x^l_i) \tag{13}$$

$$\psi^l_i = f(y^l_i) \tag{14}$$

where ϕ^l_i, ψ^l_i are hidden states in l-th layer. Collection of these representations are denoted as:

$$\Phi^l = \{\phi^l_1, ..., \phi^l_b\} \tag{15}$$

$$\Psi^l = \{\psi^l_1, ..., \psi^l_b\} \tag{16}$$

Then for each representations pair $\{\phi^l_i, \psi^l_i\}$, we compute the respective nearest neighbor sets $\mathcal{S}(\phi^l_i)$

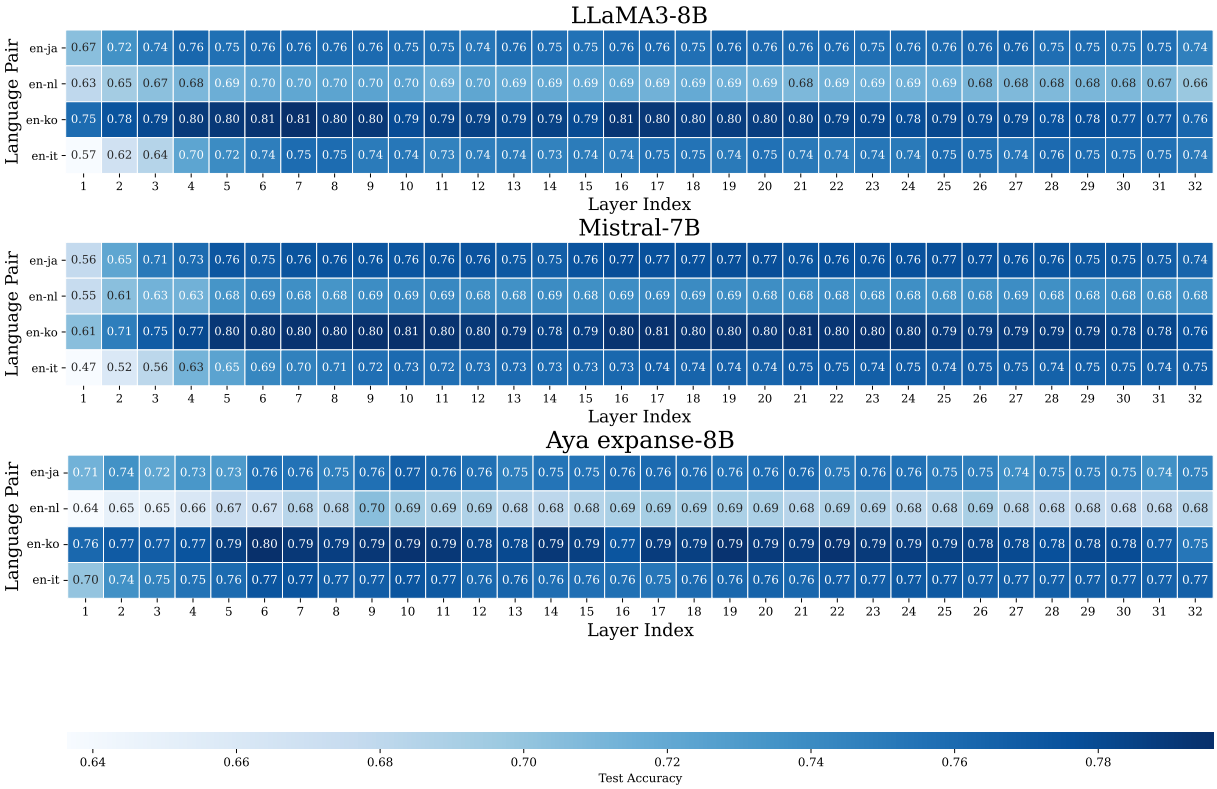


Figure 12: **Test accuracy of logistic regression model trained on parallel semantic features and non-parallel semantic features for each layer.** The y-axis indicates the language pairs, and the x-axis denotes the layer indices.

and $\mathcal{S}(\psi_i^l)$:

$$d^l_{\text{knn}}(\phi_i^l, \Phi^l \backslash \phi_i^l) = \mathcal{S}(\phi_i^l) \quad (17)$$

$$d^l_{\text{knn}}(\psi_i^l, \Psi^l \backslash \psi_i^l) = \mathcal{S}(\psi_i^l) \quad (18)$$

where d^l_{knn} returns the set of indices of its k-nearest neighbors. Finally, we measure its average intersection via:

$$m^l_{\text{nn}}(\phi_i^l, \psi_i^l) = \frac{1}{k} |\mathcal{S}(\phi_i^l) \cap \mathcal{S}(\psi_i^l)| \quad (19)$$

where $|\cdot|$ is the size of intersection.

E.2 Result of Mistral-7B

The left column of Figs. 19 and 18 show the result of computing the mutual k-NN alignment metric across layers of the models including Mistra-7B. As shown, we obtain results of Mistral-7B similar to those of LLaMA3-8B and Aya-expanse-8B (Fig. 4). Additionally, inputs from languages that belong to language families relatively close to English (Dutch, Italian) tend to form more similar latent spaces compared to those that do not.

F Visualization of Language Latent Spaces

Figs. 16 and 17 present the results of PCA applied to the hidden language representations across layers of models.

G Detecting and Controlling Transfer Neurons

G.1 The Definition of Neurons in This Study

While several studies (Suau et al., 2022; Tang et al., 2024; Wang et al., 2024; Hiraoka and Inui, 2025; Mondal et al., 2025) have investigated neuron-level analysis or identification in LLMs, these studies primarily regard the output of the non-linear activation function as activation values and treat them as the fundamental unit of neurons. Throughout the experiments in this study, however, we defined neurons as the activation values α_i^l denoted in Eqs. 1 and 2 (i.e., the output of element-wise product). We adopted this definition because, even when the output of the activation function is large, the actual activation value α_i^l can be small depending on the output of the up projection, given that the models we adopted has a gated-MLP. As a result, the impact on subsequent neurons may be attenuated, which renders α_i^l a more appropriate indicator of neuron activity in the context of our analysis.

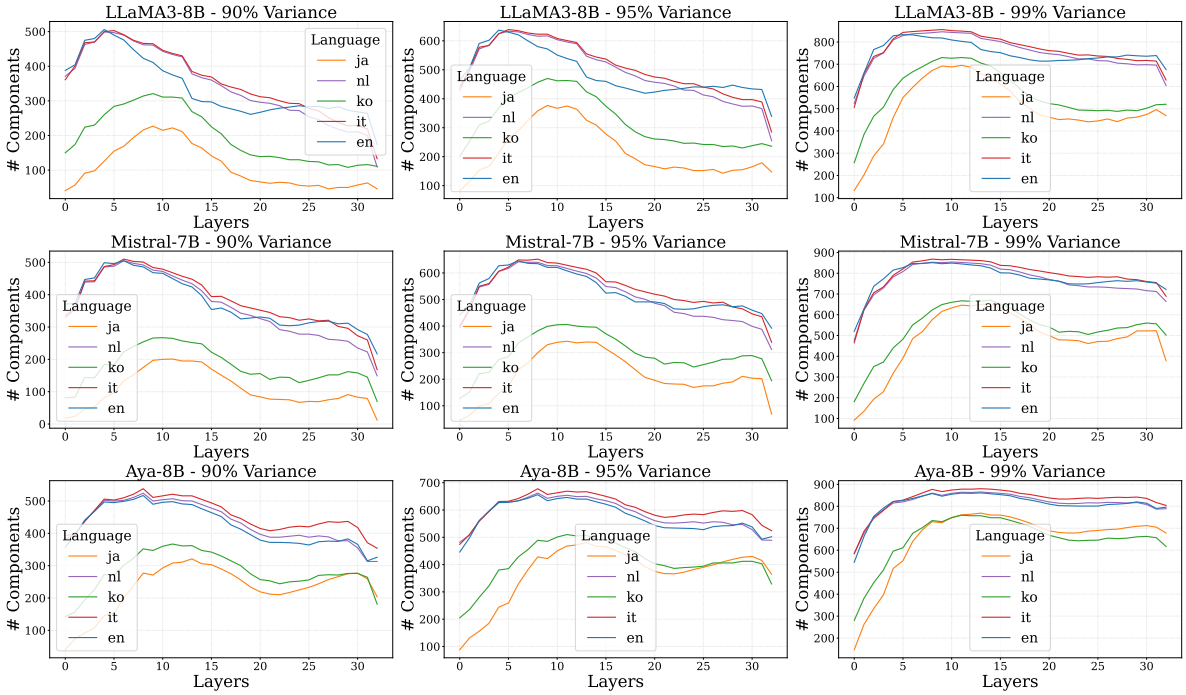


Figure 13: **Estimated dimensionality of latent spaces across layers (LLaMA3-8B, Mistral-7B, and Aya expanse-8B).** The vertical-axis represents the number of singular values (i.e., the number of orthonormal basis) required to explain 90, 95, and 99% of the variance. It is computed via SVD and cumulative explained variance ratio.

G.2 Deactivating Type-1 Transfer Neurons

G.2.1 Kernel-Based Similarity between Latent Spaces

The middle columns of Figs. 18 and 19 presents the results of kernel-based similarity between English and L2 latent spaces while deactivating top-1k Type-1 neurons, computed using the Mutual k-NN Alignment Metric formalized in Appendix E.1. As shown, deactivating the top-1k Type-1 neurons leads to a substantial drop in similarity, particularly in the middle layers where similarity originally peaked — suggesting the presence of a shared semantic latent space. This also suggests that the Type-1 neurons identified in §5 play a pivotal role in shifting representations towards this shared latent space. As explained in §5.2, we surmise that the relatively high remaining similarity in the en-nl (Dutch) and en-it (Italian) settings — even after deactivating Type-1 neurons — suggests that their specific latent spaces are already close to each other in even early layers, as shown in the PCA results (§3.1).

G.2.2 Hidden States and Activation Patterns Similarity

Figs. 20, 21, and 22 illustrate the similarity of hidden states across all models when the top 1k, 3k, and 5k Type-1 neurons are deactivated. Similarly, Figs. 23, 24, and 25 present the corresponding similarities in activation patterns.

G.2.3 Quantitative Distance among Language Latent Spaces

Figs. 26 and 27 show the distance among centroids of language latent spaces with Type-1 neurons deactivated. As shown, deactivation of Type-1 neurons hinder the movements towards English latent space significantly, compared to those of deactivation of randomly sampled neurons (baseline). This tendency is consistent across languages and models.

G.3 Deactivating Type-2 Transfer Neurons

G.3.1 Deactivating Type-2 Transfer Neurons Significantly Inhibit Spatial Transition from the Shared Semantic Latent Space to the Language-Specific Latent Spaces

Fig. 28 and 29 show the results of PCA applied to hidden representations specific to each language, while deactivating top-1k Type-2 neurons (English features are the only ones visualized without intervention). As shown, although only 0.2% of all neurons in the model were deactivated, the deactivation caused a significant delay in the movement of representations towards each language-specific latent space — especially in the final layers where most Type-2 neurons reside — compared to the

baseline (i.e., deactivating the 1k neurons randomly sampled from the same layers as the Type-2 neurons). These results support the functional impact of the detected Type-2 neurons.

G.3.2 Quantitative Distance among Language Latent Spaces

Figs. 30 and 31 show the distance among centroids of language latent spaces while deactivating Type-2 neurons. As indicated, deactivating Type-2 neurons significantly inhibit the movements towards each language specific latent space in final layers. This observation is well aligned with the PCA visualization while deactivating top-1k Type-2 neurons, as presented in Appendix G.3.1 above.

H The Nature of Transfer Neurons

H.1 Distribution

Figs. 32 and 33 show the distribution of Type-1 and Type-2 neurons.

H.2 Overlap Ratio of Transfer Neurons

Type-1 Transfer Neurons. Fig. 34 shows the overlap ratio (i.e., Jaccard Index) for Type-1 neurons across language pairs for Mistral-7B and Aya expanse-8B. As indicated, their overlap ratio increase as language representations move closer to each other towards the shared latent space in middle layers.

Type-2 Transfer Neurons. Fig. 35 shows the overlap ratio (i.e., Jaccard Index) for Type-2 neurons across language pairs for all the models. As shown, their overlap ratio decrease as language representations move to each language-specific latent space for output generation.

H.3 Detecting Language-Specific Neurons

Building on the method proposed by Kojima et al. (2024), we independently identify language-specific neurons. Kojima et al. (2024) detected such neurons by computing the *Average Precision* (i.e., *Area Under the Precision-Recall Curve*) between each neuron's activation — defined as the output of the non-linear activation function in the MLP module — and language-labeled sentences[7]. Sentences in the target language were labeled as `1`, and those in other languages were labeled as `0`. They defined both the top-1k and bottom-1k (2k neurons in total) neurons in the score ranking as language-specific neurons, considering not only those positively correlated with the target language but also those negatively correlated.

However, their method is ambiguous regarding how many neurons should be selected from the top and bottom rankings. Therefore, in our approach, we adopt the *Correlation Ratio* (Appendix H.6) instead of Average Precision as a metric, as it allows us to capture both positive and negative correlations simultaneously, with a single explicit value for each neuron that reflects the strength of the correlation.

Let S be the set of all sentences across all languages, i.e., the union of English, Japanese, Dutch, Korean, and Italian sentence sets:

$$S = \{S_{\text{en}}, S_{\text{ja}}, S_{\text{nl}}, S_{\text{ko}}, S_{\text{it}}\} \tag{20}$$

Each language contains 1k sentence samples (5k in total). For instance, to detect Japanese-specific neurons, we assign `label1` to all 1k Japanese sentences and `label0` to sentences from all other languages. Then, for each neuron, we compute the correlation ratio between its activation values and the labels of the 5k sentence samples.

Figs. 36, 37, and 38 show the distribution of detected language-specific neurons. As shown, the stronger the correlation between a neuron's activations and the target language labels, the more it tends to be located in the initial and final layers. This tendency is well aligned with the findings of previous studies. Also, as we described in §5.1, their distribution is similar to those of transfer neurons.

H.4 Language Specificity

Tabs. 4, 5, and 6 present the correlation ratios of transfer neurons with respect to language specificity across all models. The computation settings follow those described in §6.1.

H.5 Language-Family Specificity

H.5.1 Jaccard Index across Language Pairs

Tabs. 7, 8, and 9 show the Jaccard Index of transfer neurons for each language pair, representing the degree of overlap in neurons between pairs.

H.5.2 Correlation Ratio for Language-Family Specificity

To investigate language-family specificity from another perspective of view, we computed the correlation ratio by assigning `label1` to sentence pairs

[7]This method was originally proposed by Suau et al. (2022).

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.16	0.03	0.05	0.03	Type-1 TN	0.16	0.02	0.05	0.03
Type-2 TN	0.25	0.22	0.17	0.16	Type-2 TN	0.40	0.27	0.33	0.19
Top-10	**ja**	**nl**	**ko**	**it**					
Type-1 TN	0.38	0.01	0.01	0.03					
Type-2 TN	0.54	0.32	0.76	0.30					

Table 4: **Correlation ratio of top-10, top-100, and top-1k Transfer Neurons for language specificity (LLaMA3-8B).** Typically, a correlation ratio above **0.1** indicates a correlation, above **0.25** indicates a moderately strong correlation, and above **0.5** indicates a strong correlation.

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.17	0.03	0.06	0.03	Type-1 TN	0.18	0.03	0.08	0.03
Type-2 TN	0.28	0.22	0.11	0.18	Type-2 TN	0.50	0.27	0.21	0.24
Top-10	**ja**	**nl**	**ko**	**it**					
Type-1 TN	0.37	0.02	0.07	0.03					
Type-2 TN	0.80	0.43	0.52	0.41					

Table 5: **Correlation ratio of top-10, top-100, and top-1k Transfer Neurons for language specificity (Mistral-7B).**

that largely belong to the same language family, and `label0` to all others, following the settings described in Appendix H.3. In this experiment, we consider English, Dutch, and Italian as belonging to the same language-family, whereas Japanese and Korean are grouped into another language-family. Tabs. 10, 11, and 12 present the results. Language-family specificity was particularly evident in Type-2 neurons.

H.6 Correlation Ratio

Correlation ratio is a well established statistical measure of association between categorical and quantitative variables. It is computed as follows:

$$\text{corr ratio}(\eta^2) = \frac{S_B}{S_W + S_B} \tag{21}$$

where S_B (between-group sum of squares) measures the variance between label means, and S_W (within-group sum of squares) measures the variance within each label.

That is, in our case, neurons that tend to exhibit similar activation patterns for sentences within the same label (i.e., language or language-family), and distinct patterns for sentences with different labels, receive higher correlation ratio scores.

H.7 Hypothesis Testing for Reported Correlation Values

To verify the reliability of the correlation ratio scores reported in the series of experiments in this study, we perform hypothesis testing using ***An**alysis **o**f **Va**riance* (*ANOVA*). ANOVA is a statistical method used to determine whether there are significant differences between the means of multiple groups.

That is, in our context, we test whether there are significant differences in the activation values of each transfer neuron depending on the label (i.e., language or language family). Note that each transfer neuron has *p value*.

We present proportions for both Type-1 and Type-2 neurons whose activations meet following conditions: (i) statistically significant ($p < 0.05$) and (ii) correlation ratio exceeds 0.1 (meaningful correlation) and 0.25 (moderately strong correlation). The results for language specific correlation are shown in Tabs. 13, 14, 15, and 16. As presented, the proportions are well aligned with the results of correlation analysis in this paper, which shows Type-1 neurons generally do not exhibit correlation (except for Japanese ones), whereas most Type-2 neurons exhibit correlation with input languages.

We also present the results for the same testing method for the correlation with language-family specificity in Tabs. 17, 18, 19, and 20.

Additionally, as shown in Fig. 39, it turned out that approximately 80-95% of top-1k transfer neurons are statistically significant, which further strengthen our correlation analysis.

Furthermore, we conducted a *Mann–Whitney U test* as a non-parametric test under the same settings described above. While ANOVA relies on

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.19	0.02	0.05	0.03	Type-1 TN	0.32	0.02	0.03	0.02
Type-2 TN	0.24	0.22	0.17	0.24	Type-2 TN	0.43	0.37	0.40	0.36

Top-10	ja	nl	ko	it
Type-1 TN	0.27	0.03	0.03	0.02
Type-2 TN	0.80	0.50	0.70	0.33

Table 6: **Correlation ratio of top-10, top-100, and top-1k Transfer Neurons for language specificity (Aya expanse-8B).**

	ALL	ja-nl	ja-ko	nl-it	ja-it	nl-ko	it-ko
Type-1 TN	0.30	0.39	0.51	0.75	0.41	0.50	0.51
Type-2 TN	0.12	0.23	0.37	0.51	0.26	0.28	0.30

Table 7: **Jaccard index for top-1k Transfer Neurons across language pairs. (LLaMA3-8B).**

the *Central Limit Theorem*, assuming that, in our case, the distribution of the mean activation values of transfer neurons converges to a Gaussian distribution. Although it is likely that the mean activations follow the theorem, we also performed the non-parametric Mann–Whitney U test to validate our findings since activation values are outputs of non-linear functions. As a result, it yielded results that were highly consistent with those of ANOVA, further enhancing the reliability of our analysis.

I The Effects of Transfer Neurons on Internal Representations under Cross-Lingual Deactivation

Experimental Setup. To analyze cross-lingual effects of transfer neurons in hidden state space, we compute the centroids of L2 hidden states (L2 latent space) at each layer: (i) $\boldsymbol{C}^{l}_{\text{L2}}$: without any deactivation, (ii) $\boldsymbol{C}^{l-\text{L1TN}}_{\text{L2}}$: with L1 transfer neurons deactivated, and (iii) $\boldsymbol{C}^{l-\text{L2TN}}_{\text{L2}}$: with L2 transfer neurons deactivated. We then compute $\cos(\boldsymbol{C}^{l}_{\text{L2}}, \boldsymbol{C}^{l-\text{L1TN}}_{\text{L2}})$ to assess the effect of cross-lingual deactivation on both types of neurons, using $\cos(\boldsymbol{C}^{l}_{\text{L2}}, \boldsymbol{C}^{l-\text{L2TN}}_{\text{L2}})$ as a reference point.

Results. We report the results for final layer of each type of neurons (i.e., 20th layer for Type-1 neurons and final layer for Type-2 neurons) for LLaMA3-8B, Mistral-7B, and Aya expanse-8B across all the languages in Tabs. 21, 22, and 23. Deactivating Type-1 neurons for L1 had a similar impact as the L2 deactivation baseline, indicating that these neurons have a noticeable effect on the hidden state dynamics of L2 inputs. This aligns with our finding in §6.1 that Type-1 neurons are less language-specific. On the other hand, deactivating Type-2 neurons for L1 had a much smaller impact compared to the L2 deactivation baseline, indicating that these neurons have little to no influence on the hidden state dynamics of L2 inputs. A limited degree of influence was observed in cases where the target latent spaces of L1 and L2 were spatially close (e.g., L1 = Dutch and L2 = Italian). This is consistent with our finding in §6.1 that Type-2 neurons are more language-specific.

J Downstream Reasoning Tasks while Deactivating Transfer Neurons

J.1 Multilingual Knowledge QA (MKQA)

Figs. 40, 41, and 42 and Tabs 24, 25, 26, 27, and 28 show all the results for MKQA task described in §6.2. As shown, the results are highly consistent with those reported in §6.2.

J.2 MMLU-ProX

Tabs. 29, 30, 31, 32, 33, 34, 35, 36, and 37 present the results of MMLU-ProX under the same settings as MKQA task, as described in §6.2 and Appendix J.1. These are the results for Japanese, Korean and French. While our original experiments used Japanese, Korean, Italian, and Dutch, the evaluation tool used in the original paper (Xuan et al., 2025)[8] have limited language coverage. Therefore, we report results for Japanese and Korean, and additionally include French as a linguistically closer language to English such as Italian and Dutch. As shown, deactivating top-1k Type-1 neurons significantly degrade performance, whereas deactivating a baseline set of 1k randomly sampled neurons (from the same layers as the Type-1 neurons) results in negligible change compared to the orig-

[8] https://github.com/EleutherAI/lm-evaluation-harness

	ALL	ja-nl	ja-ko	nl-it	ja-it	nl-ko	it-ko
Type-1 TN	0.30	0.42	0.48	0.75	0.43	0.54	0.52
Type-2 TN	0.09	0.18	0.31	0.42	0.19	0.26	0.29

Table 8: **Jaccard index for top-1k Transfer Neurons across language pairs. (Mistral-7B).**

	ALL	ja-nl	ja-ko	nl-it	ja-it	nl-ko	it-ko
Type-1 TN	0.30	0.42	0.46	0.75	0.40	0.55	0.53
Type-2 TN	0.08	0.18	0.31	0.35	0.18	0.22	0.21

Table 9: **Jaccard index for top-1k Transfer Neurons across language pairs. (Aya expanse-8B).**

inal states (w/o intervention). These results are consistent with the results of MKQA task, further highlighting the critical role of Type-1 neurons in reasoning.

J.3 Open-Ended Text Generations while Deactivating Transfer Neurons

To evaluate the models' generation quality while nullifying the influence of top-1k transfer neurons during the generation process, we use PolyWrite (Ji et al., 2025), a multilingual open-ended generation task with no single correct answer. The following are some example questions in English.

- `Describe a day on Earth where gravity suddenly reverses for a few hours. How do people and animals react?`
- `Create a character who was once a hero but has become a villain due to a tragic misunderstanding.`
- `Write an email to your supervisor asking for feedback on a recent report you submitted. Express your interest in improving your work and ask for specific suggestions.`

We manually evaluate the quality of 50 generated texts for each language under the following settings: (a) without any intervention, (b) with top-1k Type-1 neurons intervention, and (c) with top-1k Type-2 neurons intervention. The evaluation, conducted by the authors, focus on how interventions in settings (b) and (c) alter the generated outputs compared to setting (a)[9]. We conduct generations under greedy decoding to compare outputs before and after the intervention.

[9]The authors are not proficient in Dutch, Korean, and Italian; However, we still made an effort to assess outputs of these languages by using machine translation.

Top-1k Type-1 Neurons Deactivated.

- LLaMA3-8B: Generated gibberish, suggesting that the model is no longer capable of reasoning or producing coherent output. This effect was consistent across languages.
- Mistral-7B: Fluency degrades, but not to full gibberish. In Italian, repetition spikes markedly and unique sentence variety drops, indicating looping and reduced coherence. In Dutch, outputs shorten with more line breaks and fewer unique sentences, suggesting fragmenting/early truncation. In Japanese and Korean, mild increase in repetition with overall structure mostly intact.
- Aya expanse-8B: Generally stable and coherent across languages with mild regression. Italian and Dutch keep length and repetition near normal but lose sentence variety. Japanese shows more line breaks and higher repetition, less fluent. Korean is close to normal with a small drop in variety.

Top-1k Type-2 Neurons Deactivated.

- LLaMA3-8B: Although it produced gibberish for many samples (similar to when Type-1 neurons were deactivated), in some cases it generated coherent text in other languages (e.g., English), different from the input language, yet still aligned with the question context. We show an example in Tab. 38. This phenomena support the fact that Type-2 neurons critically move internal reasoned representations to the latent space of output language in model space.
- Mistral-7B: A moderate decline in fluency is observed. In Italian, this manifests as increased looping and reduced sentence variety, while in Dutch the generations tend to be

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.14	0.08	0.11	0.08	Type-1 TN	0.15	0.11	0.12	0.11
Type-2 TN	0.25	0.23	0.22	0.23	Type-2 TN	0.31	0.25	0.27	0.25
Top-10	**ja**	**nl**	**ko**	**it**					
Type-1 TN	0.25	0.10	0.04	0.13					
Type-2 TN	0.40	0.45	0.38	0.43					

Table 10: **Correlation ratio of top-10, top-100, and top-1k Transfer Neurons for language-family specificity (LLaMA3-8B).** Typically, a correlation ratio above **0.1** indicates a correlation, above **0.25** indicates a moderately strong correlation, and above **0.5** indicates a strong correlation.

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.13	0.08	0.10	0.07	Type-1 TN	0.16	0.10	0.15	0.10
Type-2 TN	0.21	0.17	0.17	0.16	Type-2 TN	0.35	0.23	0.27	0.22
Top-10	**ja**	**nl**	**ko**	**it**					
Type-1 TN	0.21	0.16	0.14	0.14					
Type-2 TN	0.52	0.34	0.42	0.31					

Table 11: **Correlation ratio of top-10, top-100, and top-1k Transfer Neurons for language-family specificity (Mistral-7B).**

shorter, more fragmented, and frequently cut off prematurely. Nevertheless, the model occasionally produces coherent text in another language. For instance, Table 39 presents a case where the output switches from Italian to English.

- Aya expanse-8B: Although severely broken sentences are less likely to be generated compared to the other two models, the outputs tend to contain many repetitions. However, in a subset of samples, the model (1) generated answers in another language that were well aligned with the input question, or (2) produced answers in a code-switching–like form, combining the input language with another language. Some examples are shown in in Tabs. 40 and 41.

Overall, for Type-2 neurons, certain samples yielded answers in a different language than those in the without-intervention setting, yet the responses remained coherent with the input question. In particular, the generated language tended to be linguistically proximate to the input language (e.g., input: Italian → output: English; input: Japanese → output: Chinese). These results in line with the PCA visualization of the geometry of language latent spaces in the hidden state space, where latent spaces corresponding to linguistically related languages remain closely clustered when the top-1k Type-2 neurons are deactivated (see Figs. 28 and 29).

K Statements

K.1 License for Artifacts

Models.

- LLaMA3-8B: `License for Llama family`
- Mistral-7B: `apache-2.0`
- Aya expanse-8B: `cc-by-nc-4.0`

Datasets.

- Tatoeba: `cc-by-2.0`
- MKQA: `cc-by-3.0`
- MMLU-ProX: `MIT License`
- PolyWrite: `odc-by-1.0`

Evaluation Tools.

- lm-evaluation-harness: `MIT License`

Consistency of Usage. All models, datasets and a evaluation tool were used in accordance with their original intended usage.

K.2 AI Agent Usage

AI agents were used for grammar checking during the writing of this paper.

Top-1000	ja	nl	ko	it	Top-100	ja	nl	ko	it
Type-1 TN	0.11	0.07	0.08	0.07	Type-1 TN	0.14	0.08	0.08	0.08
Type-2 TN	0.19	0.15	0.17	0.15	Type-2 TN	0.28	0.24	0.26	0.23
Top-10	**ja**	**nl**	**ko**	**it**					
Type-1 TN	0.10	0.15	0.15	0.15					
Type-2 TN	0.31	0.23	0.30	0.28					

Table 12: **Correlation ratio of top-10, top-100, and top-1k Transfer Neurons for language-family specificity (Aya expanse-8B).**

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.433	0.412	0.386
Dutch	0.300	0.261	0.245
Korean	0.337	0.305	0.287
Italian	0.297	0.253	0.254

Table 13: **% of Type-1 Transfer Neurons with correlation ratio** > 0.1 **and** $p < 0.05$

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.624	0.615	0.530
Dutch	0.635	0.559	0.379
Korean	0.556	0.484	0.471
Italian	0.631	0.507	0.390

Table 14: **% of Type-2 Transfer Neurons with correlation ratio** > 0.1 **and** $p < 0.05$

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.203	0.188	0.155
Dutch	0.094	0.076	0.064
Korean	0.136	0.125	0.098
Italian	0.089	0.066	0.065

Table 15: **% of Type-1 Transfer Neurons with correlation ratio** > 0.25 **and** $p < 0.05$

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.413	0.363	0.282
Dutch	0.377	0.232	0.165
Korean	0.345	0.254	0.233
Italian	0.383	0.221	0.144

Table 16: **% of Type-2 Transfer Neurons with correlation ratio** > 0.25 **and** $p < 0.05$

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.433	0.412	0.386
Dutch	0.300	0.261	0.245
Korean	0.337	0.305	0.287
Italian	0.297	0.253	0.254

Table 17: **% of Type-1 Transfer Neurons with correlation ratio** > 0.1 **and** $p < 0.05$**, Language-family label**

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.624	0.615	0.530
Dutch	0.635	0.559	0.379
Korean	0.556	0.484	0.471
Italian	0.631	0.507	0.390

Table 18: **% of Type-2 Transfer Neurons with correlation ratio** > 0.1 **and** $p < 0.05$**, Language-family label**

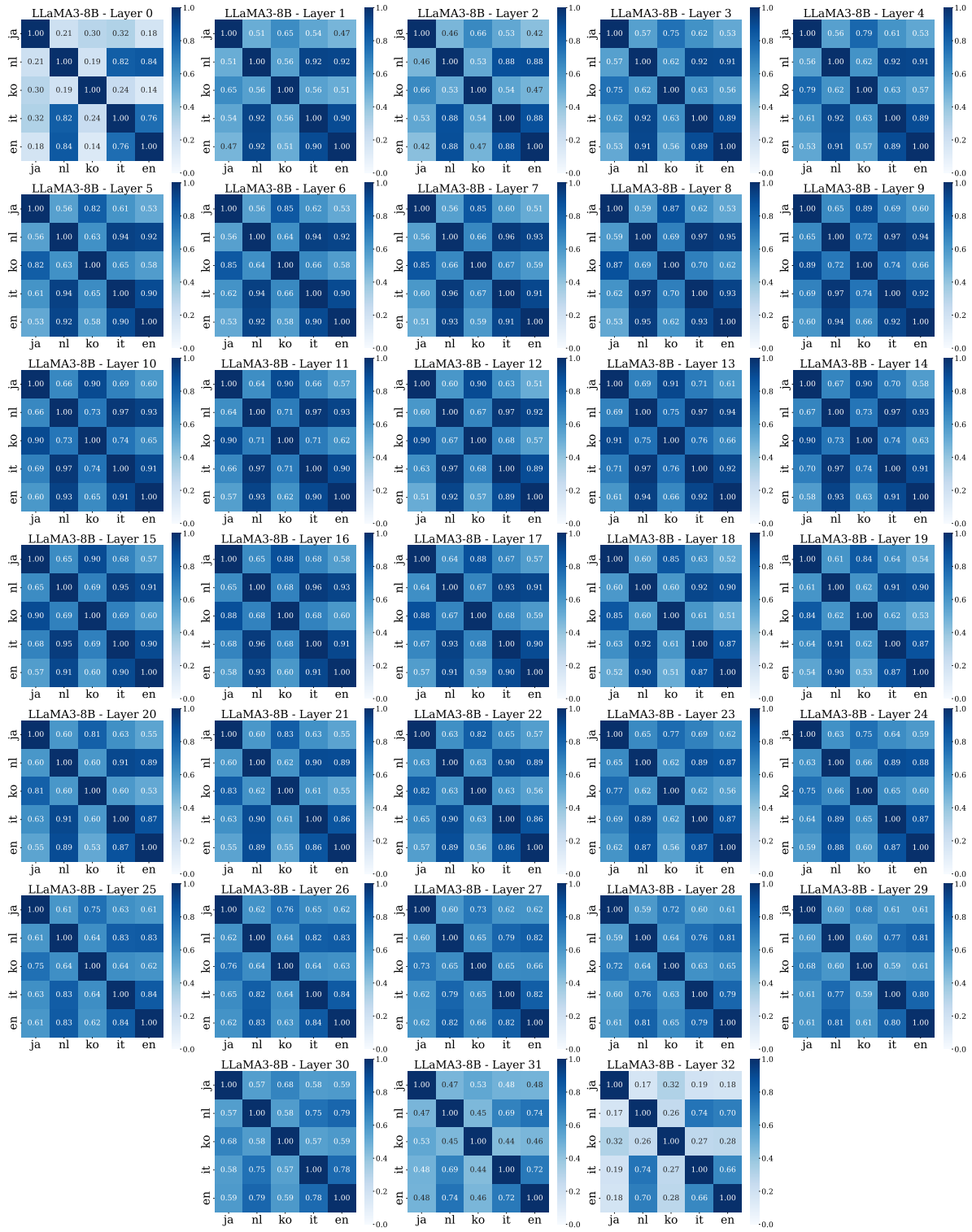


Figure 14: **The distance among centroids of language latent spaces (LLaMA3-8B).**

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.203	0.188	0.155
Dutch	0.094	0.076	0.064
Korean	0.136	0.125	0.098
Italian	0.089	0.066	0.065

Table 19: **% of Type-1 Transfer Neurons with correlation ratio** > 0.25 **and** $p < 0.05$**, Language-family label**

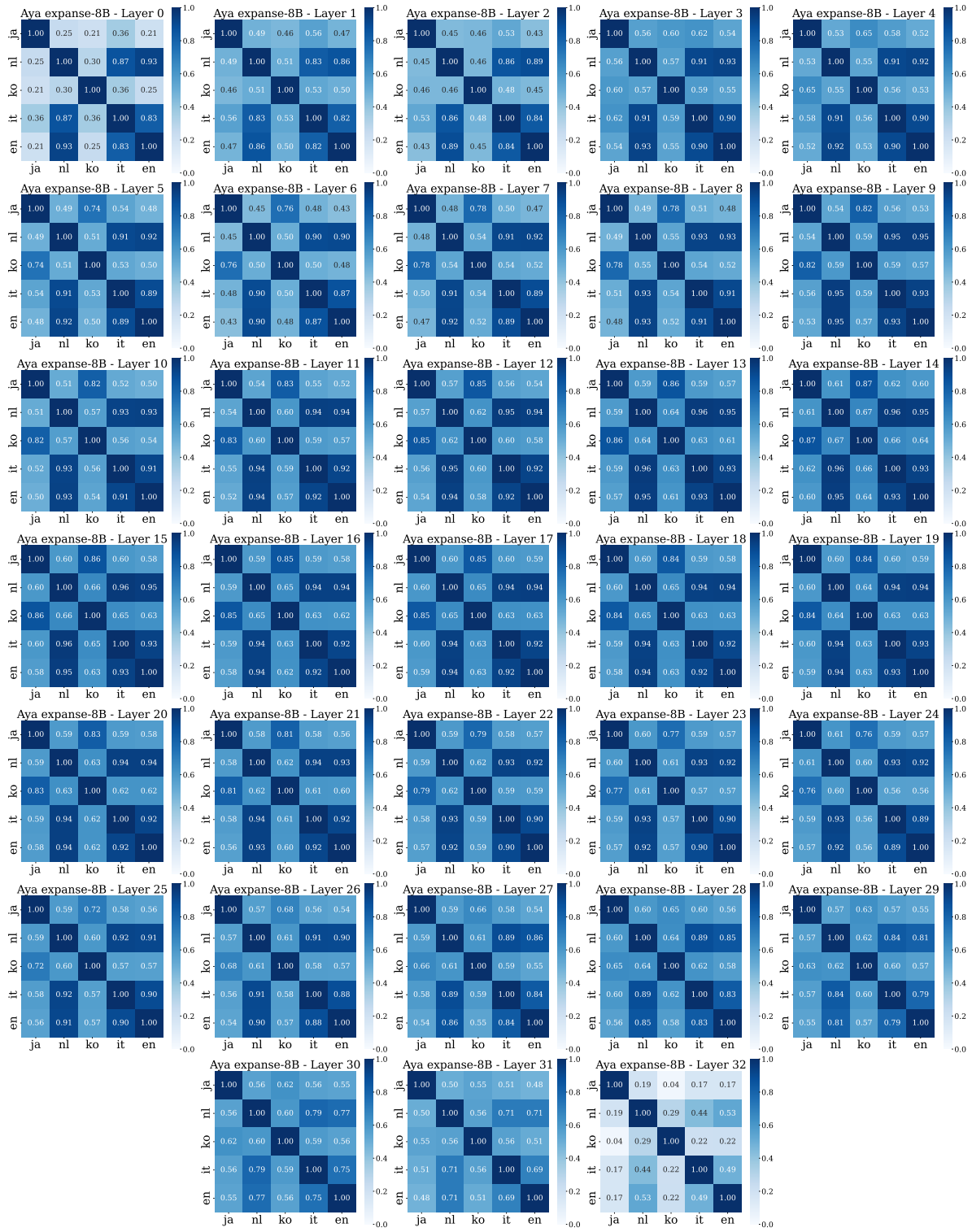


Figure 15: **The distance among centroids of language latent spaces (Aya expanse-8B).**

	LLaMA3-8B	Mistral-7B	Aya expanse-8B
Japanese	0.413	0.363	0.282
Dutch	0.377	0.232	0.165
Korean	0.345	0.254	0.233
Italian	0.383	0.221	0.144

Table 20: **% of Type-2 Transfer Neurons with correlation ratio** > 0.25 **and** $p < 0.05$**, Language-family label**

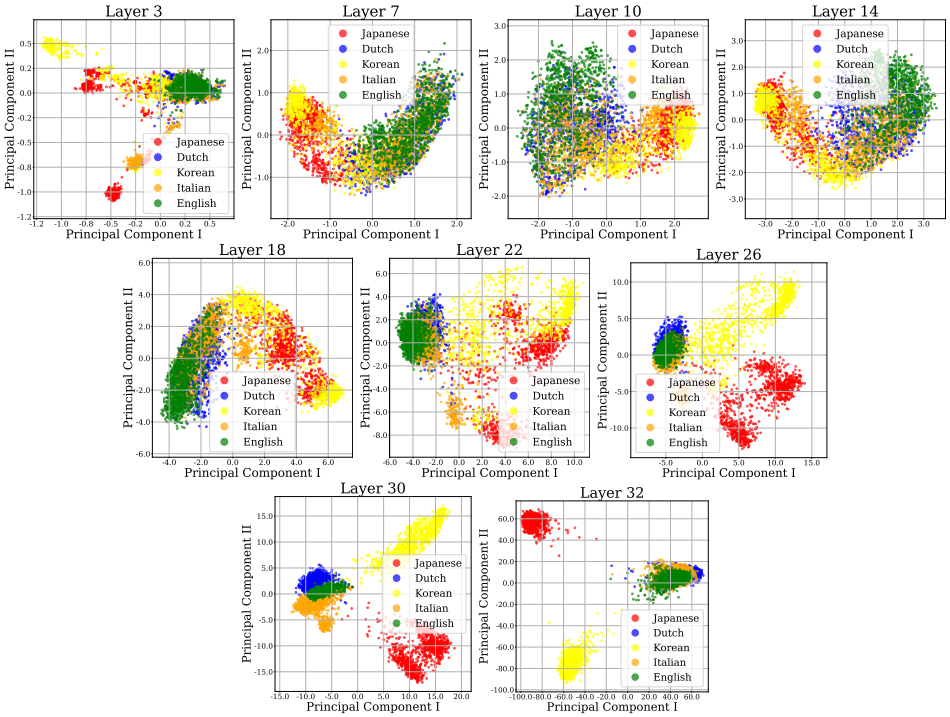


Figure 16: **The resutls of PCA applied to the hidden representations across layers (LLaMA3-8B).**

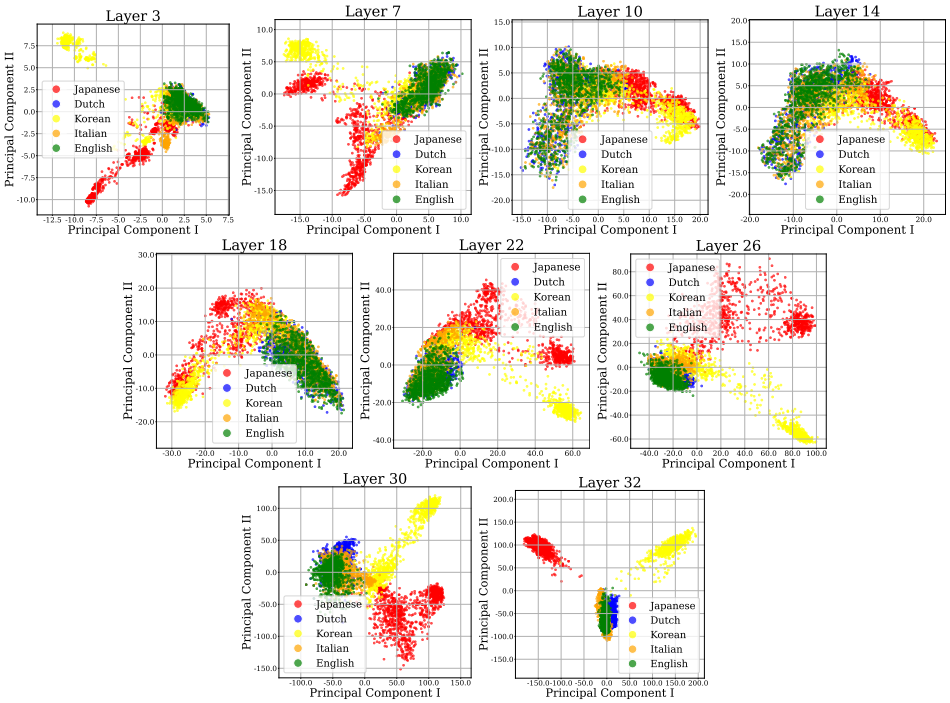


Figure 17: **The resutls of PCA applied to the hidden representations across layers (Aya expanse-8B).**

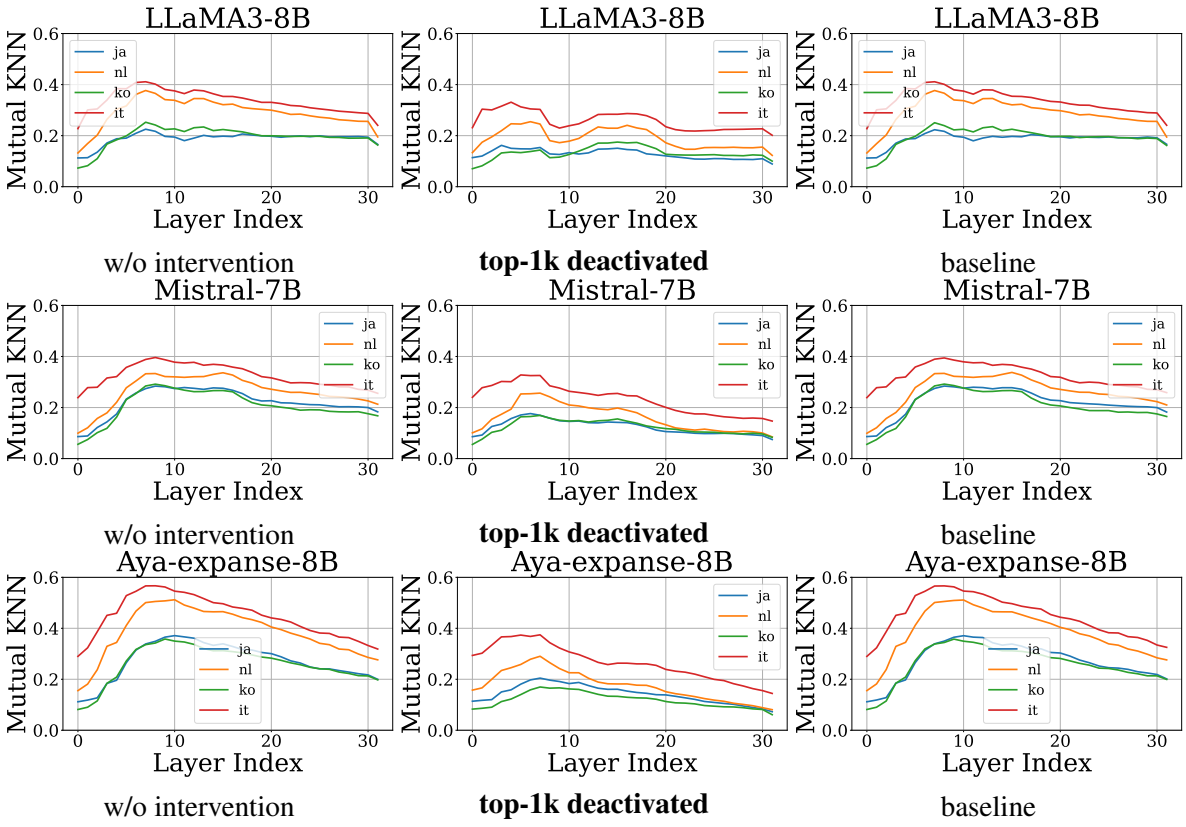


Figure 18: **Kernel-Based Similarity with Deactivation of Top-1k Type-1 Transfer Neurons (k=10).** The middle column presents the results after deactivating the top-1k Type-1 Transfer Neurons. The right column shows the results after deactivating 1k randomly sampled neurons from the same layers as Type-1 neurons for a baseline, while the left column shows the original results without any intervention.

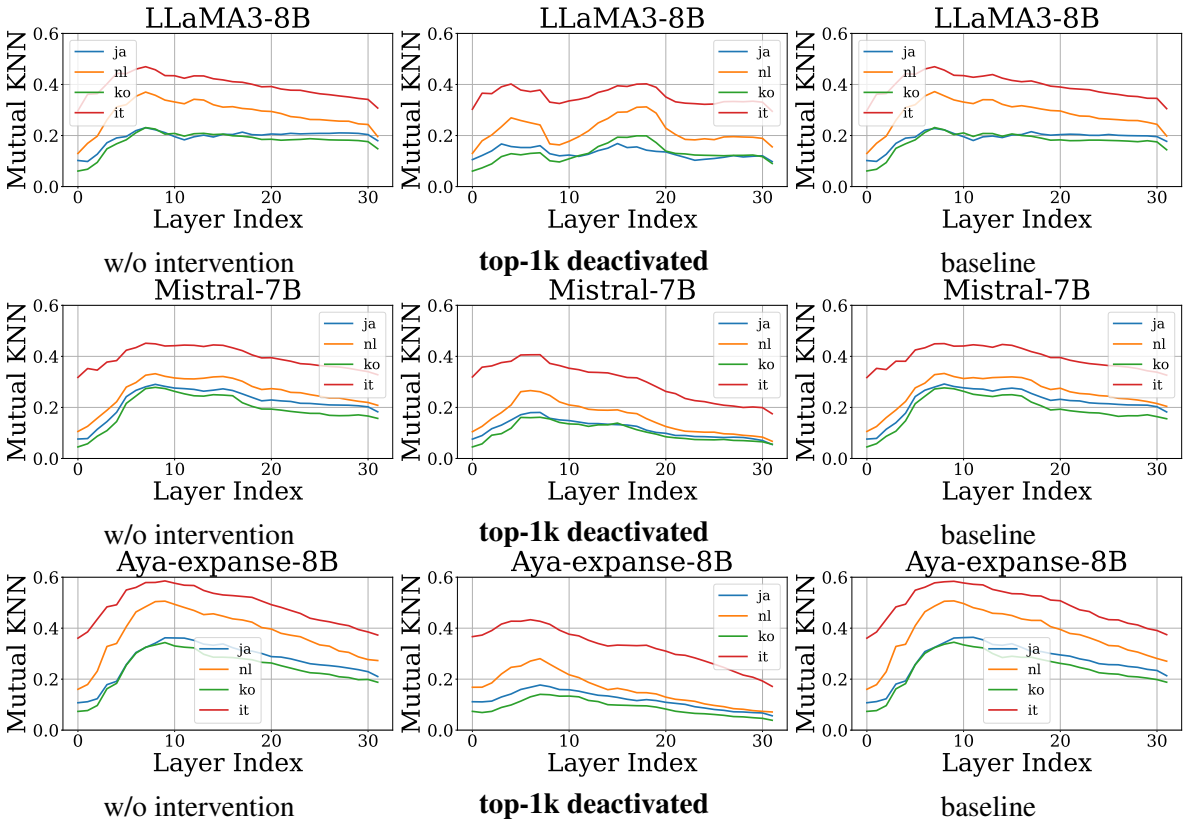
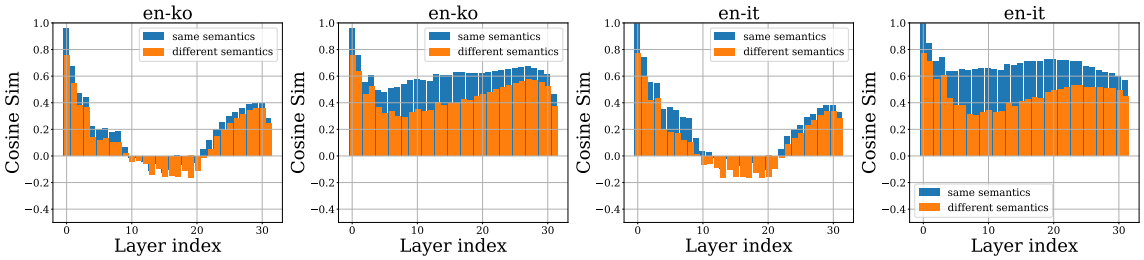
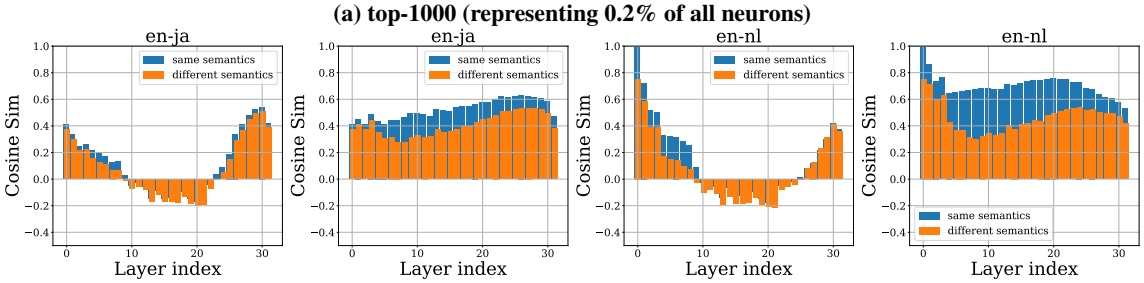


Figure 19: **Kernel-Based Similarity with Deactivation of Top-1k Type-1 Transfer Neurons (k=5).**

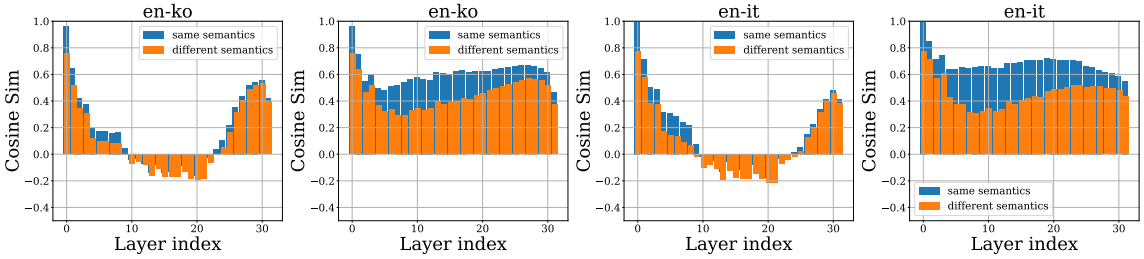


(a) en-ko (b) en-ko (baseline) (c) en-it (d) en-it (baseline)

(a) top-1000 (representing 0.2% of all neurons)

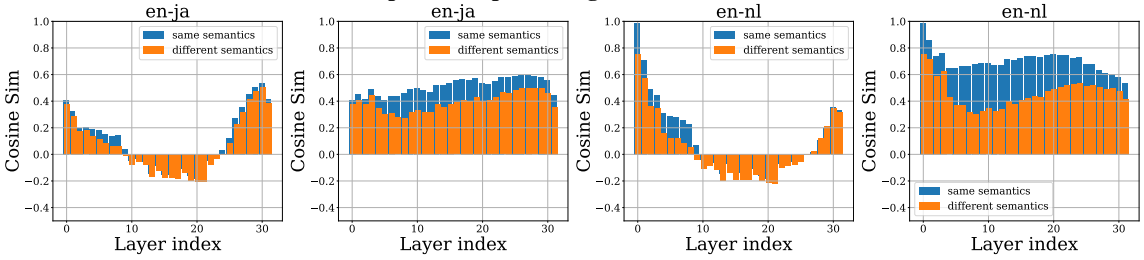


(e) en-ja (f) en-ja (baseline) (g) en-nl (h) en-nl (baseline)

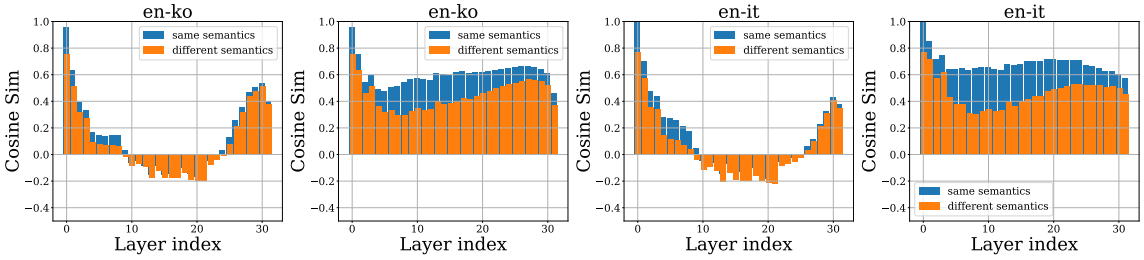


(i) en-ko (j) en-ko (baseline) (k) en-it (l) en-it (baseline)

(b) top-3000 (representing 0.6% of all neurons)



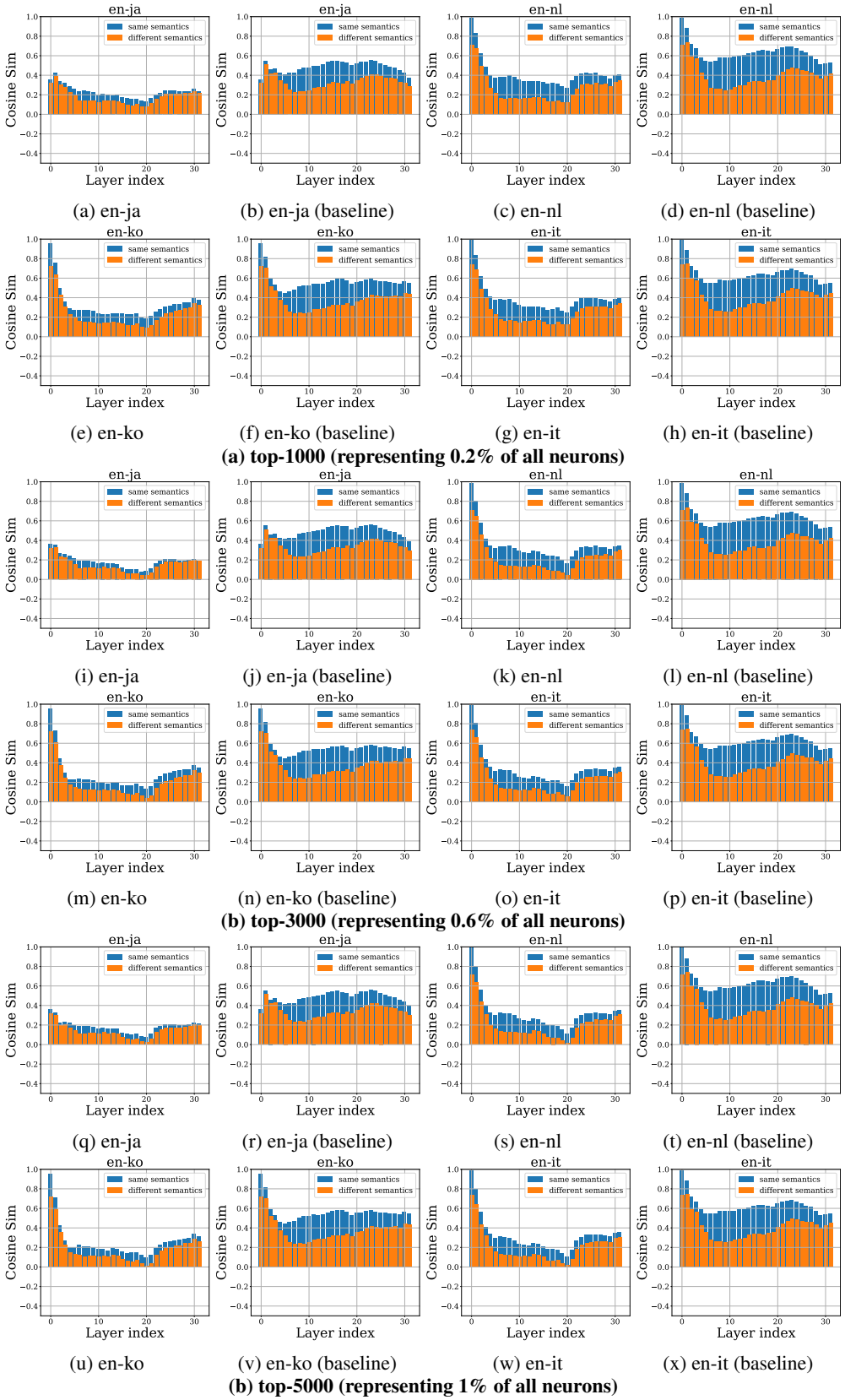
(m) en-ja (n) en-ja (baseline) (o) en-nl (p) en-nl (baseline)



(q) en-ko (r) en-ko (baseline) (s) en-it (t) en-it (baseline)

(b) top-5000 (representing 1% of all neurons)

Figure 20: **Similarity of hidden states across layers while deactivating Type-1 Transfer Neurons (LLaMA3-8B).**



(a) en-ja (b) en-ja (baseline) (c) en-nl (d) en-nl (baseline)

(e) en-ko (f) en-ko (baseline) (g) en-it (h) en-it (baseline)

(a) top-1000 (representing 0.2% of all neurons)

(i) en-ja (j) en-ja (baseline) (k) en-nl (l) en-nl (baseline)

(m) en-ko (n) en-ko (baseline) (o) en-it (p) en-it (baseline)

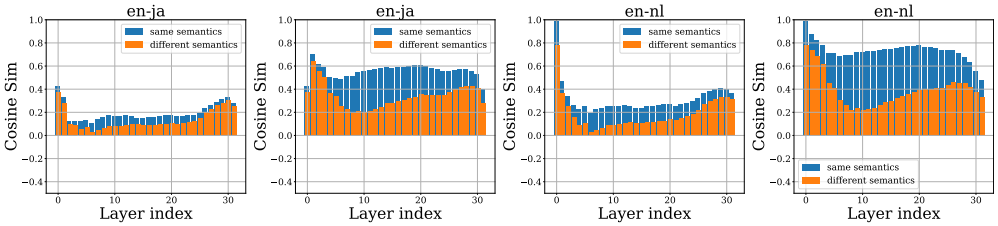
(b) top-3000 (representing 0.6% of all neurons)

(q) en-ja (r) en-ja (baseline) (s) en-nl (t) en-nl (baseline)

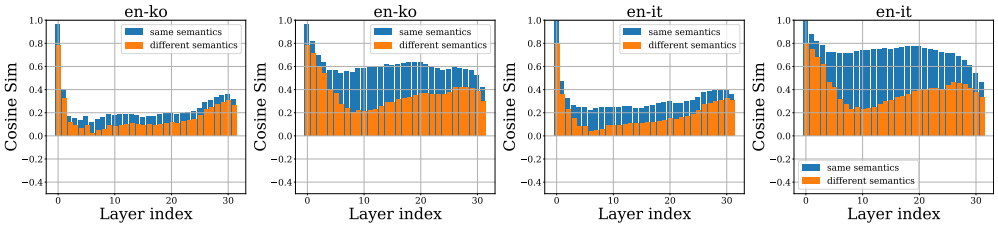
(u) en-ko (v) en-ko (baseline) (w) en-it (x) en-it (baseline)

(b) top-5000 (representing 1% of all neurons)

Figure 21: **Similarity of hidden states across layers while deactivating Type-1 Transfer Neurons (Mistral-7B).**

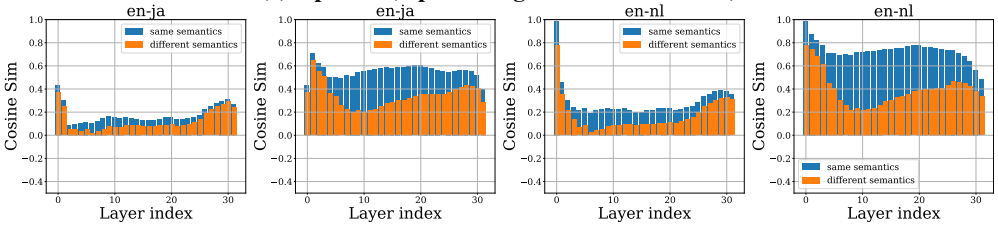


(a) en-ja (b) en-ja (baseline) (c) en-nl (d) en-nl (baseline)

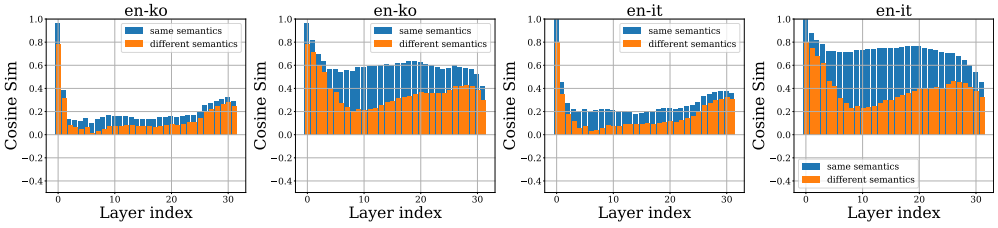


(e) en-ko (f) en-ko (baseline) (g) en-it (h) en-it (baseline)

(a) top-1000 (representing 0.2% of all neurons)

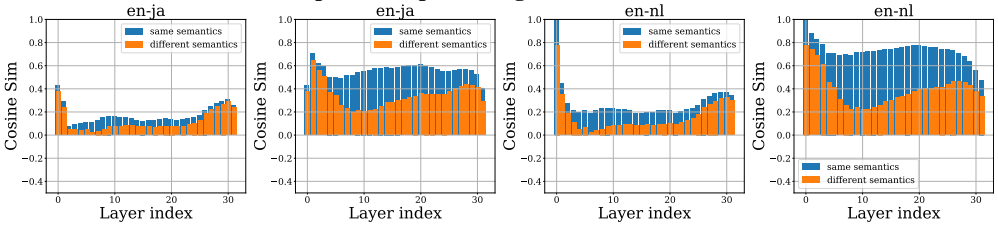


(i) en-ja (j) en-ja (baseline) (k) en-nl (l) en-nl (baseline)

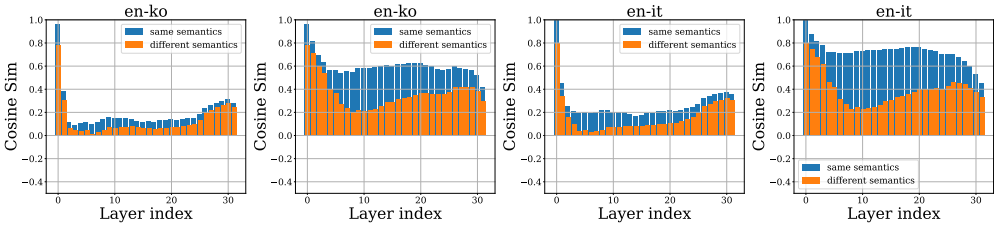


(m) en-ko (n) en-ko (baseline) (o) en-it (p) en-it (baseline)

(b) top-3000 (representing 0.6% of all neurons)



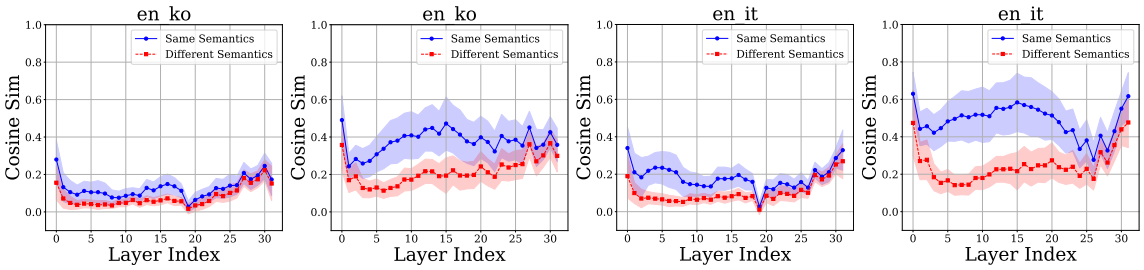
(q) en-ja (r) en-ja (baseline) (s) en-nl (t) en-nl (baseline)



(u) en-ko (v) en-ko (baseline) (w) en-it (x) en-it (baseline)

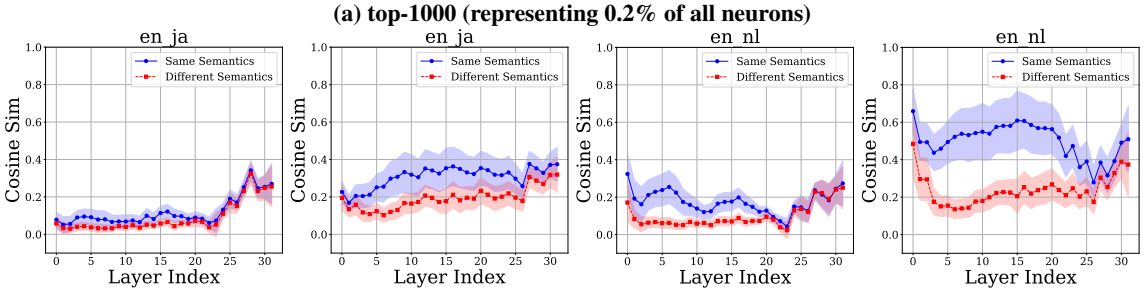
(b) top-5000 (representing 1% of all neurons)

Figure 22: **Similarity of hidden states across layers while deactivating Type-1 Transfer Neurons (Aya expanse-8B).**

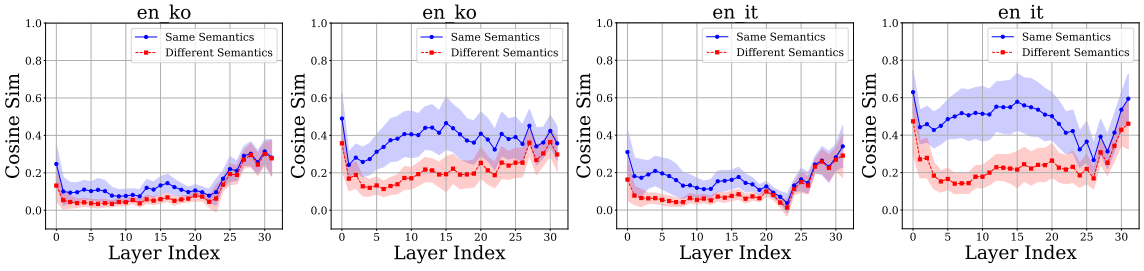


(a) en-ko (b) en-ko (baseline) (c) en-it (d) en-it (baseline)

(a) top-1000 (representing 0.2% of all neurons)

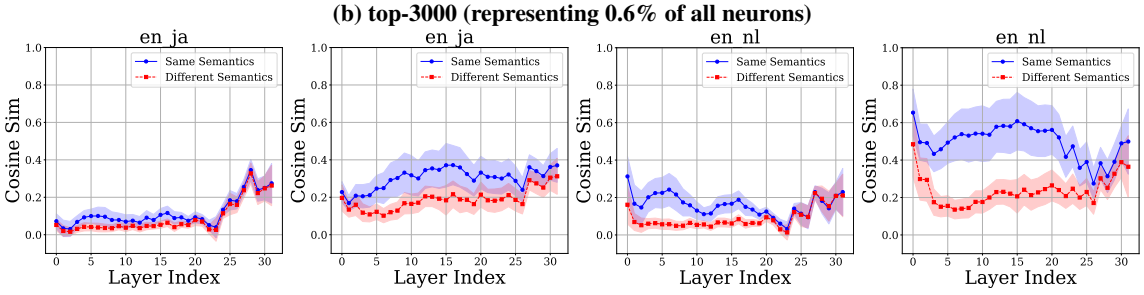


(e) en-ja (f) en-ja (baseline) (g) en-nl (h) en-nl (baseline)

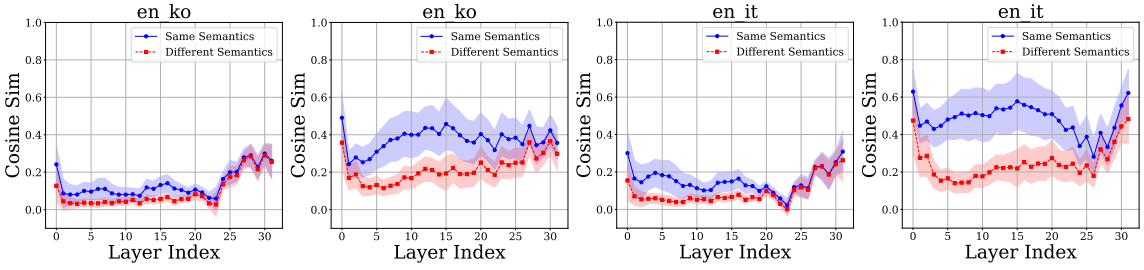


(i) en-ko (j) en-ko (baseline) (k) en-it (l) en-it (baseline)

(b) top-3000 (representing 0.6% of all neurons)



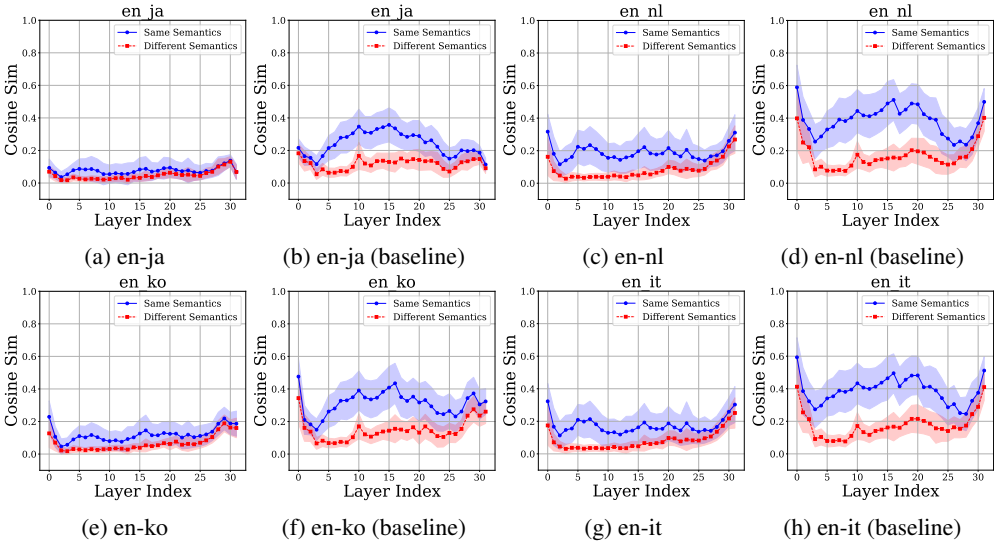
(m) en-ja (n) en-ja (baseline) (o) en-nl (p) en-nl (baseline)



(q) en-ko (r) en-ko (baseline) (s) en-it (t) en-it (baseline)

(b) top-5000 (representing 1% of all neurons)

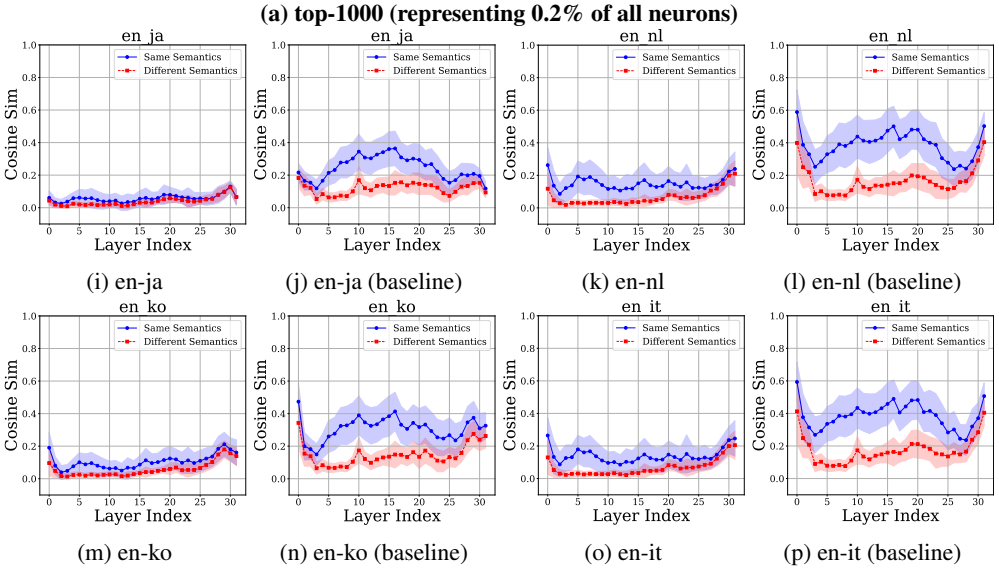
Figure 23: **Similarity of activation patterns across layers while deactivating Type-1 Transfer Neurons (LLaMA3-8B).**



(a) en-ja (b) en-ja (baseline) (c) en-nl (d) en-nl (baseline)

(e) en-ko (f) en-ko (baseline) (g) en-it (h) en-it (baseline)

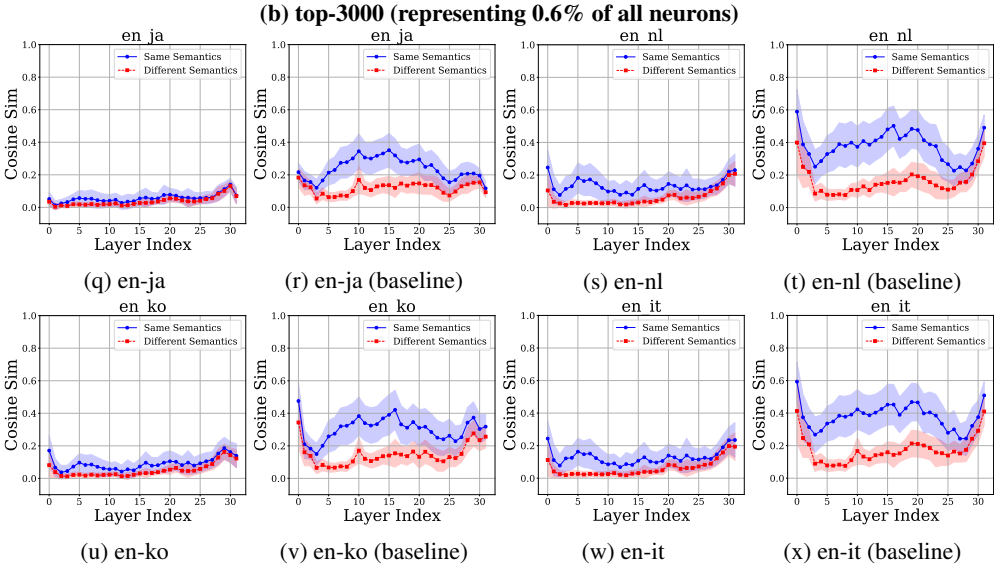
(a) top-1000 (representing 0.2% of all neurons)



(i) en-ja (j) en-ja (baseline) (k) en-nl (l) en-nl (baseline)

(m) en-ko (n) en-ko (baseline) (o) en-it (p) en-it (baseline)

(b) top-3000 (representing 0.6% of all neurons)

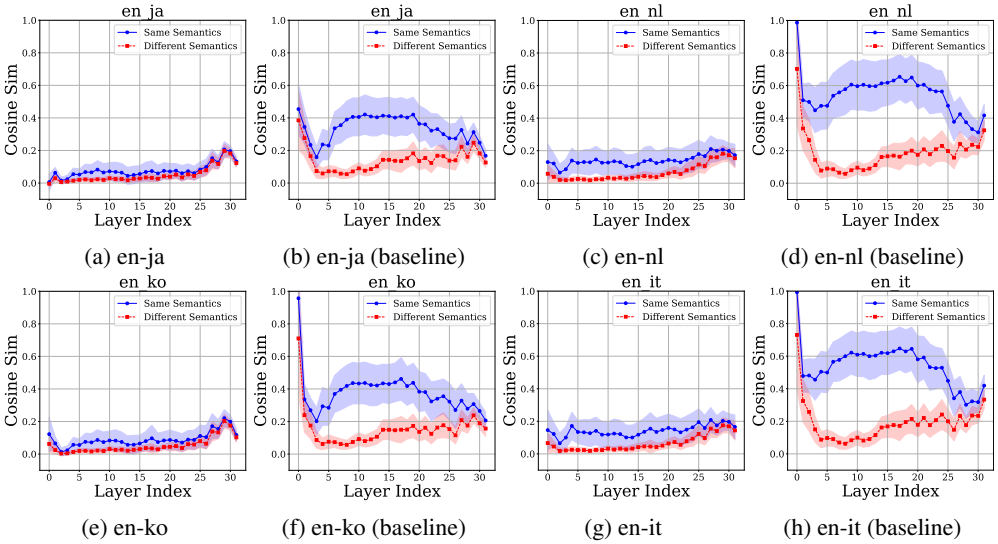


(q) en-ja (r) en-ja (baseline) (s) en-nl (t) en-nl (baseline)

(u) en-ko (v) en-ko (baseline) (w) en-it (x) en-it (baseline)

(b) top-5000 (representing 1% of all neurons)

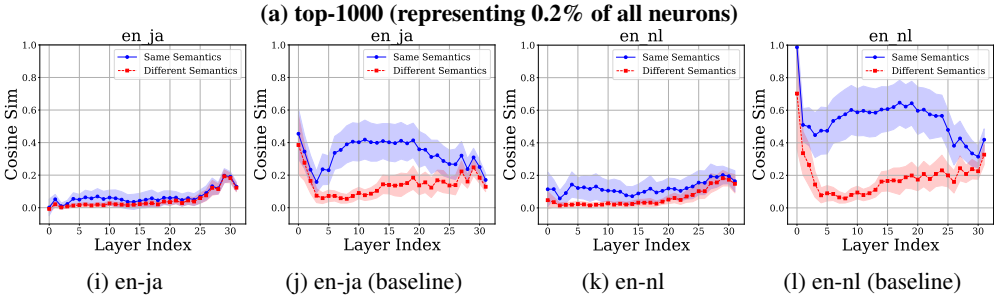
Figure 24: **Similarity of activation patterns across layers while deactivating Type-1 Transfer Neurons (Mistral-7B).**



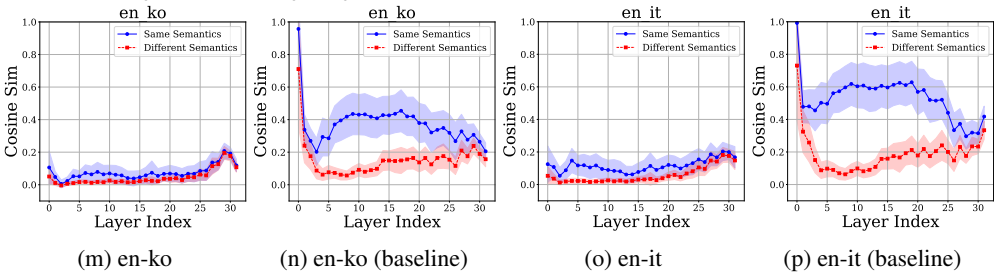
(a) en-ja (b) en-ja (baseline) (c) en-nl (d) en-nl (baseline)

(e) en-ko (f) en-ko (baseline) (g) en-it (h) en-it (baseline)

(a) top-1000 (representing 0.2% of all neurons)

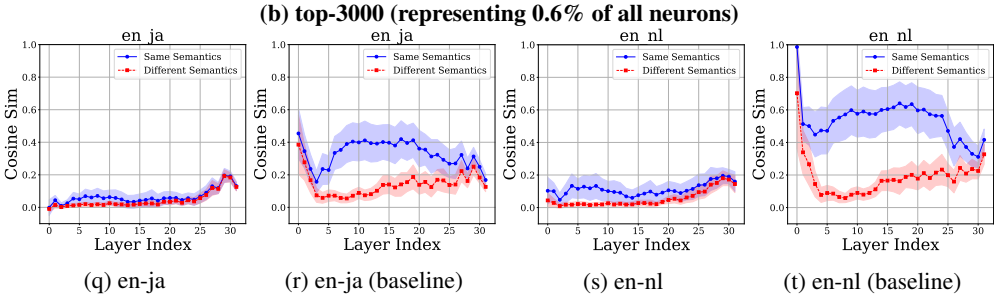


(i) en-ja (j) en-ja (baseline) (k) en-nl (l) en-nl (baseline)

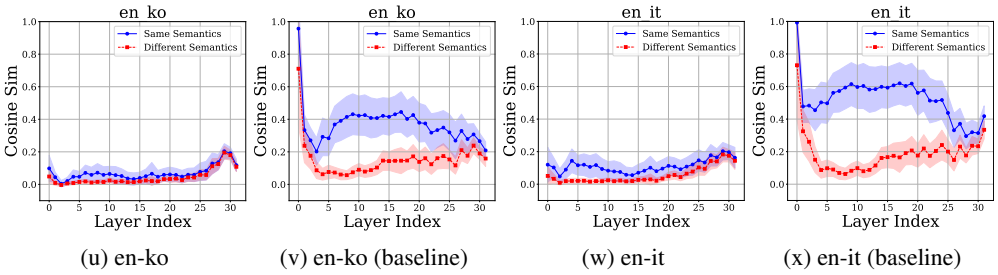


(m) en-ko (n) en-ko (baseline) (o) en-it (p) en-it (baseline)

(b) top-3000 (representing 0.6% of all neurons)



(q) en-ja (r) en-ja (baseline) (s) en-nl (t) en-nl (baseline)



(u) en-ko (v) en-ko (baseline) (w) en-it (x) en-it (baseline)

(b) top-5000 (representing 1% of all neurons)

Figure 25: **Similarity of activation patterns across layers while deactivating Type-1 Transfer Neurons (Aya expanse-8B).**

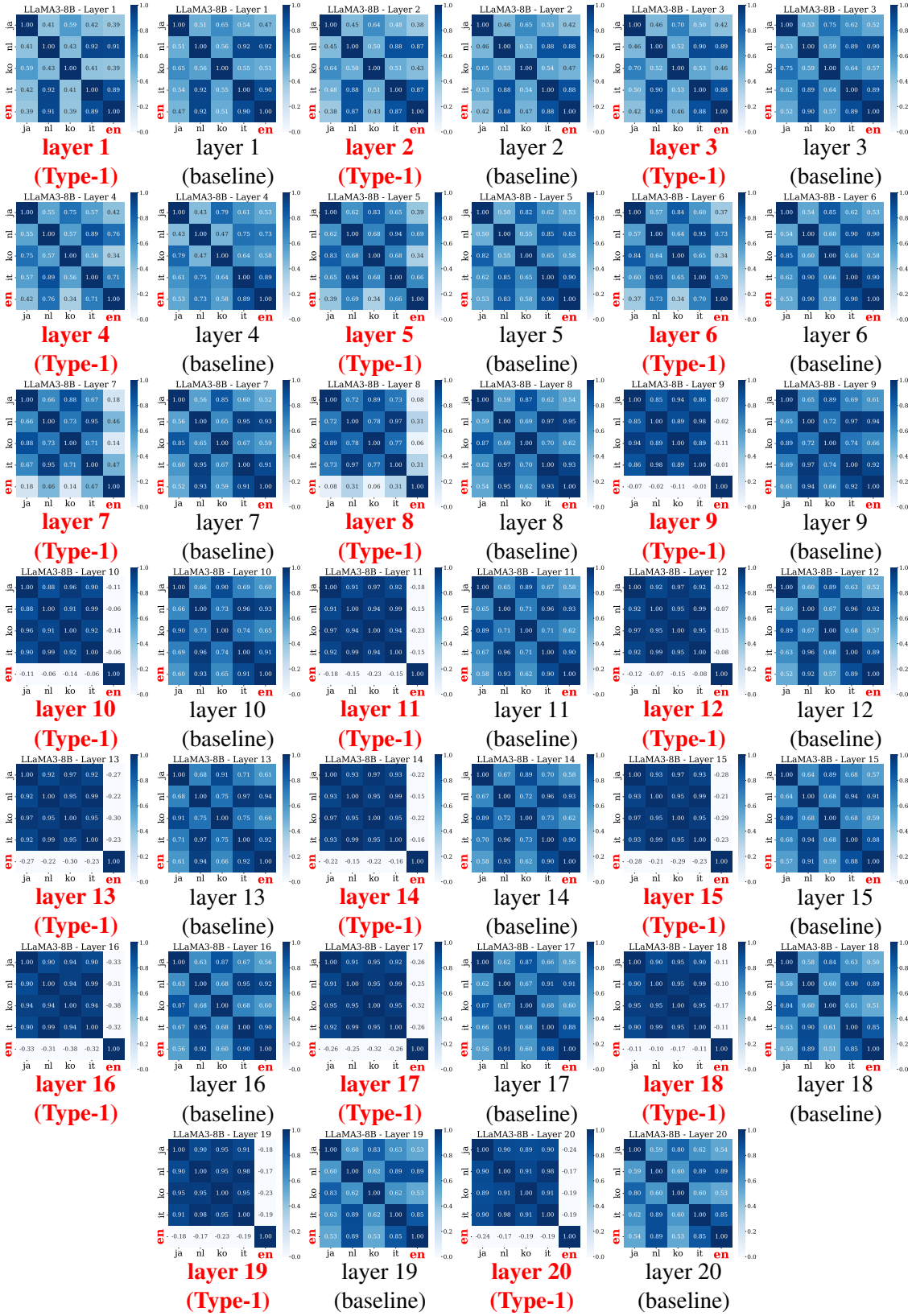


Figure 26: **Distance among language latent spaces while deactivating Top-1k Type-1 Transfer Neurons (LLaMA3-8B)**. **Layer (Type-1)** indicates the result of deactivating the top-1k Type-1 Transfer Neurons, whereas "baseline" refers to the result of deactivating 1k randomly sampled neurons from the same layers as the Type-1 Transfer Neurons.

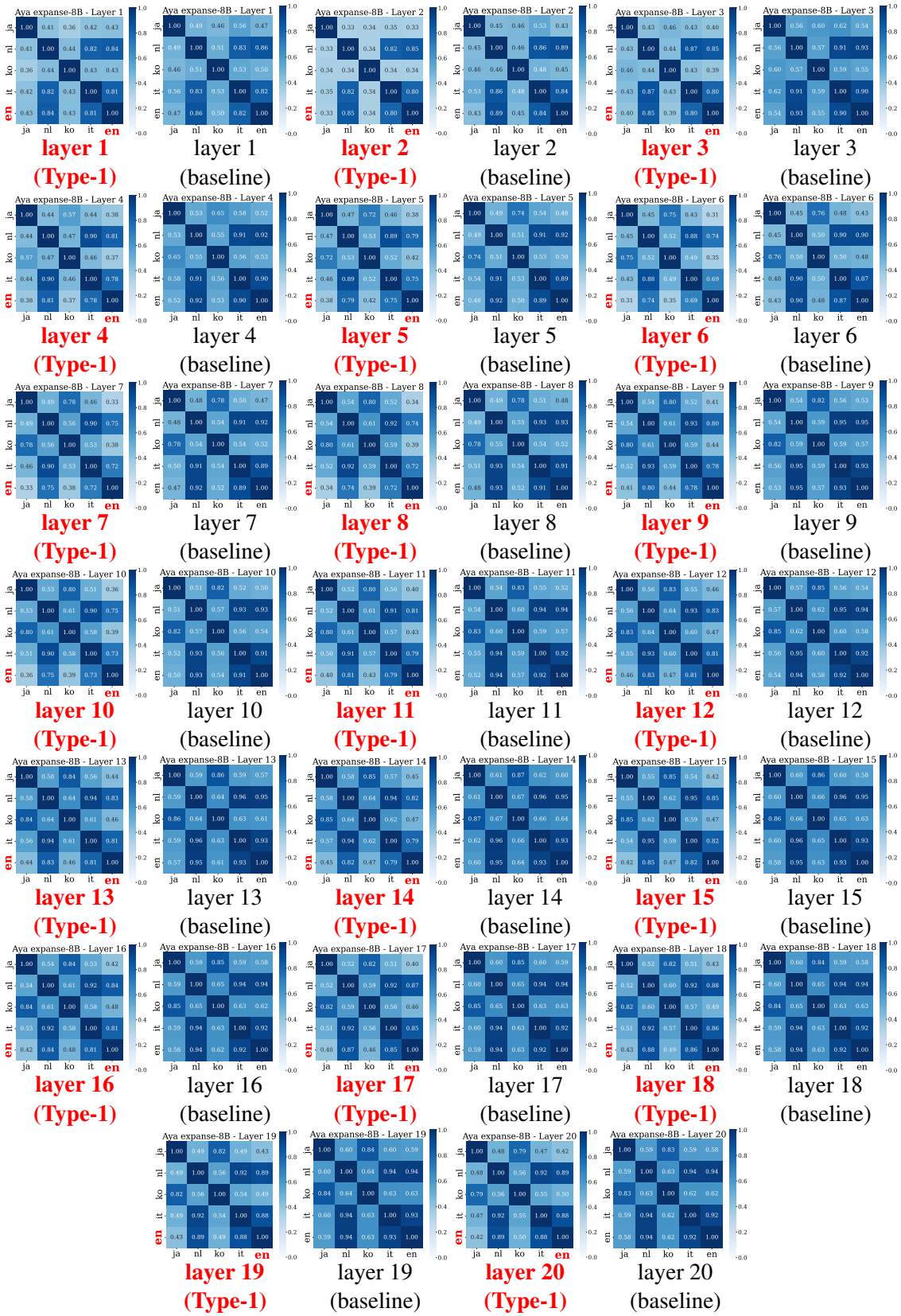
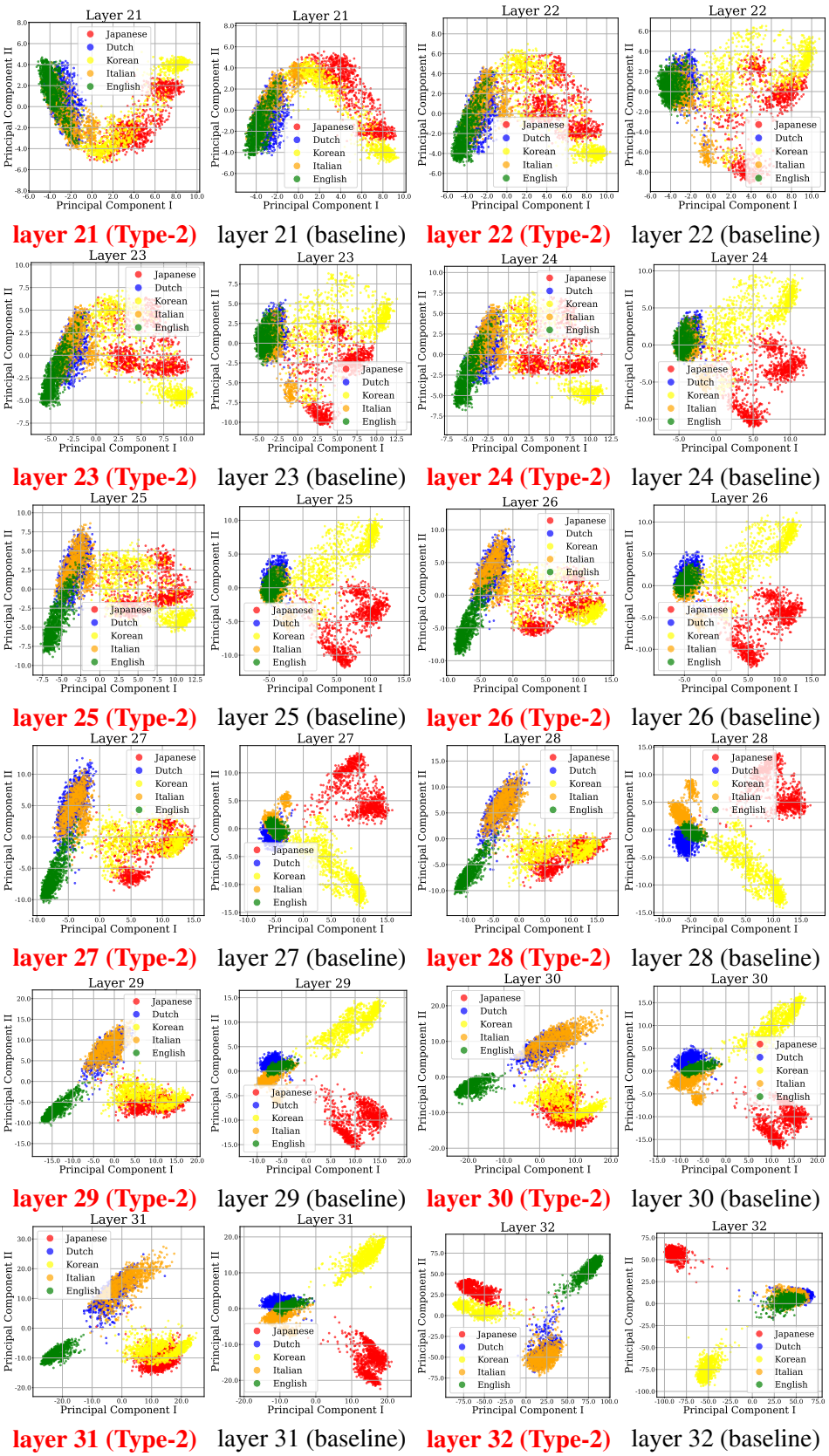


Figure 27: **Distance among language latent spaces while deactivating Top-1k Type-1 Transfer Neurons (Aya expanse-8B)**.



layer 21 (Type-2) layer 21 (baseline) **layer 22 (Type-2)** layer 22 (baseline)

layer 23 (Type-2) layer 23 (baseline) **layer 24 (Type-2)** layer 24 (baseline)

layer 25 (Type-2) layer 25 (baseline) **layer 26 (Type-2)** layer 26 (baseline)

layer 27 (Type-2) layer 27 (baseline) **layer 28 (Type-2)** layer 28 (baseline)

layer 29 (Type-2) layer 29 (baseline) **layer 30 (Type-2)** layer 30 (baseline)

layer 31 (Type-2) layer 31 (baseline) **layer 32 (Type-2)** layer 32 (baseline)

Figure 28: **The resutls of PCA while deactivating Top-1k Type-2 Transfer Neurons (LLaMA3-8B)**. **Layer (Type-2)** indicates the result of deactivating the top-1k Type-2 Transfer Neurons, whereas "baseline" refers to the result of deactivating 1k randomly sampled neurons from the same layers as the Type-2 Transfer Neurons. The English features are the only ones visualized without intervention.

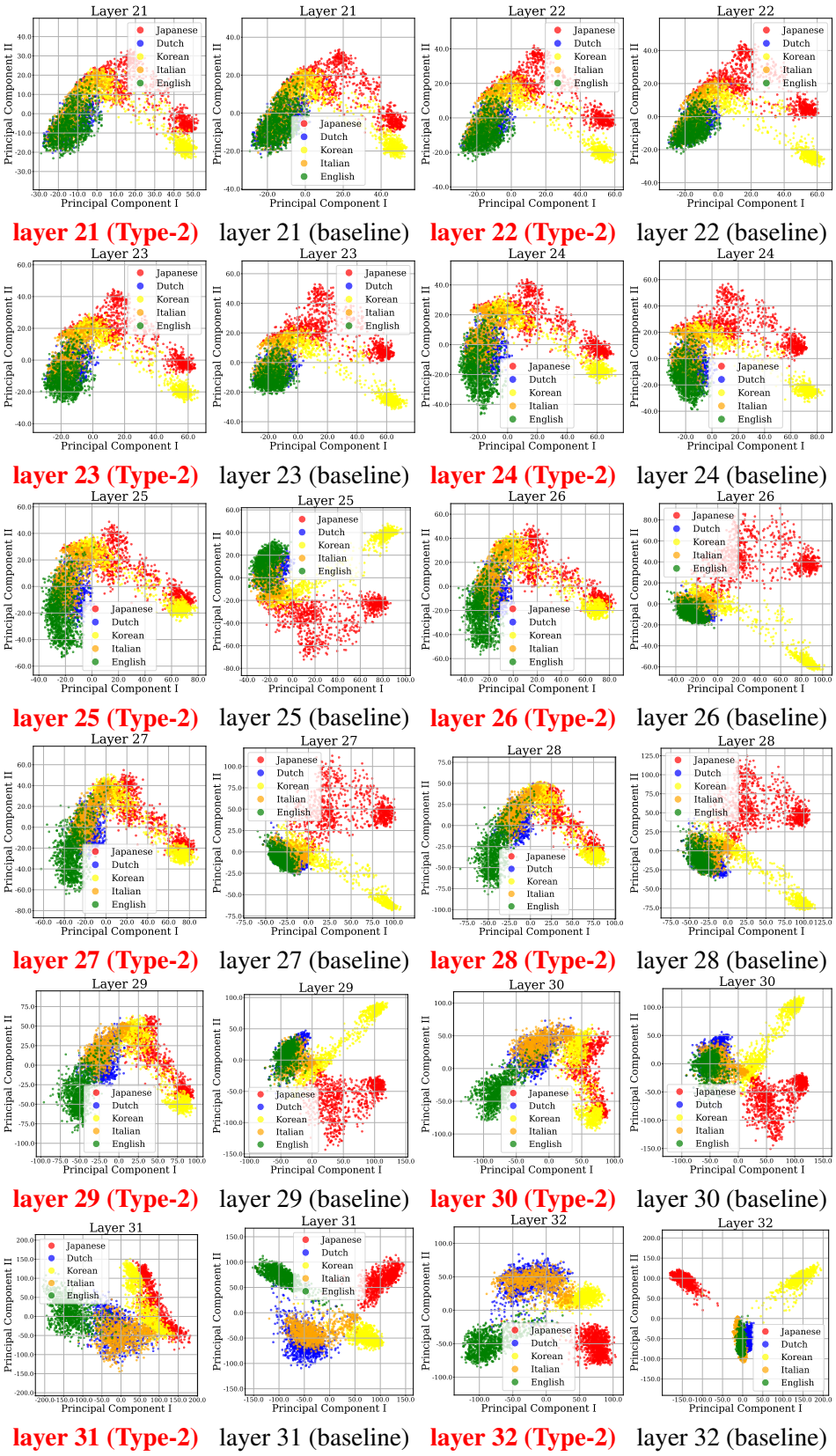


Figure 29: **The resutls of PCA while deactivating Top-1k Type-2 Transfer Neurons (Aya expanse-8B)**.

.

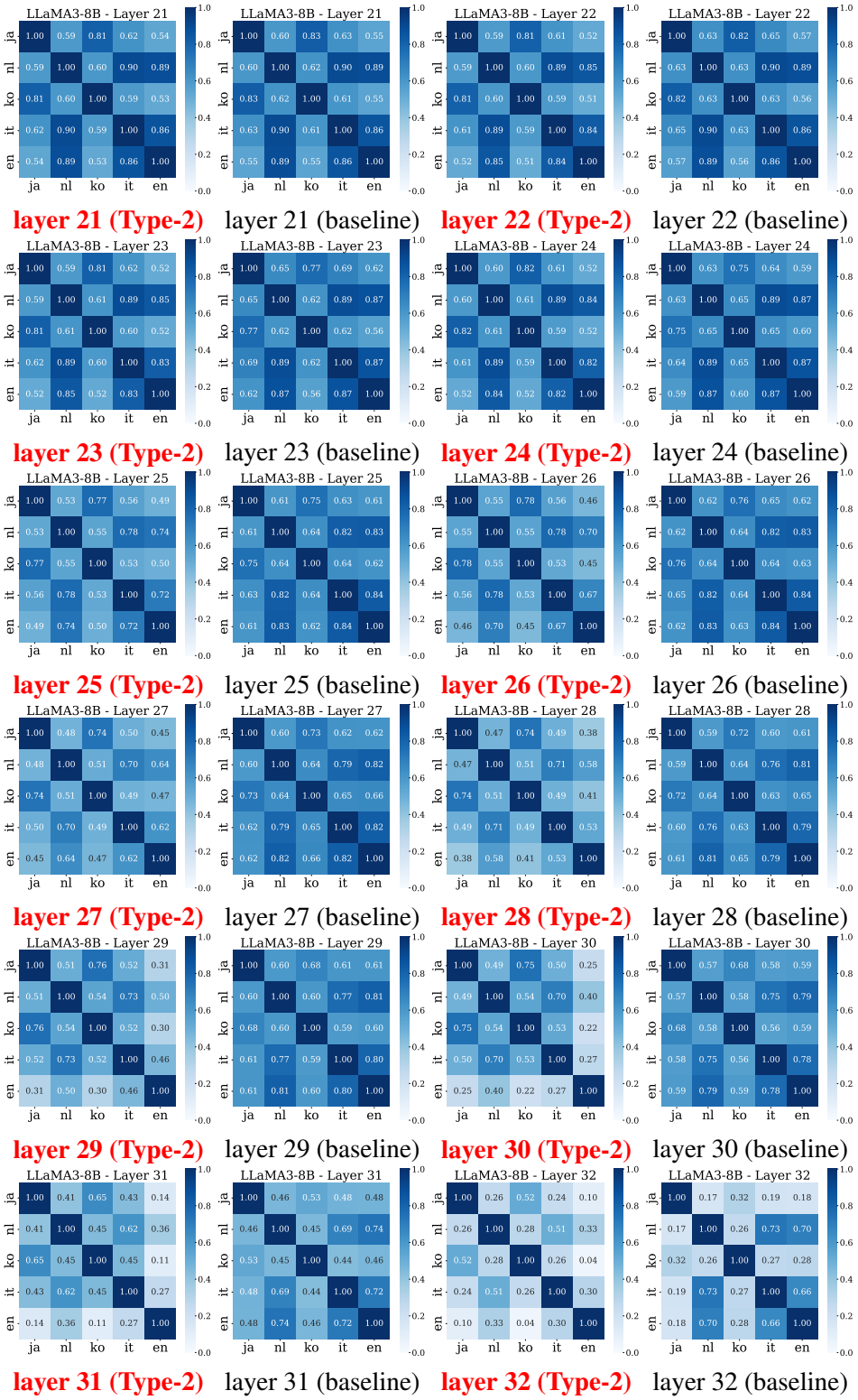


Figure 30: **Distance among centroids of language latent spaces while deactivating top-1k Type-2 Transfer Neurons (LLaMA3-8B)**. **Layer (Type-2)** indicates the result of deactivating Type-2 neurons, whereas "baseline" refers tot the result of deactivating randomly sampled neurons.

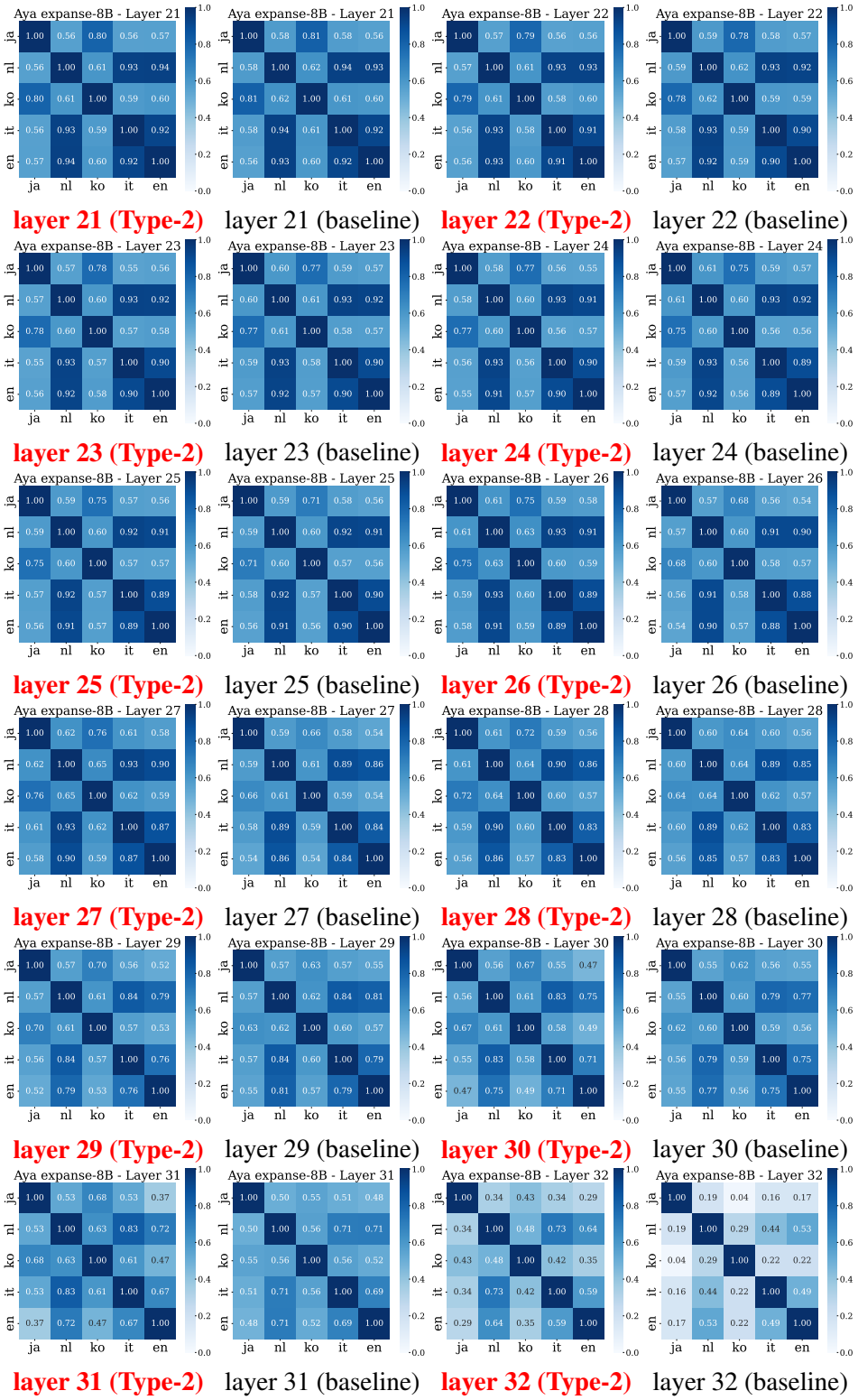
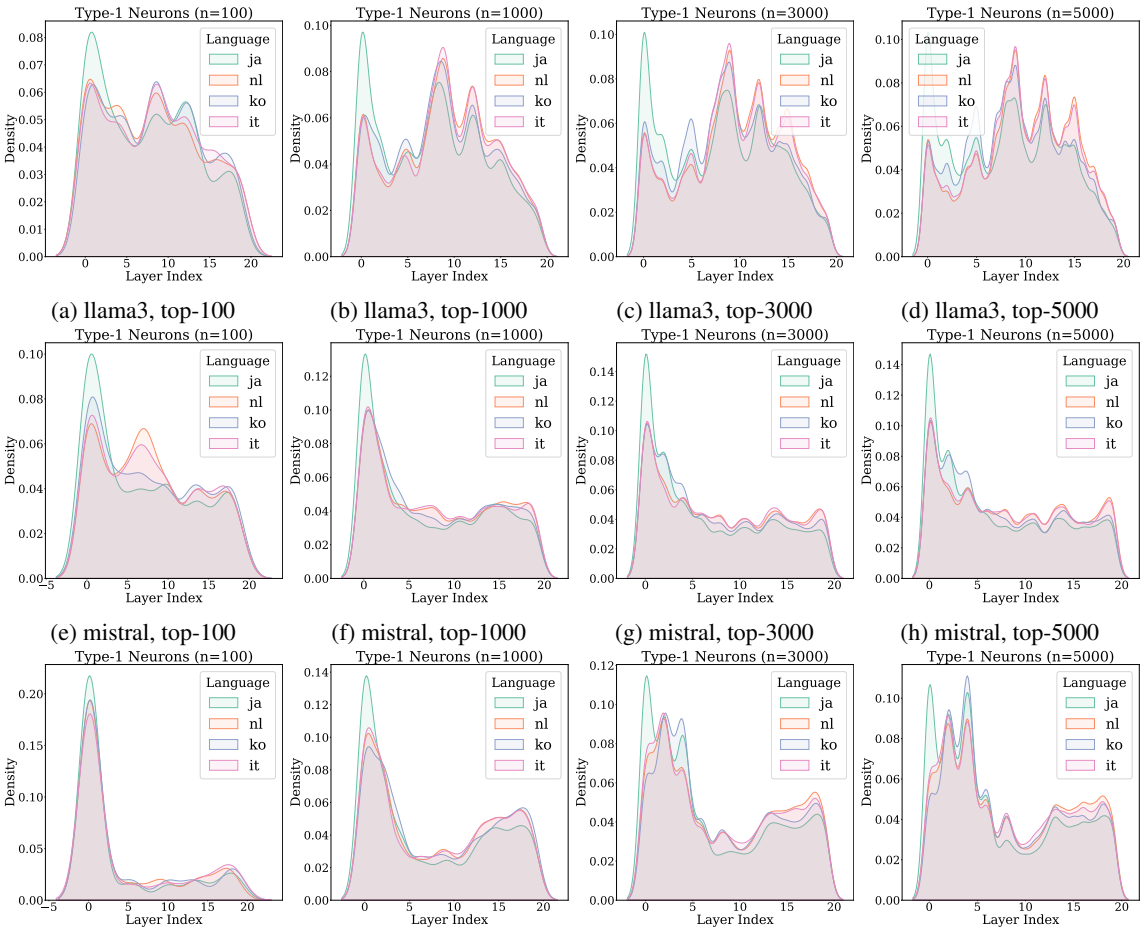


Figure 31: **Distance among centroids of language latent spaces while deactivating Top-1k Type-2 Transfer Neurons (Aya expanse-8B)**.



(a) llama3, top-100 (b) llama3, top-1000 (c) llama3, top-3000 (d) llama3, top-5000

(e) mistral, top-100 (f) mistral, top-1000 (g) mistral, top-3000 (h) mistral, top-5000

(i) aya, top-100 (j) aya, top-1000 (k) aya, top-3000 (l) aya, top-5000

Figure 32: **Distribution of Type-1 Transfer Neurons (LLaMA3-8B, Mistral-7B, and Aya expanse-8B).** 1-20 layers.

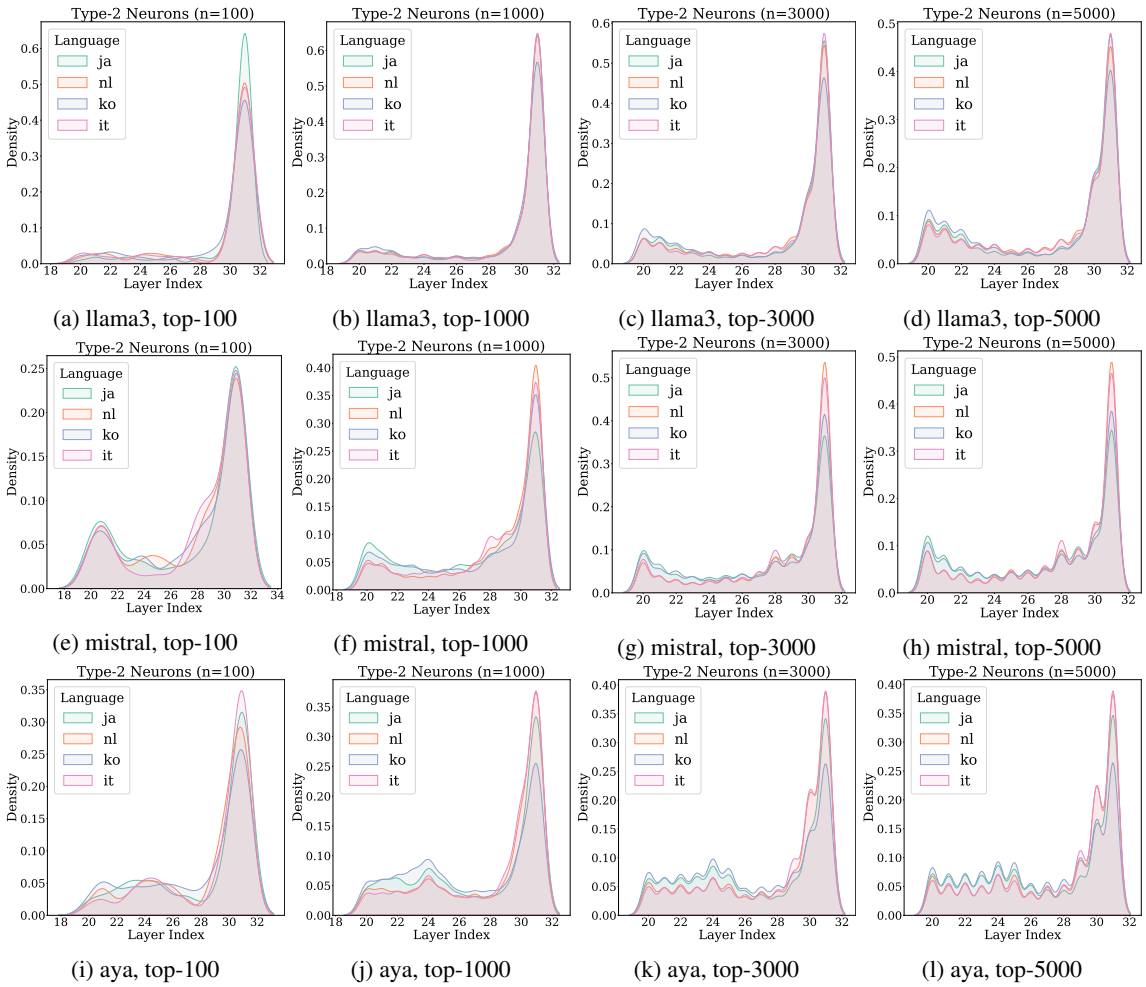


Figure 33: **Distribution of Type-2 Transfer Neurons (LLaMA3-8B, Mistral-7B, and Aya expanse-8B).** 21-32 layers.

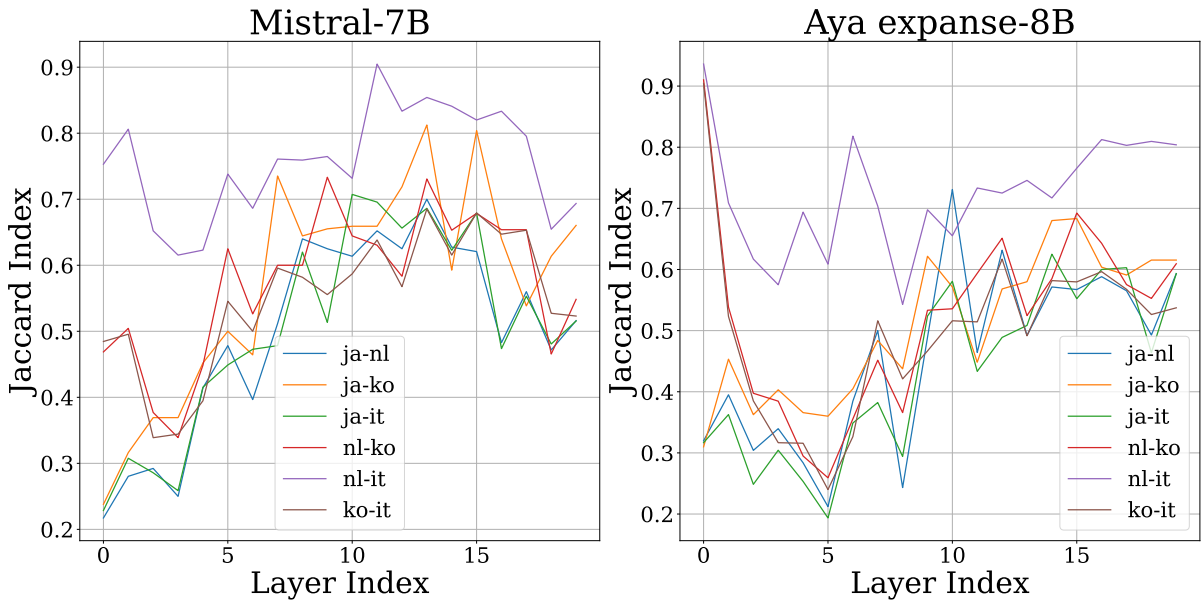


Figure 34: **Overlap ratio (Jaccard Index) of Type-1 Transfer Neurons across language pairs and decoder layers (Mistral-7B, and Aya expanse-8B)**. Higher values on the vertical axis indicate greater overlap of neurons across languages.

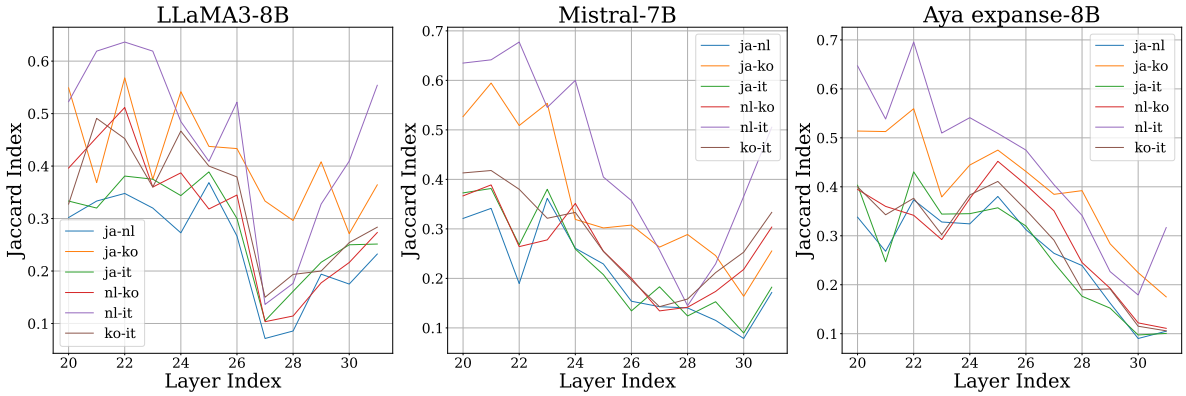


Figure 35: **Overlap ratio (Jaccard Index) of Type-2 Transfer Neurons across language pairs and decoder layers (LLaMA3-8B, Mistral-7B, and Aya expanse-8B)**.

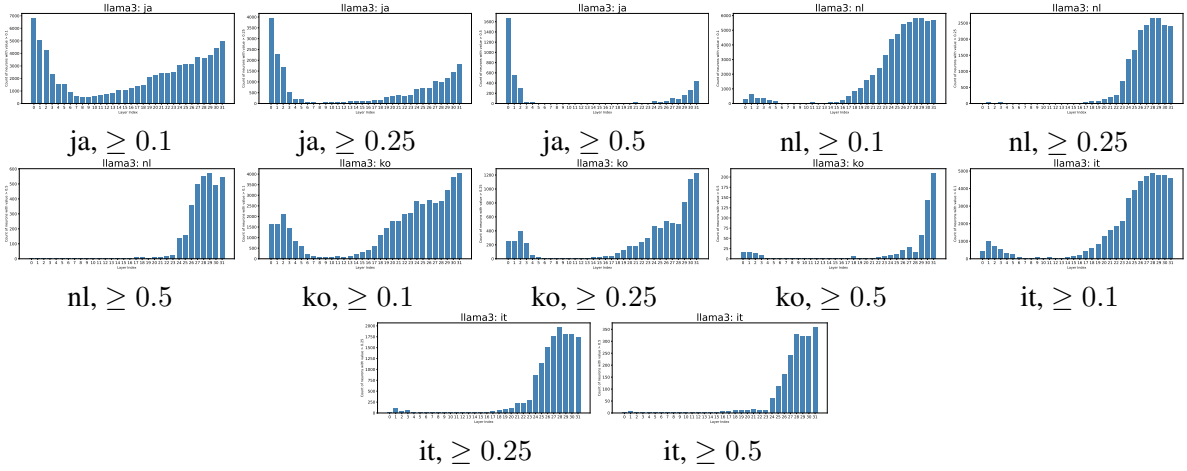


Figure 36: **Distribution of language-specific neurons (LLaMA3-8B).** Each label below the figure indicates the target language and the threshold for correlation ratio. Typically, a correlation ratio above **0.1** indicates a correlation, above **0.25** indicates a moderately strong correlation, and above **0.5** indicates a strong correlation. horizontal-axis denotes the layer indices, and vertical-axis represents the number of neurons.

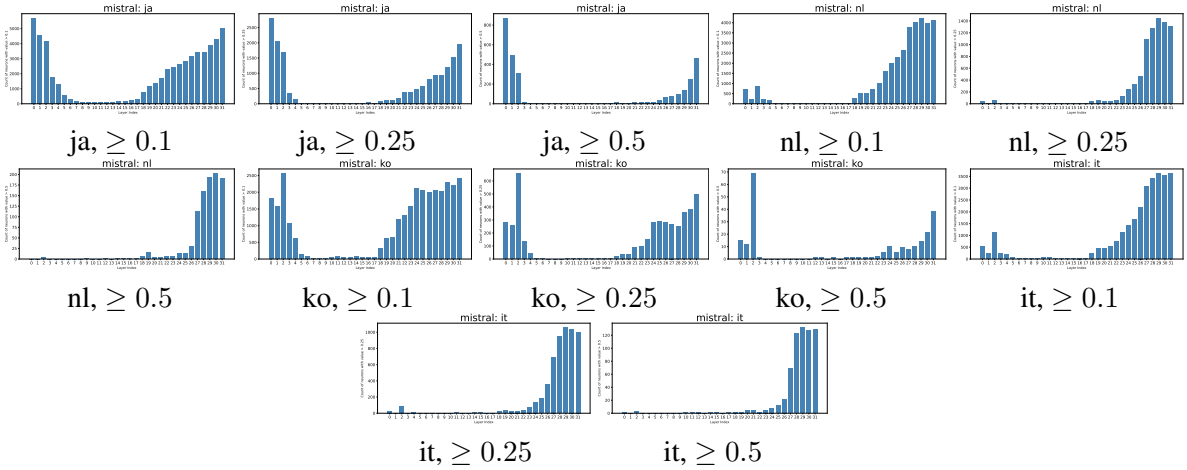


Figure 37: **Distribution of language-specific neurons (Mistral-7B).**

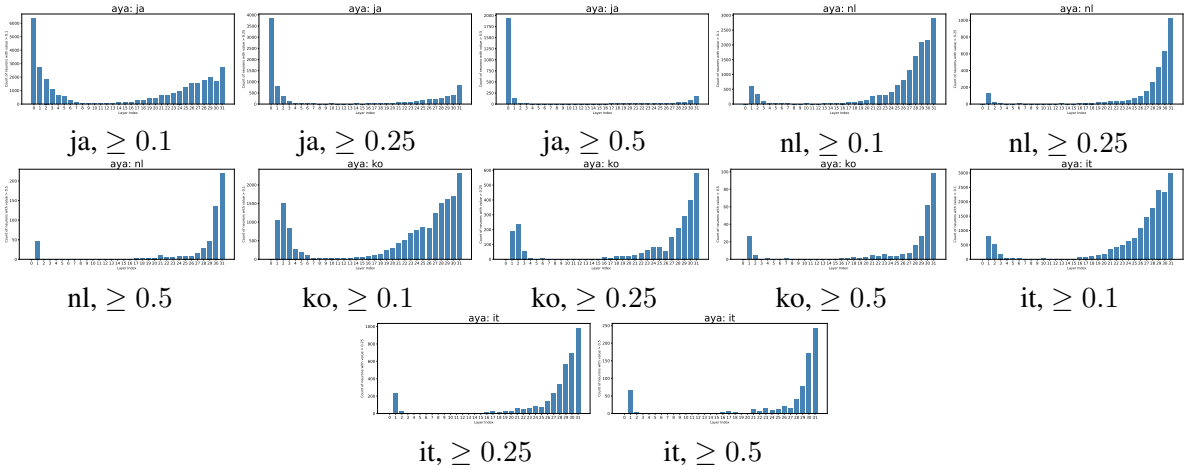
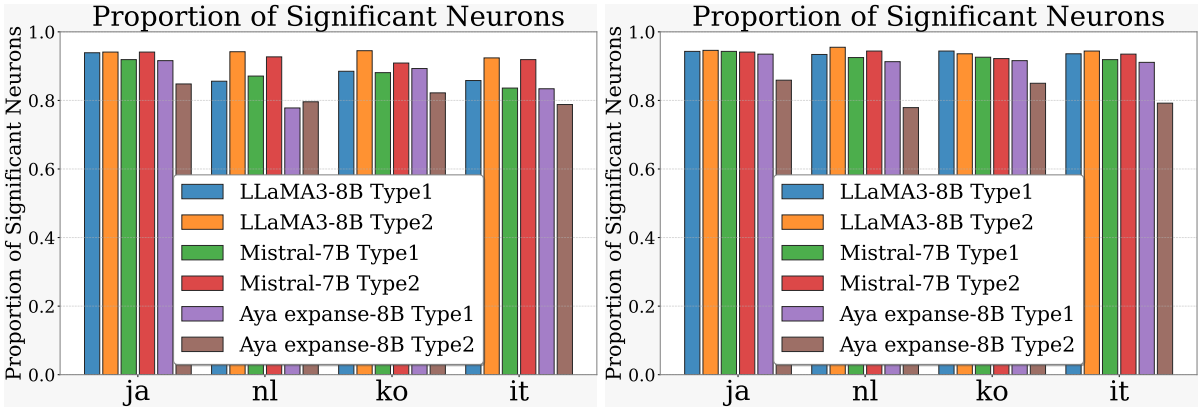


Figure 38: **Distribution of language-specific neurons (Aya expanse-8B).**



(a) Language specificity

(b) Language-family specificity

Figure 39: Proportion of statistically significant neurons among top-1k Transfer Neurons.

Lang	T1 Same	T1: ja	T1: nl	T1: ko	T1: it	T2 Same	T2: ja	T2: nl	T2: ko	T2: it
ja	**-0.021**	–	0.051	0.042	0.053	**0.386**	–	0.777	0.520	0.742
nl	**-0.090**	-0.128	–	-0.092	-0.089	**0.542**	0.710	–	0.744	0.600
ko	**0.055**	0.004	0.085	–	0.087	**0.344**	0.628	0.788	–	0.805
it	**-0.150**	-0.187	-0.150	-0.154	–	**0.520**	0.675	0.586	0.702	–

Table 21: **Hidden states similarity under cross-lingual deactivation across layers (LLaMA3-8B).** T1 refers to Type-1 neurons, while T2 refers to Type-2 neurons. The "Same" columns denote results obtained under non-cross-lingual deactivation, i.e., when the Transfer Neurons correspond to the same language as the input.

Lang	T1 Same	T1: ja	T1: nl	T1: ko	T1: it	T2 Same	T2: ja	T2: nl	T2: ko	T2: it
ja	**0.875**	–	0.879	0.852	0.884	**0.754**	–	0.981	0.912	0.960
nl	**0.845**	0.875	–	0.843	0.864	**0.880**	0.955	–	0.928	0.906
ko	**0.829**	0.856	0.869	–	0.874	**0.805**	0.864	0.949	–	0.897
it	**0.840**	0.857	0.825	0.821	–	**0.901**	0.962	0.919	0.932	–

Table 22: **Hidden states similarity under cross-lingual deactivation across layers (Mistral-7B).**

Lang	T1 Same	T1: ja	T1: nl	T1: ko	T1: it	T2 Same	T2: ja	T2: nl	T2: ko	T2: it
ja	**0.885**	–	0.896	0.890	0.891	**0.699**	–	0.945	0.936	0.947
nl	**0.951**	0.955	–	0.956	0.949	**0.755**	0.922	–	0.909	0.899
ko	**0.923**	0.933	0.930	–	0.928	**0.718**	0.886	0.936	–	0.958
it	**0.946**	0.952	0.947	0.952	–	**0.774**	0.944	0.919	0.926	–

Table 23: **Hidden states similarity under cross-lingual deactivation across layers (Aya expanse-8B).**

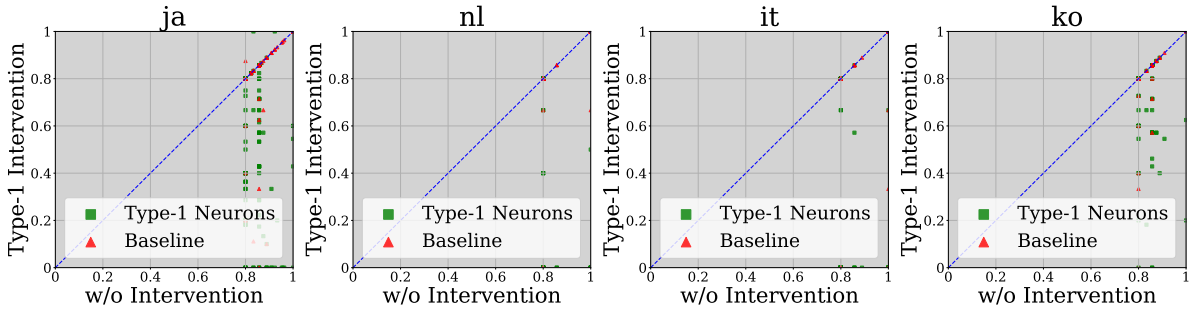


Figure 40: **Token-based F1 score when deactivating top-1k Type-1 transfer neurons (LLaMA3-8B).** Green denotes the score with intervention in the Type-1 neurons, while red represents the score with intervention in the randomly sampled neurons. Points below the $y = x$ line indicate a decrease in the score due to the intervention, whereas points above the $y = x$ line indicate an increase in the score. Points on the $y = x$ line denote no change in the score before and after the intervention.

MKQA (F1)	**ja**	**nl**	**ko**	**it**
(a) normal	0.84	0.83	0.84	0.82
(b) Type-1 Δ	-0.30	-0.41	-0.15	-0.62
(c) baseline Δ	-0.03	-0.09	-0.04	-0.07

Table 24: **Changes in token-based F1 scores for questions with original scores above 0.8 (LLaMA3-8B).**

MKQA (F1)	**ja**	**nl**	**ko**	**it**
(a) normal	0.68	0.72	0.68	0.72
(b) Type-1 Δ	-0.06	-0.13	-0.05	-0.21
(c) baseline Δ	-0.01	-0.01	$\pm$0	-0.02

Table 25: **Changes in token-based F1 scores for questions with original scores above 0.5 (Mistral-7B).**

MKQA (F1)	**ja**	**nl**	**ko**	**it**
(a) normal	0.90	0.96	0.90	0.96
(b) Type-1 Δ	-0.09	-0.06	-0.05	-0.28
(c) baseline Δ	-0.03	-0.02	-0.01	-0.02

Table 26: **Changes in token-based F1 scores for questions with original scores above 0.8 (Mistral-7B).**

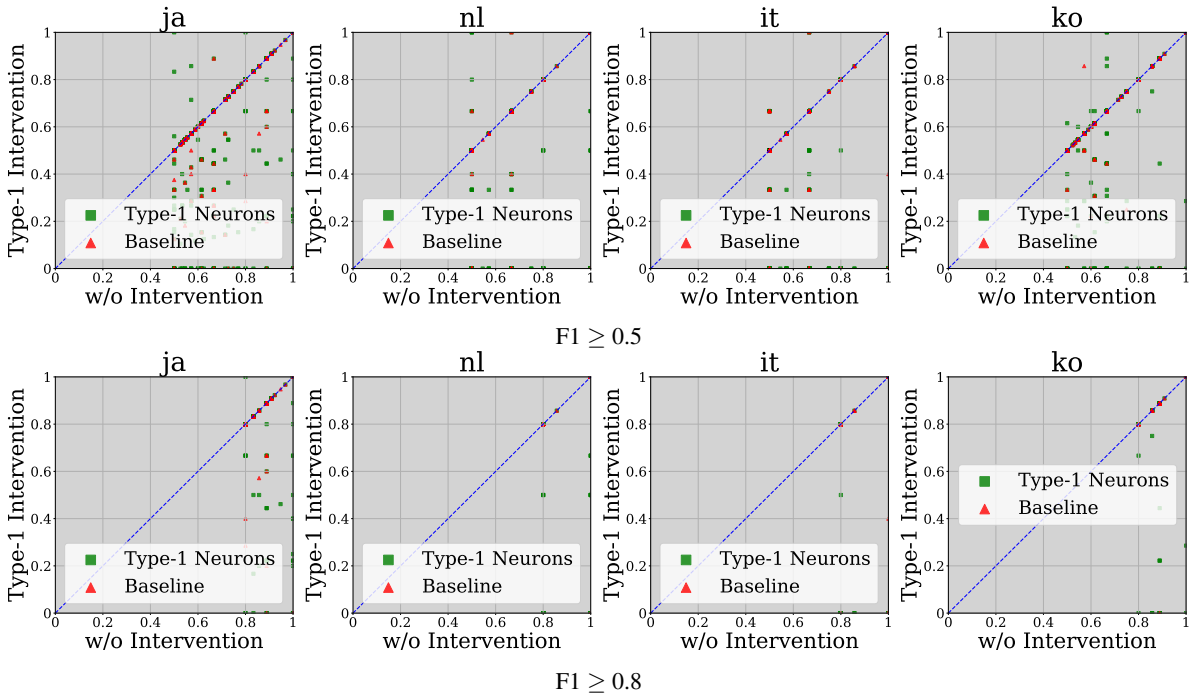


Figure 41: **Token-based F1 score when deactivating top-1k Type-1 transfer neurons (Mistral-7B).**

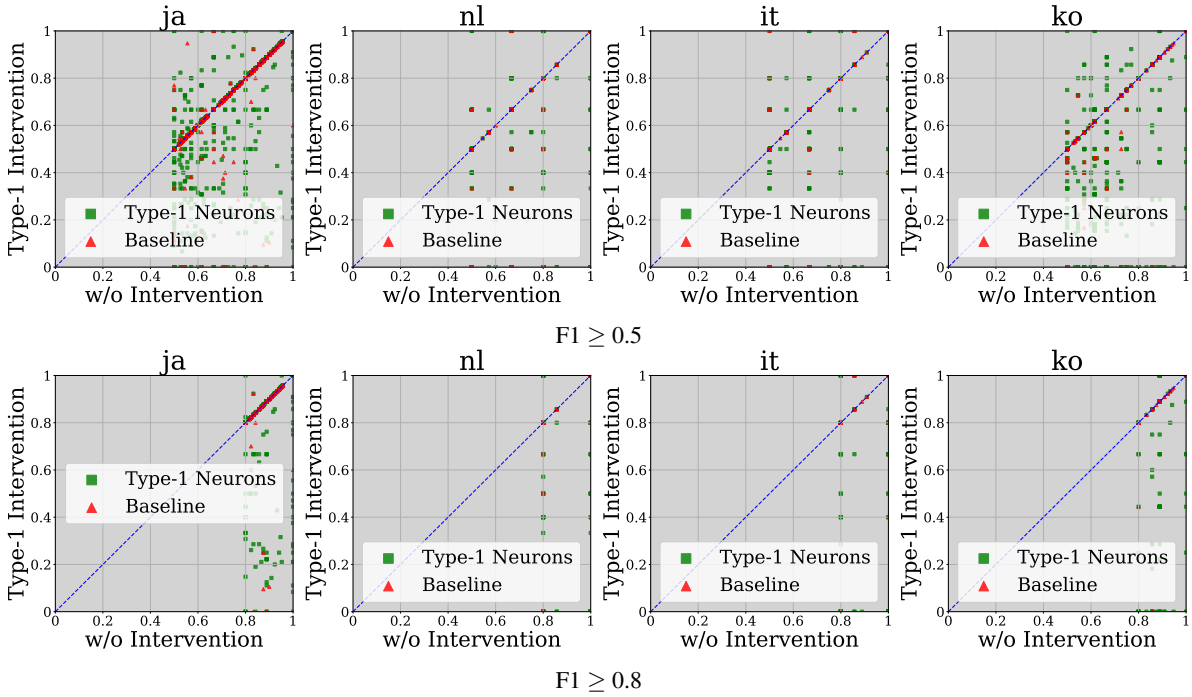


Figure 42: **Token-based F1 score when deactivating top-1k Type-1 transfer neurons (Aya expanse-8B).**

MKQA (F1)	ja	nl	ko	it
(a) normal	0.71	0.72	0.69	0.73
(b) Type-1 Δ	-0.05	-0.16	-0.11	-0.28
(c) baseline Δ	±0	-0.02	±0	-0.01

Table 27: **Changes in token-based F1 scores for questions with original scores above 0.5 (Aya expanse-8B).**

MKQA (F1)	ja	nl	ko	it
(a) normal	0.90	0.93	0.89	0.94
(b) Type-1 Δ	-0.06	-0.19	-0.14	-0.26
(c) baseline Δ	$\pm$0	-0.02	$\pm$0	-0.02

Table 28: **Changes in token-based F1 scores for questions with original scores above 0.8 (Aya expanse-8B).**

Table 29: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (LLaMA3-8B, **Japanese**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0**.	0.1996	0.2033
- biology	0.	0.2873	0.2887
- business	0.	0.2180	0.2294
- chemistry	0.	0.1590	0.1564
- computer_science	0.	0.2756	0.2707
- economics	0.	0.2915	0.3009
- engineering	0.	0.1362	0.1414
- health	0.	0.2052	0.2038
- history	0.	0.1942	0.2152
- law	0.	0.1043	0.1178
- math	0.	0.1887	0.1828
- other	0.	0.1959	0.2045
- philosophy	0.	0.2285	0.2265
- physics	0.	0.1316	0.1355
- psychology	0.	0.3283	0.3308

Table 30: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (LLaMA3-8B, **Korean**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0**.	0.2097	0.2115
- biology	0.	0.2580	0.2524
- business	0.	0.2408	0.2598
- chemistry	0.	0.1793	0.1661
- computer_science	0.	0.2829	0.2561
- economics	0.	0.2867	0.2950
- engineering	0.	0.1754	0.1816
- health	0.	0.1747	0.1921
- history	0.	0.1890	0.2073
- law	0.	0.1283	0.1220
- math	0.	0.2095	0.2250
- other	0.	0.2348	0.2424
- philosophy	0.	0.2084	0.1944
- physics	0.	0.1486	0.1470
- psychology	0.	0.3108	0.2995

Table 31: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (LLaMA3-8B, **French**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0**.	0.2728	0.2705
- biology	0.	0.4170	0.4338
- business	0.	0.2738	0.2801
- chemistry	0.	0.1820	0.1988
- computer_science	0.	0.2683	0.2390
- economics	0.	0.3744	0.3756
- engineering	0.	0.2054	0.2002
- health	0.	0.2780	0.2838
- history	0.	0.3123	0.2992
- law	0.	0.1512	0.1293
- math	0.	0.2494	0.2428
- other	0.	0.3202	0.3139
- philosophy	0.	0.3026	0.2926
- physics	0.	0.2271	0.2248
- psychology	0.	0.4110	0.4085

Table 32: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (Mistral-7B, **Japanese**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0.0717**	0.1712	0.1750
- biology	0.0293	0.2594	0.2664
- business	0.1179	0.2155	0.2231
- chemistry	0.0221	0.1210	0.1193
- computer_science	0.1488	0.2341	0.2366
- economics	0.1197	0.2393	0.2512
- engineering	0.0485	0.1620	0.1465
- health	0.0815	0.1718	0.1659
- history	0.0919	0.1470	0.1575
- law	0.0083	0.0949	0.1116
- math	0.0607	0.1392	0.1540
- other	0.1061	0.1753	0.1742
- philosophy	0.1062	0.1764	0.1723
- physics	0.0370	0.1055	0.070
- psychology	0.1441	0.2820	0.2882

Table 33: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (Mistral-7B, **Korean**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0.0454**	0.1606	0.1593
- biology	0.0014	0.1520	0.1632
- business	0.0963	0.2243	0.2281
- chemistry	0.0053	0.1042	0.0989
- computer_science	0.0829	0.2561	0.2439
- economics	0.1126	0.2192	0.2121
- engineering	0.0444	0.1486	0.1465
- health	0.0757	0.1645	0.1587
- history	0.0367	0.1286	0.1312
- law	0.0083	0.1137	0.0980
- math	0.0200	0.1340	0.1295
- other	0.0714	0.1634	0.1688
- philosophy	0.0641	0.1383	0.1242
- physics	0.0200	0.1162	0.1309
- psychology	0.0677	0.2845	0.2845

Table 34: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (Mistral-7B, **French**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0.1477**	0.2462	0.2460
- biology	0.1311	0.4254	0.4045
- business	0.1584	0.2459	0.2446
- chemistry	0.0972	0.1572	0.1705
- computer_science	0.1488	0.2488	0.2537
- economics	0.1765	0.3436	0.3389
- engineering	0.0939	0.1806	0.1734
- health	0.2227	0.2693	0.2780
- history	0.2283	0.2625	0.2756
- law	0.0667	0.1397	0.1335
- math	0.0888	0.1969	0.1895
- other	0.1861	0.2825	0.2955
- philosophy	0.2204	0.2565	0.2545
- physics	0.1532	0.2009	0.1971
- psychology	0.2531	0.3960	0.4048

Table 35: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (Aya expanse-8B, **Japanese**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0.1685**	0.2636	0.2648
- biology	0.2734	0.4184	0.4017
- business	0.1762	0.2801	0.2842
- chemistry	0.1166	0.2147	0.2138
- computer_science	0.1610	0.2585	0.2610
- economics	0.1339	0.2133	0.2097
- engineering	0.1589	0.2116	0.2260
- health	0.1426	0.2402	0.2402
- history	0.1706	0.2808	0.2756
- law	0.1074	0.1616	0.1814
- math	0.1651	0.3168	0.3101
- other	0.1926	0.2587	0.2554
- philosophy	0.1583	0.2325	0.2305
- physics	0.1463	0.2325	0.2394
- psychology	0.3070	0.4173	0.4148

Table 36: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (Aya expanse-8B, **Korean**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0.1854**	0.2696	0.2702
- biology	0.2148	0.3250	0.3543
- business	0.1952	0.2864	0.2864
- chemistry	0.1466	0.2094	0.2191
- computer_science	0.1244	0.3000	0.30981
- economics	0.2666	0.3318	0.3412
- engineering	0.1476	0.2652	0.2487
- health	0.1587	0.2213	0.2183
- history	0.1496	0.2598	0.2415
- law	0.1241	0.1814	0.1700
- math	0.1858	0.3198	0.3183
- other	0.1991	0.2630	0.2727
- philosophy	0.1743	0.1924	0.1864
- physics	0.1586	0.2232	0.2186
- psychology	0.3434	0.4110	0.4123

Table 37: MMLU-ProX: Type-1 deactivated vs Baseline deactivated vs w/o intervention (Aya expanse-8B, **French**)

Tasks	Type-1 Score	Baseline Score	w/o intervention
OVERALL	**0.2371**	0.3191	0.3207
- biology	0.3501	0.5397	0.5300
- business	0.2383	0.3131	0.3105
- chemistry	0.1590	0.2182	0.2279
- computer_science	0.2220	0.3049	0.3000
- economics	0.3235	0.4218	0.4336
- engineering	0.1827	0.2693	0.2869
- health	0.2606	0.3290	0.3115
- history	0.1444	0.3360	0.3150
- law	0.1470	0.1741	0.18777
- math	0.2865	0.3434	0.3501
- other	0.2522	0.3290	0.3323
- philosophy	0.2064	0.2866	0.2786
- physics	0.1964	0.2594	0.2594
- psychology	0.3446	0.4511	0.4398

Question (Japanese)	プロジェクトの期限に間に合わなかったことを上司に謝罪するメールを作成します。遅延の理由を説明し、同じことが二度と起こらないようにするための手順を説明します。 *Write an email to your supervisor apologizing for not meeting the project deadline. Explain the reasons for the delay and describe the steps that will be taken to ensure that the same issue does not occur again.*
Output without any intervention (Japanese)	いつもお世話になっております。\n 今日は、プロジェクトの期限に間に合わなかったことをお詫び申し上げます。\n その理由は、... *Thank you always for your support.\n Today, I would like to apologize for not meeting the project deadline.\n The reason is, ...*
Output with top-1k Type-2 neurons deactivated (English)	Dear Mr. Smith,\n \n I'm writing to apologize for missing the project's deadline. The reason for the delay was ...

Table 38: **Example of generated text in which the output language changed from the input language after deactivating the top-1k Type-2 neurons (LLaMA3-8B).**

Question (Italian)	Scrivi un'email a un venditore che ha consegnato prodotti errati o danneggiati. Spiega educatamente il problema, fornisci dettagli e richiedi un rimborso o una sostituzione. *Write an email to a seller who has delivered incorrect or damaged products. Politely explain the problem, provide details, and request a refund or a replacement.*
Output without any intervention (Italian)	Grazie per aver contattato il nostro team di supporto.\n\n Siamo molto scioccati per il problema che hai avuto con il nostro prodotto. Siamo dispiaciuti per il disagio che hai subito e siamo pronti a fornire una soluzione. ... *Thank you for contacting our support team. We are very shocked about the issue you experienced with our product. We are sorry for the inconvenience you have suffered, and we are ready to provide a solution. ...*
Output with top-1k Type-2 neurons deactivated (English)	Hello,\n\nI recently purchased a product from your store and unfortunately, it arrived damaged. I'm writing to request a replacement or a refund for the item.\n\nHere are the details of the issue: ...

Table 39: **Example of generated text in which the output language changed from the input language after deactivating the top-1k Type-2 neurons (Mistral-7B).**

Question (Japanese)	気候変動は地球に大きな影響を与えています。気候変動の主な原因は何だと思いますか。また、気候変動は環境にどのような影響を与える可能性がありますか。 *Climate change is having a significant impact on the Earth. What do you think are the main causes of climate change, and what potential impacts might it have on the environment?*
Output without any intervention (Japanese)	気候変動の主な原因は、人間の活動による温室効果ガスの排出です。特に、化石燃料の燃焼（石炭、石油、天然ガスなど）や森林破壊が大きな要因となっています。これらの活動によって二酸化炭素（CO2）、メタン（CH4）、亜酸化窒素（N2O）などの温室効果ガスが大量に放出され、地球の大気中に蓄積しています。... *The main cause of climate change is the emission of greenhouse gases from human activities. In particular, the burning of fossil fuels (such as coal, oil, and natural gas) and deforestation are major factors. These activities release large amounts of greenhouse gases, including carbon dioxide (CO), methane (CH), and nitrous oxide (NO), which accumulate in the Earth's atmosphere. ...*
Output with top-1k Type-2 neurons deactivated (Chinese)	氣候變動的主要原因之一是人類活動，特別是燃燒化石燃料（如煤、石油和天然氣）和森林砍伐。這些活動釋放出大量的室氣體（如二氧化碳和甲），導致地球大氣層中的室效應強，進而導致全球氣上升。... *One of the main causes of climate change is human activity, particularly the burning of fossil fuels (such as coal, oil, and natural gas) and deforestation. These activities release large amounts of greenhouse gases (such as carbon dioxide and methane), which intensify the greenhouse effect in the Earth's atmosphere and, in turn, lead to a rise in global temperatures. ...*

Table 40: **Example of generated text in which the output language changed from the input language after deactivating the top-1k Type-2 neurons (Aya expanse-8B).**

Question (Korean)	인구 과잉은 많은 도시에서 심각한 문제가 되고 있습니다. 인구 과잉의 원인은 무엇이며 도시 생활에 미치는 영향은 무엇입니까? *Overpopulation is becoming a serious problem in many cities. What are the causes of overpopulation, and what effects does it have on urban life?*
Output without any intervention (Korean)	21세기의 많은 도시들은 급속한 인구 증가로 인해 인구 과잉 문제를 겪고 있습니다. 인구 과잉은 특정 지역이나 도시에 너무 많은 사람들이 모여 사는 현상을 말합니다. 이 현상의 원인과 영향에 대해 살펴보겠습니다. ... *Many cities in the 21st century are experiencing problems of overpopulation due to rapid population growth. Overpopulation refers to the phenomenon in which too many people live concentrated in a particular area or city. Let us examine the causes and effects of this phenomenon. ...*
Output with top-1k Type-2 neurons deactivated (Korean and Japanese)	人口過剰, 즉 인구 과잉은 한 지역에서나 국가 전체에서, 그 지역이나 국가의 생태학적 용량을 초과하는 인구 수를 의미합니다. 人口過剰의 원인과 도시 생활에 미치는 영향은 다음과 같습니다.\n\n 1. 人口增加率과 資源有限性: ... *Overpopulation, that is, excessive population, refers to a number of people in a region or an entire country that exceeds the ecological capacity of that region or country. The causes of overpopulation and its effects on urban life are as follows.\n\n 1. Population growth rate and limited resources: ...*

Table 41: **Example of generated text in code-switching–like form in Korean and Japanese after deactivating the top-1k Type-2 neurons (Aya expanse-8B).**