Linguistic Neuron Overlap Patterns to Facilitate Cross-lingual Transfer on Low-resource Languages

Yuemei Xu[1], **Kexin Xu**[1], **Jian Zhou**[1], **Ling Hu**[1], **Lin Gui**[2]
[1] School of Information Science and Technology, Beijing Foreign Studies University
[2] Department of Informatics, King's College London
{xuyuemei, xukexin, bwzj, huling}@bfsu.edu.cn
lin.1.gui@kcl.ac.uk

Abstract

The current Large Language Models (LLMs) face significant challenges in improving their performance on low-resource languages and urgently need data-efficient methods without costly fine-tuning. From the perspective of language-bridge, we propose a simple yet effective method, namely BridgeX-ICL, to improve the zero-shot Cross-lingual In-Context Learning (X-ICL) for low-resource languages. Unlike existing works focusing on language-specific neurons, BridgeX-ICL explores whether sharing neurons can improve cross-lingual performance in LLMs. We construct neuron probe data from the ground-truth MUSE bilingual dictionaries, and define a subset of language overlap neurons accordingly to ensure full activation of these anchored neurons. Subsequently, we propose an HSIC-based metric to quantify LLMs' internal linguistic spectrum based on overlapping neurons, guiding optimal bridge selection. The experiments conducted on 4 cross-lingual tasks and 15 language pairs from 7 diverse families, covering both high-low and moderate-low pairs, validate the effectiveness of BridgeX-ICL and offer empirical insights into the underlying multilingual mechanisms of LLMs. The code is publicly available at https://github.com/xuyuemei/BridgeX-ICL.

1 Introduction

Although Large Language Models (LLMs) have demonstrated impressive multilingual capacities, there is still significant space for improving the performance on low-resource languages (Huang et al., 2024; Al Nazi et al., 2024). To address this issue, especially avoiding costly post-training (Muller et al., 2021; Yong et al., 2023), it is critical to fully investigate the multilingual understanding and transferring ability in LLMs.

Recent research has increasingly focused on data-efficient methods, particularly Cross-lingual In-Context Learning (X-ICL) (Winata et al., 2021; Tanwar et al., 2023; Al Nazi et al., 2024; Cahyawijaya et al., 2024), which surprisingly works well on low-resource languages, likely because LLMs are in-context low-resource language learners (Brown et al., 2020b; Cahyawijaya et al., 2024). For instance, in the Arabic-to-Hebrew Bilingual Lexicon Induction (BLI) task, the zero-shot baseline accuracy in LLaMA 3 is 47.0%. However, simply specifying English as a bridge language in a zero-shot setting boosts accuracy to 64.5%, which even significantly outperforms the two-shot X-ICL. This observation motivates us to further explore: *How can we improve cross-lingual capabilities of LLMs on low-resource languages by selecting an optimal bridge language in X-ICL*? Should the selection be purely data-driven, favoring high-resource bridge languages (Vulic et al., 2020)? Or can human linguistic knowledge, such as language genealogy, or established evolutionary taxonomies, offer a more effective alternative (Stanczak et al., 2022; Wang et al., 2024)?

To systematically investigate this issue, we leverage linguistic neurons (Tang et al., 2024) that handle language features to guide optimal bridge language selection in X-ICL. However, there are two limitations when applying neuron-based interpretation (Cao et al., 2024; Tang et al., 2024; Liu et al., 2024) on low-resource languages:

• **Inaccurate neuron activation.** Current work often relies on multilingual corpora like Wikipedia (Foundation, 2024) to probe internal neurons, without verifying whether LLMs truly understand the multilingual input. This may lead to **unreliable neuron activations**, particularly for low-resource languages. When LLMs poorly under-

This work was supported by the National Social Science Foundation (No.24CYY107), the Fundamental Research Funds for the Central Universities (No.2024TD001), and the National Natural Science Foundation of China (No. 62576120).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27658–27673
November 4-9, 2025

stand the probe input, they may instead activate neurons for processing unfamiliar or noisy input.

• **Lacking guidance for cross-lingual transfer.** Recent work argues that language-specific neurons do not facilitate cross-lingual transfer (Mondal et al., 2025). This raises a critical question: if language-specific neurons cannot, can sharing neurons improve cross-lingual transfer in LLMs? This exploration is also important for transferring language neuron research to actionable strategies to enhance the multilinguality of LLMs.

Motivated by this, we propose a simple yet effective bridge method, BridgeX-ICL, to improve LLMs' cross-lingual capabilities, especially on low-resource languages. To address the **inaccurate activation issue**, we construct probe data by leveraging the ground-truth bilingual lexicon MUSE (Conneau et al., 2017). We collect bilingual word pairs from MUSE that LLMs can translate accurately and use them to prompt the models for bidirectional translations, generating answers in both language directions. To address the **cross-lingual guidance issue**, we first explore overlap neurons' features and their impact on cross-lingual transfer, and then propose a bridge selection strategy based on the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005). Furthermore, we measure the linguistic spectrum in LLMs based on overlapping-neurons and compare it with human language genealogy from Glottolog Trees (Hammarström et al., 2023). We conduct extensive experiments on 4 cross-lingual tasks and 15 language pairs from 7 diverse families. Our main contributions and findings are as follows:

- To the best of our knowledge, this is the first work to explore language-bridge for zero-shot X-ICL to improve LLMs' performance on low-resource languages.
- We construct accurate neuron probe data and use it to fully activate the anchored overlap neurons. We also propose an HSIC-based metric to quantify the similarity between overlapping neurons and specific neurons for optimal bridge selection in X-ICL.
- We validate the generalization of BridgeX-ICL on 4 cross-lingual tasks and 15 language pairs. Here are empirical findings: 1) Strong neural overlaps align with human linguistic taxonomy within language families, but do not consistently hold across families. 2) Overlapping neurons embody shared semantic information, regardless of the language within or between families. 3) BridgeX-ICL improves the performance on cross-lingual tasks of BLI and MRC across 15 language pairs by an average of 6.02% and 5.25% over zero-shot baselines. 4) English is selected as the optimal bridge in 9 out of 15 language pairs in LLaMA (7 out of 15 in Mistral), indicating that high-resource, Latin-script languages tend to be the default bridge. We also find that non-Latin script languages like Chinese also show potential as effective bridges.

2 Related Work

2.1 Cross-lingual In-context Learning

LLMs face significant challenges when applied to low-resource languages (Costa-jussà et al., 2022; Muennighoff et al., 2023; Huang et al., 2024), mainly due to insufficient training data and *the curse of multilinguality* (Conneau et al., 2020). To address these issues without updating model parameters, Cross-lingual In-context Learning (X-ICL), an extension of in-context learning (ICL), has recently gained attention (Brown et al., 2020a). Prior studies (Winata et al., 2021; Tanwar et al., 2023; Al Nazi et al., 2024; Cahyawijaya et al., 2024) have demonstrated that LLMs act as effective few-shot multilingual learners, with few-shot ICL even outperforming fine-tuned language-specific models on several tasks (Winata et al., 2021). However, few-shot X-ICL's performance is highly dependent on the context and the selection of examples, especially for unconventional or ambiguous languages (Philippy et al., 2023; Al Nazi et al., 2024). Consequently, existing research mainly focuses on optimizing few-shot example selection. To the best of our knowledge, we are the first to explore X-ICL explicitly from the perspective of leveraging language bridges.

2.2 Linguistic Neuron in LLMs

Recent research (Stanczak et al., 2022; Tang et al., 2024; Liu et al., 2024; Cao et al., 2024; Wang et al., 2024) has revealed that language-related neurons exist in FFN layers of transformer architecture. Deactivating these neurons will have a vital impact on LLMs' multilingual capacities. Beyond uncovering multilingual mechanisms, some research has gone to explore the neuron pattern across languages (Wang et al., 2024; Stanczak

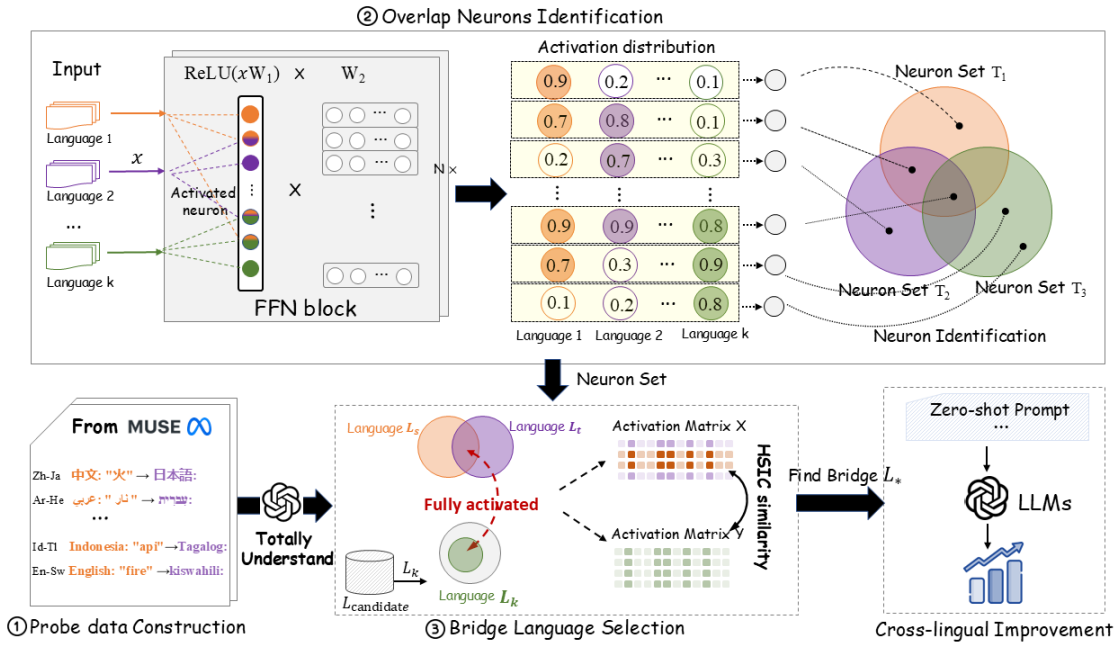


Figure 1. **An illustration of BridgeX-ICL approach**, consisting of three steps: Neuron probe data construction; Language neurons and their overlappings detection; Optimal bridge L_* selection based on HSIC similarity.

et al., 2022) and its impact on cross-lingual performance (Mondal et al., 2025; Zhang et al., 2025). Specifically, Wang et al. (2024) observed that similar languages may not exhibit significant neuron sharing in LLMs like BLOOM, suggesting that neuron sharing does not fully align with language similarity. Furthermore, recent work argues that language-specific neurons do not facilitate cross-lingual transfer (Mondal et al., 2025). This raises a critical question: whether sharing neurons can improve cross-lingual performance in LLMs. Motivated by these findings, we aim to further investigate LLM-internal neuron sharing across languages and its impact. In particular, we define a subset of language-overlapping neurons and explore whether they can serve as internal bridges to support cross-lingual inference.

3 Methodology

3.1 Task Statement

Given a set of languages $\mathcal{L} = \{L_1, ...L_{|\mathcal{L}|}\}$, this work aims to measure the linguistic genealogy implicitly learned by LLMs from language overlapping neurons, then use the quantified linguistic similarity to guide the bridge language selection in X-ICL.

Figure 1 depicts three main steps of our approach: ① neuron probe data construction; ② language neurons and their overlapping detection; ③ bridge language selection, guided by the observed pattern of overlapping neurons and a modified HSIC dependency estimation, which selects the optimal L_* from the candidate set $\mathcal{L}_{\text{candidate}}$ to facilitate X-ICL from a source language L_s to a target language L_t.

3.2 Probe Data Construction

We employ two types of probe data for different purposes of language neurons identification and optimal bridge selection. The former is task independent, targeting language neurons, and can use existing multilingual corpora. In our work, we adopt FLORES+ (NLLB Team et al., 2024), a high-quality parallel corpus released by Meta, and combine its development set and test set to obtain $2,000$ parallel sentences for each language.

Bridge selection for X-ICL needs to consider both language neurons and those that contribute to cross-lingual tasks. Inspired by findings on task-specific neurons (Song et al., 2024), we propose that certain neurons directly influence cross-lingual transfer, and their manipulation and measurement should not rely solely on monolingual corpora. Therefore, we construct probe data by leveraging bilingual word translations as follows.

Probe Data Design. We collect d (i.e., 100) word pairs that LLMs can accurately translate. These word pairs are fed into the LLMs in both directions of $L_1 \rightarrow L_2$ and $L_2 \rightarrow L_1$, ensuring

neurons linked to L_1 and L_2 are fully activated. Instead of feeding word pairs directly, we prompt LLMs to generate translations, which guarantees accurate neuron activation. Examples of probe data for 3 language pairs are shown below.

Examples of probe data

中文: “火” → 日本語:
Indonesia: “api” → Tagalog:
English: “fire” → Kiswahili:

3.3 Linguistic Overlap Neurons

3.3.1 Neurons in LLMs

Neuron identification follows (Tang et al., 2024), which assumes that language neurons are mainly located in the Feed-Forward Network (FFN) layers. Given the transformation at the i-th layer:

$$\boldsymbol{h}_i = \sigma(\tilde{\boldsymbol{h}}_i \boldsymbol{W}_1^i) \cdot \boldsymbol{W}_2^i \tag{1}$$

where $\tilde{\boldsymbol{h}}_i$ is the hidden state input to the i-th layer and $\sigma(\cdot)$ denotes the activation function. $\boldsymbol{W}_1^i \in \mathbb{R}^{d \times N}$ and $\boldsymbol{W}_2^i \in \mathbb{R}^{N \times d}$ are the learned parameters. Here, a neuron is defined as a linear transformation of a single column in $\boldsymbol{W}_1^i$ and there are N neurons in each layer. The activation value of the j-th neuron is $\sigma(\tilde{\boldsymbol{h}}_i \boldsymbol{W}_1^i)_j$. If this value exceeds 0, the neuron is considered an activated neuron.

3.3.2 Overlap Neuron Identification

First, we identify neurons $\mathcal{T}_k$ associated with each language L_k. Unlike existing work (Wang et al., 2024; Mondal et al., 2025) using **L**anguage **A**ctivation **P**robability **E**ntropy (LAPE) (Tang et al., 2024) to identify neurons with high activation probability for one language but low for others, which is less effective to capture neuron relationships across languages, we identify neurons set $\mathcal{T}_k$ based on activation frequency. Let $f_{k,j}$ denote the activation frequency of neuron n_j when processing tokens from language L_k. Neurons with the top $\tau \cdot N$ activation frequencies are selected into $\mathcal{T}_k$ based on a threshold τ.

Overlap Neuron Definition. For languages L_u and L_v, with associated neuron sets $\mathcal{T}_u$ and $\mathcal{T}_v$, the overlap neurons are defined as the interaction of $\mathcal{T}_u$ and $\mathcal{T}_v$. At the i-th FFN layer, we have $\mathcal{T}_{u,v}(i) = \mathcal{T}_u(i) \cap \mathcal{T}_v(i)$.

Linguistic Similarity Calculation. The linguistic similarity between L_u and L_v is quantified through the activation frequencies of their overlapping neurons. Let $\boldsymbol{f_u} = \{f_{u,1}, f_{u,2}, ..., f_{u,|\mathcal{T}_{u,v}|}\}$ denote the activation frequency vector of neurons in $\mathcal{T}_{u,v}$ when processing tokens from L_u. The linguistic similarity between L_u and L_v is calculated:

$$\text{sim}(\mathcal{T}_u, \mathcal{T}_v) = \frac{\boldsymbol{f_u} \cdot \boldsymbol{f_v}}{\|\boldsymbol{f_u}\| \|\boldsymbol{f_v}\|} \tag{2}$$

By computing pairwise similarities for all languages in $\mathcal{L}$, we obtain a comprehensive linguistic spectrum of the model.

3.3.3 Overlap Neuron Pattern

Second, we use the constructed probe data from FLORES+ to explore overlapping neurons' features and their generalized impact on cross-lingual transfer so that we can utilize them to guide bridge language selection. We have two observations:

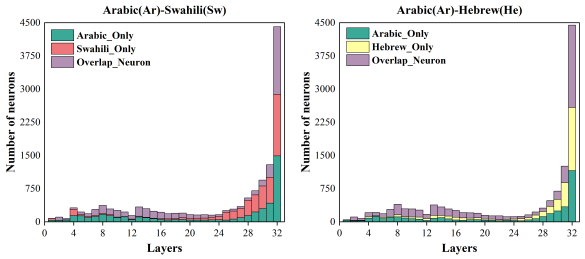


Figure 2. **Language-overlapping neurons** on distant pair (Arabic-Swahili) and close pair (Arabic-Hebrew).

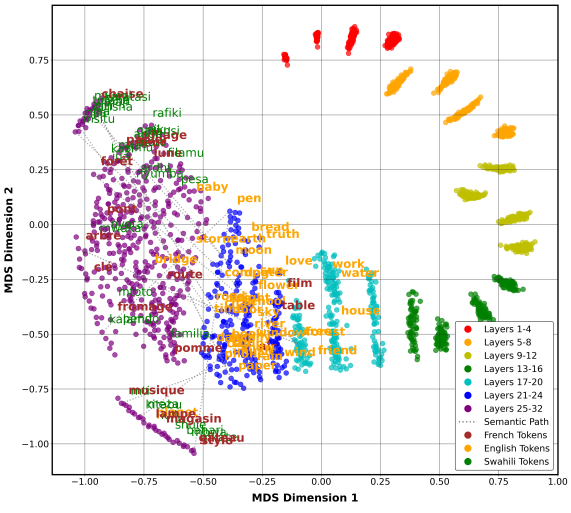


Figure 3. **Layer-wise latent embeddings projected with MDS** in French-Swahili translation. A rainbow-colored path traces the latent embeddings across 32 layers. The predicted Swahili tokens are in green, and their correct English tokens are in orange.

• Similar languages share more neurons than distant ones. For example, Arabic-Hebrew within the

same language family has more overlapping neurons than Arabic-Swahili across families, as presented in Figure 2. This suggests the potential of neural overlap to measure language distance.

• Overlap neurons are predominantly concentrated in the middle and final layers, serving distinct roles of semantic understanding and language decoding for next-token prediction. This neural function is further evidenced by neurons deactivation shown in Figure 7. Specifically, neurons in final layers are task-related and responsible for cross-lingual generation. To examine whether middle-layer neurons handle semantic understanding, we further employ a technique called *logit lens* (Nostalgebraist, 2020) to visualize the latent semantic embeddings across layers. We visualize French-Swahili translation using 60 word pairs that LLaMA 3 translates accurately. We extract the model's latent embeddings at each layer for next-token prediction and project them into a 2D space using classical multidimensional scaling (MDS), presented in Figure 3. The embedding trajectory is marked in a rainbow-colored path (e.g., red = layers 1-4, violet = layers 25-32). We can observe French inputs and their corresponding correct English next tokens cluster in middle layers, indicating that LLMs rely on the knowledge in high-resource languages like English to perform cross-lingual reasoning there. Neurons in middle layers should be prioritized over those in final layers when quantifying language similarity.

3.4 Bridge Language Selection

Based on the above observations, we leverage the constructed probe data with d samples per language pair to identify the optimal bridge language L_* to facilitate X-ICL from source language L_s to target language L_t.

Given L_s and L_t and their overlap neurons $\mathcal{T}_{s,t}$ identified in section 3.3.2, we obtain the activation value matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{T}_{s,t}| \times 2d}$ by prompting the LLM with d samples in both directions for balanced neuron activation in L_s and L_t. We also obtain the activation matrix $\mathbf{Y} \in \mathbb{R}^{|\bar{\mathcal{T}}_y| \times 2d}$ for a candidate bridge language $L_y \in \mathcal{L}_{\text{candidate}}$. Here, $\bar{\mathcal{T}}_y = \mathcal{T}_y - \mathcal{T}_{s,t} - \mathcal{T}_{y'}$ represents the set of language-specific neurons in L_y, excluding those shared with $\mathcal{T}_{s,t}$ or with any other language $L_{y'} \neq L_y$.

We employ HSIC (Gretton et al., 2005) to measure the nonlinear dependency between activation matrices $\mathbf{X}$ and $\mathbf{Y}$. Average pooling will be performed to standardize matrices of $\mathbf{X}$ and $\mathbf{Y}$ to have the same row dimension n. The HSIC is formally calculated as: $\text{HSIC}(\mathbf{X}, \mathbf{Y}) = n^{-2}\text{Tr}(\mathbf{KHLH})$, where $\text{Tr}(\cdot)$ is the trace operation, $\mathbf{K}, \mathbf{L} \in \mathbb{R}^{n \times n}$ are learned kernel matrices for $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{H} = \mathbf{I}_{n \times n} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is a centering matrix, where $\mathbf{I}_{n \times n}$ is the identity matrix of size $n \times n$, $\mathbf{1}_n$ is a vector of n ones. Rather than computing HSIC over the entire activation matrices, we adopt a bidirectional maximum matching strategy that measures the strongest dependency between a single neuron vector $\mathbf{x}_i \in \mathbf{X}$ ($\mathbf{y}_j \in \mathbf{Y}$) and the entire distribution of the other, where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$, computed as:

$$\text{H}(\bar{\mathcal{T}}_y, \mathcal{T}_{s,t}) = \frac{1}{2}\Big(\max_i \text{HSIC}(\mathbf{x}_i, \mathbf{Y}) + \max_j \text{HSIC}(\mathbf{X}, \mathbf{y}_j)\Big) \tag{3}$$

We compute the dependency scores layer by layer and average them across the middle K layers to estimate the selection probability of L_y:

$$p(L_y | L_s \to L_t) = \frac{1}{K}\sum_{i=1}^{K} \text{H}\left(\bar{\mathcal{T}}_y(i), \mathcal{T}_{s,t}(i)\right) \tag{4}$$

where K is determined according to embedding semantic similarity and discussed in section C.3. Finally, the optimal bridge L^* is selected by:

$$L_* = \arg \max_{L_y \in \mathcal{L}_{\text{candidate}}} p(L_y | L_s \to L_t) \tag{5}$$

4 Experiment

4.1 Experiment Setup

Implementation. We evaluate BridgeX-ICL on 4 cross-lingual tasks and 15 languages covering 7 diverse language families: **Indo-European**: English (En), German (De), French (Fr), Italian (It), Portuguese (Pt), Spanish (Es); **Uralic**: Finnish (Fi), Hungarian (Hu); **Afro-Asiatic**: Arabic (Ar), Hebrew (He); **Austronesian**: Indonesian (Id), Tagalog (Tl); **Sino-Tibetan**: Chinese (Zh); **Japonic**: Japanese (Ja); **Niger-Congo**: Swahili (Sw).

As the evaluation focuses on LLMs' cross-lingual transfer on low-resource languages, we take He, Tl, Sw, and Ja as target languages to build 15 cross-lingual pairs, covering moderate-to-low (e.g., Ar-Sw) and high-to-low (e.g., En-He) pairs, both within and across language families. The classification of high-, moderate-, and low-resource languages is based on their proportion in LLMs' training corpora, following previous work (Cieri et al., 2016). Since the bridge language should be well supported by LLMs, we select 6

languages in the Indo-European family as candidate bridges. We also conducted an exploratory experiment using bridges not in the Indo-European family discussed in section 5.1.

Datasets. To evaluate the generalization of bridge selection beyond the BLI task, we further evaluate cross-lingual Machine Reading Comprehension (MRC) using the Belebele dataset (Bandarkar et al., 2024). To verify robustness, we additionally consider two cross-lingual tasks: Cross-Lingual Question Answering (CLQA) and Cross-Lingual Natural Language Inference (XNLI). Since CLQA and XNLI cover only six of our evaluated language pairs, their experimental results are reported in Appendix B.

To evaluate low-resource languages, a key challenge lies in the lack of evaluation benchmarks. Although the ground-truth MUSE (Conneau et al., 2017) provides 110 bilingual dictionaries for the BLI task, it does not cover the 15 language pairs we tested. To address this, we used English as a pivot to build L_s-L_t dictionary from L_s-English and L_t-English. For languages not in MUSE (e.g., Swahili), we extracted word pairs from wiktionary_bli (Izbicki, 2022) to build En-Sw. We verified all word pairs using both Google and Microsoft translators to ensure quality and selected $1,000$ word pairs for each language pair that are consistently validated by both systems [1].

Metrics. For the BLI task, we use the Precision@N metric, which measures the accuracy of the model's top-N candidate translations. In this study, N is set to 1. For the MRC task, we use accuracy to evaluate whether the model selects the correct answer from multiple choices.

LLMs. We conducted experiments on two open-source LLMs: LLaMA-3-8B (Dubey et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). Their training corpora cover 176 and 53 languages, respectively, which include all the experimental low-resource languages and allow us to explore the underlying linguistic mechanisms.

Baselines. Baselines are divided into *zero-shot*, *few-shot*, and *zero-shot with bridge*. Specifically, *zero-shot* is the basic prompt setup, and *few-shot* builds on the zero-shot prompt by adding 1, 2, 3, or 4 samples. For zero-shot with bridge approach, we compare BridgeX-ICL with 5 baselines described below. 1) **Phylogenetic Distance Source/-Target** (Ph.D Source or Ph.D Target): Select the bridge language closest to the source or target language according to language genealogy of Glottolog Trees (Hammarström et al., 2023). 2) **English Bridge**: Use English as the bridge language. 3) **Sharing Matters**: Wang et al.(2024) used activation values to find shared neurons across languages. We select language with the most shared neurons as the bridge language. 4) **IoU**: Use Intersection over Union (IoU) (Tan et al., 2024), also known as Jaccard index, to measure linguistic distance. Given neuron sets $\mathcal{T}_u, \mathcal{T}_v$ associated with language L_u, L_v, $\text{IoU}(\mathcal{T}_u, \mathcal{T}_v) = |\mathcal{T}_u \cap \mathcal{T}_v| / |\mathcal{T}_u \cup \mathcal{T}_v|$. Language with the highest average IoU score to L_s and L_t is selected. 5) **LAPE_overlap**: Use entropy-based LAPE (Tang et al., 2024) to identify language-specific neurons. We then compute cosine similarity on overlap neurons between the bridge and source/target. The language with the highest average similarity is selected.

		Sino-Tibetan	Japonic	Afro-Asiatic		Austronesian		Uralic		Niger-Congo	Indo-European					
		zh	ja	ar	he	id	tl	fi	hu	sw	en	de	fr	it	pt	es
Sino-Tibetan	zh	1.000	0.644	0.487	0.189	0.367	0.271	0.201	0.281	0.024	0.505	0.420	0.408	0.302	0.387	0.444
Japonic	ja	0.644	1.000	0.492	0.229	0.333	0.236	0.253	0.293	0.031	0.446	0.453	0.373	0.346	0.373	0.366
Afro-Asiatic	ar	0.487	0.492	1.000	0.470	0.421	0.333	0.311	0.283	0.161	0.376	0.437	0.479	0.458	0.470	0.451
	he	0.189	0.229	0.470	1.000	0.199	0.215	0.197	0.211	0.137	0.047	0.265	0.274	0.300	0.271	0.243
Austronesian	id	0.367	0.333	0.421	0.199	1.000	0.428	0.350	0.309	0.298	0.368	0.413	0.350	0.379	0.424	0.371
	tl	0.271	0.236	0.333	0.215	0.428	1.000	0.258	0.232	0.371	0.219	0.251	0.297	0.302	0.374	0.359
Uralic	fi	0.201	0.253	0.311	0.197	0.350	0.258	1.000	0.414	0.181	0.217	0.383	0.291	0.320	0.302	0.249
	hu	0.281	0.293	0.283	0.211	0.309	0.232	0.414	1.000	0.111	0.292	0.430	0.362	0.355	0.374	0.341
Niger-Congo	sw	0.024	0.031	0.161	0.137	0.298	0.371	0.181	0.111	1.000	0.029	0.119	0.024	0.069	0.092	0.000
Indo-European	en	0.505	0.446	0.376	0.047	0.368	0.219	0.217	0.292	0.029	1.000	0.469	0.518	0.461	0.538	0.525
	de	0.420	0.453	0.437	0.265	0.413	0.251	0.383	0.430	0.119	0.469	1.000	0.549	0.548	0.547	0.502
	fr	0.408	0.373	0.479	0.274	0.350	0.297	0.291	0.362	0.024	0.518	0.549	1.000	0.734	0.720	0.726
	it	0.302	0.346	0.458	0.300	0.379	0.302	0.320	0.355	0.069	0.461	0.548	0.734	1.000	0.760	0.706
	pt	0.387	0.373	0.470	0.271	0.424	0.374	0.302	0.374	0.092	0.538	0.547	0.720	0.760	1.000	0.838
	es	0.444	0.366	0.451	0.243	0.371	0.359	0.249	0.341	0.000	0.525	0.502	0.726	0.706	0.838	1.000

(a) Linguistic spectrum in LLaMA 3

		Sino-Tibetan	Japonic	Afro-Asiatic		Austronesian		Uralic		Niger-Congo	Indo-European					
		zh	ja	ar	he	id	tl	fi	hu	sw	en	de	fr	it	pt	es
Sino-Tibetan	zh	1.000	0.314	0.150	0.150	0.183	0.183	0.479	0.479	0.126	0.075	0.076	0.051	0.084	0.051	0.062
Japonic	ja	0.314	1.000	0.436	0.436	0.534	0.531	0.619	0.619	0.367	0.218	0.220	0.149	0.244	0.149	0.178
Afro-Asiatic	ar	0.150	0.436	1.000	0.914	0.707	0.702	0.295	0.295	0.696	0.413	0.417	0.280	0.463	0.280	0.336
	he	0.150	0.436	0.914	1.000	0.707	0.702	0.295	0.295	0.696	0.413	0.417	0.280	0.463	0.280	0.336
Austronesian	id	0.183	0.534	0.707	0.707	1.000	0.890	0.361	0.361	0.594	0.354	0.357	0.241	0.397	0.241	0.289
	tl	0.183	0.531	0.702	0.702	0.890	1.000	0.360	0.360	0.590	0.351	0.354	0.238	0.393	0.238	0.286
Uralic	fi	0.479	0.619	0.295	0.295	0.361	0.360	1.000	0.943	0.248	0.148	0.149	0.101	0.166	0.101	0.121
	hu	0.479	0.619	0.295	0.295	0.361	0.360	0.943	1.000	0.248	0.148	0.149	0.101	0.166	0.101	0.121
Niger-Congo	sw	0.126	0.367	0.696	0.696	0.594	0.590	0.248	0.248	1.000	0.470	0.475	0.318	0.528	0.318	0.382
Indo-European	en	0.075	0.218	0.413	0.413	0.354	0.351	0.148	0.148	0.470	1.000	0.806	0.482	0.662	0.482	0.580
	de	0.076	0.220	0.417	0.417	0.357	0.354	0.149	0.149	0.475	0.806	1.000	0.489	0.669	0.489	0.588
	fr	0.051	0.149	0.280	0.280	0.241	0.238	0.101	0.101	0.318	0.482	0.489	1.000	0.562	0.829	0.743
	it	0.084	0.244	0.463	0.463	0.397	0.393	0.166	0.166	0.528	0.662	0.669	0.562	1.000	0.562	0.670
	pt	0.051	0.149	0.280	0.280	0.241	0.238	0.101	0.101	0.318	0.482	0.489	0.829	0.562	1.000	0.792
	es	0.062	0.178	0.336	0.336	0.289	0.286	0.121	0.121	0.382	0.580	0.588	0.743	0.670	0.792	1.000

(b) Language similarity from Glottolog Trees

Figure 4. **Comparison of linguistic spectrum** calculated based on overlapping neurons in LLaMA 3 and language similarity derived from Glottolog Phylogenetic Trees, including 15 languages from 7 families. Darker blue indicates a higher language similarity.

[1]The constructed BLI dictionaries are available at: https://github.com/xuyuemei/BLI-.

4.2 Main Results

4.2.1 LLMs' Linguistic Spectrum Discussion

This section discusses the linguistic similarities across 15 languages from 7 families, calculated based on overlapping neurons in LLaMA 3 and Mistral, as presented in Figure 4(a) and Figure 8 in Appendix C.2, respectively. To evaluate how closely the linguistic spectrum learned by LLMs aligns with that of human languages, we leverage Glottolog Phylogenetic Trees (Hammarström et al., 2023), which encode hierarchical relationships among 8,000+ human languages, to derive human language similarity in Figure 4(b). The detailed process for computing linguistic similarity based on Glottolog Trees is in Appendix A. In Figure 4, darker blue indicates stronger similarity, and the diagonal denotes self-similarity (1.0).

Linguistic spectrum learned by LLMs is not fully aligned with human language phylogeny. We observe strong neural similarities within language families, highlighted by a red text box in Figure 4, which matches human linguistic taxonomy. For example, Arabic (Ar) and Hebrew (He), within the Afro-Asiatic family, exhibit a high neuron overlap (0.470), greater than Arabic-Swahili with 0.161 similarity. A similar pattern appears with Indonesian (Id) and Tagalog (Tl), both from the Austronesian family. In addition, high-resource Indo-European languages, such as French-Italian (Fr-It) and Portuguese-Spanish (Pt-Es), show the highest overlap scores, with the darkest blue in the bottom-right corner of the heatmap. But this alignment breaks down in high-to-low resource language pairs, and some cross-family pairs display unexpected high similarity scores, likely reflecting training data distribution rather than intrinsic linguistic relationships.

LLMs build their own distinct understanding of language relationships. The calculated linguistic spectra of LLaMA 3 and Mistral are similar but not identical. The two models may choose different bridges for the same language pair, as discussed later. We observe that Arabic has the strongest similarity (0.479) with French in the Romance family, rather than with Hebrew (0.470) from its own Afro-Asiatic family. This counterintuitive result is likely due to the linguistic relationships learned by LLMs being primarily shaped by the distribution of languages in training corpora, as noted in (Philippy et al., 2023).

4.2.2 Cross-lingual Results Analysis

Table 1 compares the performance of BridgeX-ICL against various baselines on the BLI task across 15 language pairs.

We observe the following findings: **1) LLMs exhibit poor and imbalanced performance on low-resource languages.** For example, LLaMA 3 achieves its best BLI performance of 69.90 on the Ar-Ja pair, but its worst of 25.90 on the Id-Sw pair. **2) LLMs are few-shot multilingual learners**, as also noted in (Al Nazi et al., 2024). However, few-shot X-ICL does not consistently yield stable gains. For example, one-shot sometimes performs worse than zero-shot, and performance often stabilizes or may decline once the number of shots exceeds 3. This suggests that when applying few-shot X-ICL, 3-shots will be enough. **3) Zero-shot with bridge is a simple yet data-efficient strategy for low-resource languages**. BridgeX-ICL finds 9 optimal bridges out of 15 language pairs, achieving average performance of two-shot X-ICL across all pairs, followed by the English-bridge method. While phylogenetic distance source/target methods are the least effective. It seems using English as the default bridge is cost-effective, which will be discussed in section 5.2.

5 Discussion

5.1 Candidate Bridge Language Selection

To validate the rationale for selecting Indo-European languages as bridge candidates, Table 2 presents an exploratory experiment in which all languages in our study were evaluated as potential bridges, using 6 representative language pairs.

It shows that Indo-European languages on average outperform the nine non-Indo-European languages. Interestingly, some non-Latin-script languages, such as Chinese, also demonstrate potential as effective bridges. The results provided valuable insights: only languages well supported by LLMs functioned as effective bridges. Based on this observation, we selected 6 languages in Indo-European as the final candidates in our study.

5.2 Application of Bridge Language

Beyond the BLI task, Table 3 evaluates BridgeX-ICL on the MRC cross-lingual task. The prompt for zero-shot with bridge in MRC is detailed in Appendix F. Appendix B further evaluates BridgeX-ICL on CLQA and XNLI cross-lingual tasks. Results show our approach works well and benefits

Table 1. Comparison of BLI task improvement on 15 language pairs. The highest gains are marked with **bold** in few-shot and zero-shot with bridge methods. '-' indicates the selected bridge is either the source or target language.

LLaMA-3-8B																
Method		**Zh-Ja**	**Zh-He**	**Zh-Tl**	**Zh-Sw**	**Ar-Ja**	**Ar-He**	**Ar-Tl**	**Ar-Sw**	**Id-Ja**	**Id-He**	**Id-Tl**	**Id-Sw**	**En-He**	**En-Tl**	**En-Sw**
Zero-shot		67.10	44.10	42.60	31.20	69.90	47.00	46.70	39.10	62.50	44.70	49.30	25.90	56.90	60.00	28.80
Few-shot	One-shot	+3.20	+12.60	+1.20	+3.00	+4.20	+13.50	-4.80	+0.60	+7.40	+7.50	+0.70	+6.00	+18.60	-6.30	+4.40
	Two-shot	+9.40	+16.90	+2.20	+5.10	**+9.40**	+13.90	-0.30	+3.50	+16.00	+15.90	+5.90	**+6.10**	+23.90	-5.50	+6.50
	Three-shot	+6.70	**+22.20**	**+3.70**	+6.80	+7.80	**+16.90**	**+1.50**	**+3.70**	**+16.70**	**+22.90**	+10.10	-4.30	**+26.30**	-4.00	+6.90
	Four-shot	**+12.50**	+20.50	+3.20	**+7.00**	+7.70	+15.80	+0.80	+3.20	+14.30	+22.00	**+10.60**	-6.00	+26.10	**-3.70**	**+7.40**
Zero-shot with bridge	Ph.D Source	+2.80	+6.70	+3.80	+1.50	-9.50	-	-7.40	-0.60	-11.90	-3.20	-	-3.90	+12.30	-3.30	-1.70
	Ph.D Target	+2.80	+9.60	-0.10	**+5.00**	-0.50	-	-7.10	-	-12.10	+3.80	-	+1.20	**+16.40**	-2.30	**+2.30**
	English Bridge	**+10.80**	+12.60	**+11.40**	+4.80	**+10.70**	**+17.50**	**+10.10**	+3.30	**+3.60**	+9.30	+10.40	+2.50	-	-	-
	Sharing Matters	+9.50	**+14.50**	+8.40	+2.60	+6.40	+17.10	+6.10	**+3.90**	+2.60	+15.20	+7.70	+2.40	+12.30	-3.30	-1.70
	IoU Score	**+10.80**	+12.60	**+11.40**	+4.80	**+10.70**	+13.20	+5.30	+3.50	+2.30	+15.20	+7.40	+3.80	+9.70	-6.10	-1.80
	LAPE_overlap	+10.50	+13.90	**+11.40**	+3.00	**+10.70**	**+17.50**	**+10.10**	+3.30	**+3.60**	+9.30	+10.40	+2.50	+12.30	-6.10	-1.80
	Ours	**+10.80**	+12.60	**+11.40**	+4.80	**+10.70**	**+17.50**	**+10.10**	+3.30	**+3.60**	**+16.60**	**+11.70**	**+4.10**	+14.90	**-1.30**	-2.60
Mistral-7B																
Method		**Zh-Ja**	**Zh-He**	**Zh-Tl**	**Zh-Sw**	**Ar-Ja**	**Ar-He**	**Ar-Tl**	**Ar-Sw**	**Id-Ja**	**Id-He**	**Id-Tl**	**Id-Sw**	**En-He**	**En-Tl**	**En-Sw**
Zero-shot		57.80	26.20	34.10	8.40	52.50	32.30	28.00	9.10	48.40	36.20	40.60	8.60	47.80	45.70	8.20
Few-shot	One-shot	-7.20	-0.40	-1.90	+1.60	-8.10	-7.40	-2.00	+2.10	+5.60	+2.50	-3.70	+1.20	+0.40	-13.80	+2.40
	Two-shot	+0.40	+0.80	-2.90	+2.10	-5.10	-7.00	-0.40	+1.60	+10.10	+3.20	-0.60	+1.60	+0.50	-3.90	**+3.20**
	Three-shot	+3.00	+0.70	-1.60	+2.20	**-2.90**	**-6.50**	**+0.20**	+2.00	**+10.70**	+3.30	+0.80	+2.00	0.00	-2.80	+3.00
	Four-shot	**+3.20**	**+1.20**	**-0.50**	**+2.60**	**-2.90**	-6.60	0.00	**+2.20**	**+10.70**	**+4.10**	**+2.50**	**+2.60**	**+1.20**	**-2.00**	+2.50
Zero-shot with bridge	Ph.D Source	-3.60	-0.20	+0.90	+0.90	-8.70	-	-2.20	+0.80	-5.70	-5.70	-	-0.40	**-0.10**	+2.30	+2.10
	Ph.D Target	-3.60	-1.60	-0.60	+0.20	-9.50	-	+4.90	-	-4.70	-7.20	-	-0.20	-9.00	+1.40	+1.20
	English Bridge	**+7.90**	**+8.60**	**+6.40**	+1.20	**+8.90**	**+2.70**	**+8.60**	+1.20	**+9.60**	**+2.20**	+2.50	+0.20	-	-	-
	Sharing Matters	**+7.90**	**+8.60**	**+6.40**	+1.20	+4.60	**+2.70**	**+8.60**	+1.20	**+9.60**	**+2.20**	+2.50	+0.20	**-0.10**	+1.50	+0.90
	IoU Score	+0.50	+5.30	+4.90	**+2.20**	+5.90	+1.10	+7.50	**+1.60**	+2.40	-2.80	-0.50	+1.00	-1.00	+1.50	**+2.20**
	LAPE_overlap	+0.50	+3.50	+2.20	**+2.20**	**+8.90**	**+2.70**	+7.60	+1.20	**+9.60**	**+2.20**	+1.00	+0.20	**-0.10**	**+4.50**	+2.00
	Ours	**+7.90**	**+8.60**	**+6.40**	+1.20	**+8.90**	**+2.70**	**+8.60**	+1.20	+2.40	-3.30	**+3.30**	**+1.00**	**-0.10**	+1.50	+2.00

Table 2. Candidate bridge language selection comparing Indo-European languages with 9 non-Indo-European languages. '-' indicates the selected bridge is either the source or target language.

	Zero-shot	**Zh**	**Ja**	**Ar**	**He**	**Id**	**Tl**	**Fi**	**Hu**	**Sw**	**Indo-European (Avg.)**
Zh-Ja	67.10	-	-	+3.30	+0.80	0	-5.10	+2.80	+5.20	-9.90	**+9.43**
Zh-He	44.10	-	+7.90	+9.60	-	+4.80	-0.20	+6.70	+10.80	-8.80	**+14.37**
Ar-Tl	46.70	+4.50	+3.10	-	-7.40	-7.10	-	+3.40	+4.90	-6.10	**+6.88**
Ar-He	47.00	+14.90	+11.70	-	-	-5.20	+1.40	+11.60	+15.30	-3.00	**+15.72**
Id-Ja	62.50	-2.40	-	-9.00	-5.20	-	-11.90	-12.10	-9.60	-24.80	**+1.40**
Id-Sw	25.90	+3.20	+1.50	+1.20	-0.80	-	-3.90	-0.50	+1.30	-	**+3.43**

more from LLaMA 3 than from Mistral. For example, BridgeX-ICL improves the performance of LLaMA 3 by an average of 6.03% over the zero-shot baseline across 15 language pairs, while the average improvement on Mistral is 4.48%.

English is selected as the optimal bridge in 9 out of 15 language pairs in LLaMA 3 (6 out of 15 in Mistral). This is partly due to the unbalanced language abilities of LLMs across 5 candidate Indo-European bridges. As discussed in Figure 3, another key factor is the model's inherent preference for English-pivot during cross-lingual transfer.

5.3 Ablation Study

In this part, we conduct ablation study to evaluate the impact of the constructed neuron probe data and the proposed HSIC similarity metric on bridge selection, using the BLI task as an example.

Table 4 presents the ablation results by comparing "w/o $*$" with our constructed probe data,

Table 3. Evaluation on MRC cross-lingual task. Red color highlights the different bridge selections, and **bold** marks the highest gains at each language pair.

	LLaMA-3-8B			Mistral-7B		
	Bridge	Zero-shot	Ours	Bridge	Zero-shot	Ours
Zh-Ja	En	61.80	-0.40	En	66.20	**+6.20**
Zh-He	En	56.00	+7.20	En	50.20	**+10.20**
Zh-Tl	En	57.20	+3.00	En	60.60	**+7.80**
Zh-Sw	En	48.60	**+10.40**	Pt	43.00	+1.00
Ar-Ja	En	52.20	+3.20	En	48.40	**+7.40**
Ar-He	En	51.20	+4.00	En	42.00	**+5.20**
Ar-Tl	En	46.40	**+7.20**	En	47.60	+4.60
Ar-Sw	En	40.60	**+12.60**	En	30.40	+9.80
Id-Ja	En	56.20	**+8.20**	Pt	63.40	+2.00
Id-He	Es	58.20	**+7.00**	Es	45.80	+6.40
Id-Tl	Fr	58.40	**+4.20**	De	55.60	+3.00
Id-Sw	Fr	49.80	**+5.20**	Pt	39.60	-2.80
En-He	Es	70.60	**+8.00**	Es	61.20	+3.80
En-Tl	Fr	72.80	**+3.20**	Es	72.20	+2.40
En-Sw	Fr	64.60	**+7.40**	Fr	51.20	+0.20

where "w/o $*$" denotes replacing the constructed probe data with the bilingual tokens extracted from FLORES+ (NLLB Team et al., 2024). The detailed construction of "w/o $*$" is provided in Ap-

Table 4. The impact of neuron probe data. 'w/o ∗' denotes replacing our constructed probe data with bilingual tokens extracted from the FLORES+ dataset.

	LLaMA-3-8B		Mistral-7B	
	w/o ∗	Ours	w/o ∗	Ours
Zh-Ja	76.40	77.90 ↑	63.40	65.70 ↑
Zh-He	56.70	56.70 –	32.70	34.80 ↑
Zh-Tl	51.10	54.00 ↑	38.20	40.50 ↑
Zh-Sw	36.40	36.00 ↓	10.60	9.60 ↓
Ar-Ja	77.80	80.60 ↑	58.40	61.40 ↑
Ar-He	64.10	64.50 ↑	33.40	35.00 ↑
Ar-Tl	52.00	56.80 ↑	33.30	35.50 ↑
Ar-Sw	41.60	42.40 ↑	10.70	10.70 –
Id-Ja	65.10	66.10 ↑	50.80	58.00 ↑
Id-He	59.90	61.30 ↑	33.40	38.40 ↑
Id-Tl	57.00	61.00 ↑	40.20	40.20 –
Id-Sw	28.30	30.00 ↑	9.20	9.20 –
En-He	66.60	71.80 ↑	47.70	47.70 –
En-Tl	58.70	58.70 –	47.70	48.00 ↑
En-Sw	26.20	26.20 –	10.20	10.30 ↑

pendix D. The results highlight the crucial impact of probe data on effective neuron manipulation. Furthermore, Table 6 in Appendix D compares HSIC with Cosine similarity, showing that HSIC better captures the dependency between language-overlapping neurons and specific neurons.

5.4 Overlapping Neuron Distribution

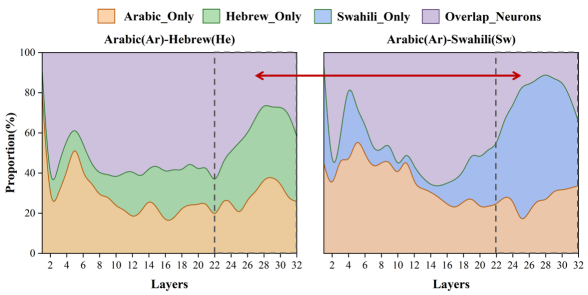


Figure 5. **Distribution of overlap neurons** in language pairs within and across families in LLaMA 3.

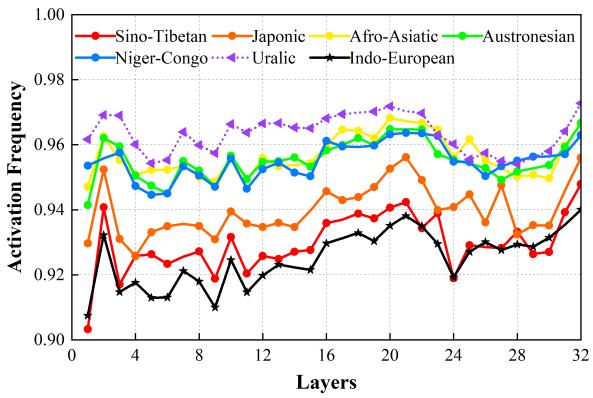


Figure 6. **Layer-wise activation frequency of neurons** viewed from language families in LLaMA 3.

This section analyzes the distribution of overlap neurons. Figure 5 compares neurons in language pairs within the same family (e.g., Ar-He) and across families (e.g., Ar-Sw). Obviously, Ar-He shares more overlapping neurons. Similar observations can be found in comparing language pairs from different source languages to a same target language (Figure 10 in Appendix C.4)

From the perspective of language families, Figure 6 examines the activated behaviors of neurons in low-resource languages. Obviously, low-resource languages within the Uralic family have the highest activation frequency, while Indo-European languages have the lowest. We hypothesize LLMs activate neurons more frequently for processing low-resource languages due to their perceived difficulty.

6 Conclusion

In this work, we explore whether sharing neurons can improve LLMs' cross-lingual performance on low-resource languages. We propose a simple yet effective language-bridge approach with the help of neuron interpretation. To ensure accurate and full activation of overlap neurons across languages, we construct probing data from the ground-truth MUSE dictionaries. By quantifying neuron similarity, we seek the optimal bridge for X-ICL and conduct extensive experiments to validate its efficacy and generalization.

Limitations

This work focuses on sharing neurons across languages and relies on the evaluated datasets to validate the effectiveness of our approach. Due to the lack of comprehensive benchmarks for low-resource languages, our experiments cover only 15 language pairs and select 4 low-resource languages from distinct families as target languages and test their performance on 4 cross-lingual tasks. Second, our study reveals that high-quality probe data is crucial to accurately analyze neuron behaviors of low-resource languages, while the proposed linguistic distance measurement is probe-data-induced, offering qualitative but not quantitative insights. Finally, although bridge selection should ideally follow linguistic phylogeny, we aim to select bridges that LLMs can best exploit, inevitably reflecting their training biases.

References

Zabir Al Nazi, Md Rajib Hossain, and Faisal Al Mamun. 2024. Evaluation of open and closed-source llms for a low-resource language with zero-shot, few-shot , and chain-of-thought prompting.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and 1 others. 2020b. Language models are probing learners. In *NeurIPS*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. 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 405–433.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *arXiv preprint arXiv:2411.17401*.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

Wikimedia Foundation. 2024. Wikimedia downloads.

Russell D. Gray, Alexei J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science*, 323(5913):479–483.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 5.1.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, and 1 others. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Jaime Huerta-Cepas, François Serra, and Peer Bork. 2016. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635–1638.

Mike Izbicki. 2022. Aligning word vectors on low-resource languages with Wiktionary. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 107–117, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Weize Liu, Yinlong Xu, Hongxia Xu, and other. 2024. Unraveling babel: Exploring multilingual activation patterns of llms and their applications. In *EMNLP*, pages 11855–11881. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. *arXiv preprint arXiv:2503.17456*.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, and et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. LessWrong.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5877–5891.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024. Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7113.

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

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, and 1 others. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *ACL (1)*, pages 5701–5715. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307.

Ivan Vulic, Edoardo Maria Ponti, Robert Litschko, Goran Glavas, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *EMNLP (1)*, pages 7222–7240. Association for Computational Linguistics.

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

Søren Wichmann and Eric W. Holman. 2009. *Temporal stability of linguistic typological features*. Lincom Europa.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, and 1 others. 2023. Bloom+ 1: Adding language support to bloom for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. Exploring translation mechanism of large language models. *arXiv preprint arXiv:2502.11806*.

A Appendix: Linguistic Similarity Based on Glottolog Phylogenetic Trees

We leverage Glottolog version 5.1 (Hammarström et al., 2023) as a foundational phylogenetic framework to calculate the linguistic similarity of human languages. It has two key steps: data preprocessing and similarity calculation.

Data Preprocessing. The preprocessing pipeline consists of three key steps: 1) Locate Glottocode identifiers with regex pattern matching to ensure unambiguous language node identification. 2) Standardize node naming with underscores (e.g., [sini1245] $\rightarrow$ _slaini1245_), ensuring consistent formatting in downstream phylogenetic analyses. 3) Mitigate encoding conflicts through temporary file caching. These steps preserve accurate parsing of phylogenetic tree while maintaining computational compatibility.

Similarity Calculation. The proposed metric integrates two well-established principles from historical linguistics: node distance normalization and depth-adjusted compensation.

First, building upon Wichmann & Holman's framework for typological stability assessment (Wichmann and Holman, 2009), we compute the inter-language distance $d(L_1, L_2)$ between languages L_1 and L_2 using ETE3's optimized tree traversal algorithms (Huerta-Cepas et al., 2016). We then normalize the distance to make it comparable across language families, calculated as:

$$S_{\text{distance}} = 1 - \min\left(1, \frac{d(L_1, L_2)}{\hat{D}}\right) \quad (6)$$

where $\hat{D}$ is the family-specific maximum. For example, $\hat{D} = 80$ for Sino-Tibetan languages, reflecting their deep internal divergence, whereas $\hat{D} = 75$ for Indo-European languages, due to their relatively shallower subgroup structure.

Second, depth-adjusted compensation aims to mitigate biases introduced by uneven tree depth and family-specific structural variation. Following the work (Gray et al., 2009) to calculate depth disparity factor $\delta(L_1, L_2)$, we measure the depth $\alpha_{\text{depth}}(L_1, L_2)$ between L_1 and L_2 as:

$$\alpha_{\text{depth}} = 1 - \frac{\delta(L_1, L_2)}{\max(\text{depth}(L_1), \text{depth}(L_2))} \quad (7)$$

The final language similarity score is computed as:

$$\text{Sim}(L_1, L_2) = S_{\text{distance}} \times \alpha_{\text{depth}} \quad (8)$$

B Appendix: Supplementary Cross-lingual Tasks

To further verify robustness, we conduct additional experiments using the bridge languages selected by our method on downstream tasks, including Cross-Lingual Question Answering (CLQA) and Cross-Lingual Natural Language Inference (XNLI), as shown in Table 5. Due to the limited availability of cross-lingual benchmarks covering our target low-resource languages (e.g., Tagalog), the evaluation is restricted to 6 language pairs for CLQA and 3 pairs for XNLI.

Table 5. Evaluation on CLQA and XNLI cross-lingual tasks. **Bold** highlights improved performance.

	CLQA		XNLI	
	Zero-shot	Ours	Zero-shot	Ours
Zh-Ja	42.80	-0.60	-	-
Zh-Sw	38.20	**+5.20**	35.90	**+1.30**
Ar-Ja	41.60	**+6.20**	-	-
Ar-Sw	34.20	**+2.40**	35.50	**+3.70**
Id-Ja	34.60	**+3.60**	-	-
Id-Sw	34.80	**+5.20**	36.60	**+0.60**

C Appendix: Neuron Patterns

C.1 Deactivation Overlap Neurons

Figure 7 presents the distribution of overlap neurons and their deactivation effects on the Chinese-Hebrew (Zh-He) BLI task.

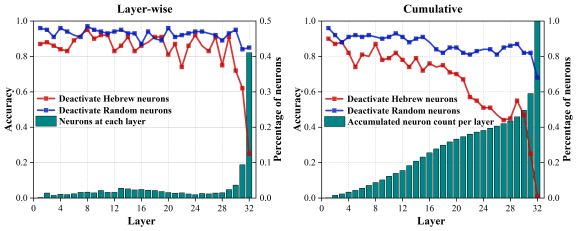


Figure 7. **Overlap neuron distributions and their deactivation effects** on the Chinese-Hebrew BLI task.

C.2 Linguistic Spectrum in Mistral

C.3 Parameter K Discussion

Here we discuss which k middle layers should be selected to quantify linguistic similarity. According to observations in section 3.3.3, neurons in middle layers should be prioritized over final layers when measuring language similarity. We use the embedding semantic similarity metric to determine K. For example, we analyze Arabic-Hebrew

		Sino-Tibetan	Japonic	Afro-Asiatic		Austronesian		Uralic		Niger-Congo	Indo-European					
		zh	ja	ar	he	id	tl	fi	hu	sw	en	de	fr	it	pt	es
Sino-Tibetan	zh	1.000	0.700	0.486	0.387	0.426	0.280	0.344	0.474	0.057	0.480	0.486	0.441	0.422	0.459	0.442
Japonic	ja	0.700	1.000	0.398	0.375	0.380	0.270	0.370	0.476	0.062	0.338	0.442	0.336	0.370	0.346	0.349
Afro-Asiatic	ar	0.486	0.398	1.000	0.671	0.445	0.353	0.350	0.382	0.231	0.293	0.349	0.398	0.353	0.375	0.408
	he	0.387	0.375	0.671	1.000	0.357	0.280	0.293	0.333	0.135	0.217	0.312	0.274	0.295	0.313	0.289
Austronesian	id	0.426	0.380	0.445	0.357	1.000	0.600	0.466	0.478	0.299	0.369	0.475	0.427	0.462	0.487	0.464
	tl	0.280	0.270	0.353	0.280	0.600	1.000	0.365	0.317	0.299	0.207	0.278	0.243	0.289	0.353	0.362
Uralic	fi	0.344	0.370	0.350	0.293	0.466	0.365	1.000	0.530	0.222	0.204	0.432	0.317	0.396	0.340	0.309
	hu	0.474	0.476	0.382	0.333	0.478	0.317	0.530	1.000	0.057	0.369	0.594	0.491	0.507	0.510	0.478
Niger-Congo	sw	0.057	0.062	0.231	0.135	0.299	0.299	0.222	0.057	1.000	0.000	0.065	0.031	0.097	0.070	0.070
Indo-European	en	0.480	0.338	0.293	0.217	0.369	0.207	0.204	0.369	0.000	1.000	0.540	0.583	0.542	0.571	0.571
	de	0.486	0.442	0.349	0.312	0.475	0.278	0.432	0.594	0.065	0.540	1.000	0.611	0.610	0.590	0.574
	fr	0.441	0.336	0.398	0.274	0.427	0.243	0.317	0.491	0.031	0.583	0.611	1.000	0.752	0.742	0.729
	it	0.422	0.370	0.353	0.295	0.462	0.289	0.396	0.507	0.097	0.542	0.610	0.752	1.000	0.779	0.771
	pt	0.459	0.346	0.375	0.313	0.487	0.353	0.341	0.510	0.070	0.571	0.590	0.742	0.779	1.000	0.879
	es	0.442	0.349	0.408	0.289	0.464	0.362	0.309	0.478	0.070	0.571	0.574	0.729	0.771	0.879	1.000

Figure 8. **Mistral's linguistic spectrum** across 15 languages from 7 families. The color intensity represents the degree of overlap between language pairs.

and Chinese-Hebrew translation pairs by prompting LLMs with the same semantical words in Arabic and Chinese to generate the corresponding Hebrew translation. We then compute the layer-wise embedding semantic similarity between the two pairs and identify layers in which this similarity is insensitive to variations in the predicted tokens, reflecting the inherent distance between the languages. As shown in Figure 9, we find embedding similarity is stable in the middle layers 10-21 of LLaMA 3 and is not affected by token-level variations. Therefore, K layers is set to be 10-21 in LLaMA 3 and 15-23 in Mistral.

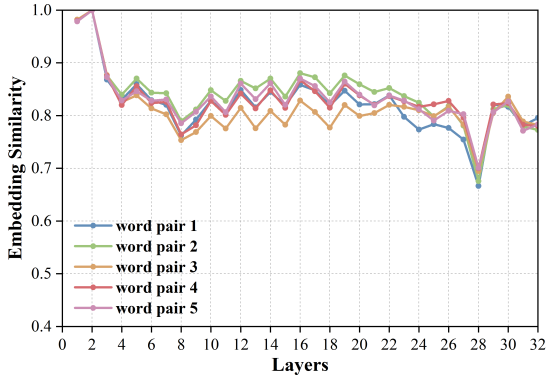


Figure 9. **Embedding semantic similarity** between Arabic-Hebrew and Chinese-Hebrew translations when predicting the same token at each layer of LLaMA 3.

C.4 Overlap Neuron Distribution

Figure 10 presents the distribution of overlap neurons across different language pairs in LLaMA 3.

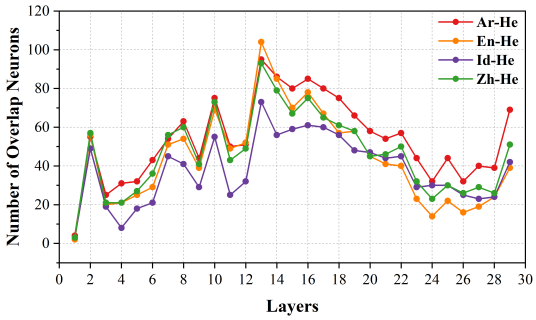


Figure 10. **Distribution of overlap neurons** across different language pairs in LLaMA 3.

D Appendix: Ablation Results

This section presents the detailed experimental ablation to evaluate the impact of neuron probe data construction and the HSIC similarity metric on the bridge language selection. Table 4 and Table 6 present the results of ablation experiments on the BLI task. "w/o $*$" denotes replacing our probe data with a simplified version based on FLORES+. For example, Figure 11 illustrates the construction of "w/o $*$" probe data for Indonesian-Hebrew.

Table 6. Performance comparison of using HSIC and Cosine similarity metrics on the BLI task.

	LLaMA-3-8B			
	Bridge	HSIC	Bridge	Cosine
Zh-Ja	En	77.90 ↑	De	76.60
Zh-He	En	56.70 ↓	De	58.60
Zh-Tl	En	54.00 ↑	De	51.00
Zh-Sw	En	36.00 ↑	De	33.80
Ar-Ja	En	80.60 ↑	De	76.30
Ar-He	En	64.50 ↑	De	64.10
Ar-Tl	En	56.80 ↑	De	52.80
Ar-Sw	En	42.40 ↓	De	43.00
Id-Ja	En	66.10 ↑	De	65.10
Id-He	Es	61.30 ↑	De	59.90
Id-Tl	Fr	61.00 ↑	De	57.00
Id-Sw	Fr	30.00 ↑	De	28.30
En-He	Es	71.80 ↓	Pt	72.30
En-Tl	Fr	58.70 ↑	Pt	55.80
En-Sw	Fr	26.20 ↑	It	21.50

E Appendix: Neuron Semantic Analysis

In this section, we analyze the semantic similarity of overlapping neurons across two language groups: Hebrew-Tagalog-Swahili (He-Tl-Sw, from different language families) and Portuguese-Spanish-Italian (Pt-Es-It, from the same language family). Our goal is to examine whether overlapping neurons cluster together within the same language family or also across different families.

Examples of Ablation experiment (Id-He)

with probe data:
Indonesia: "api" → עִבְרִית:

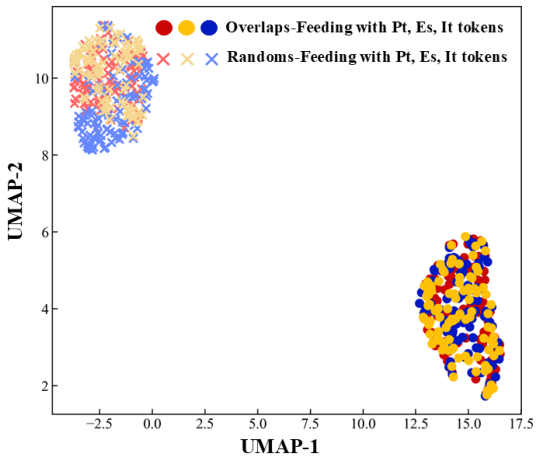
w/o probe data (From Flores):
Sejarah atau tawarik (artinya "mengusut, pengetahuan yang diperoleh melalui penelitian") adalah kajian tentang masa lampau, khususnya bagaimana kaitann dengan manusia. הִיסטוֹרִיָה

Figure 11. **Example of "w/o ∗" probe data** in Indonesian-Hebrew.

For comparison, we select 100 overlapping neurons and randomly sample the same number of non-overlapping neurons for comparison. We feed the model with m parallel sentences and then record the neuron activation frequency for each sentence, obtaining three $m \times 100$ activation matrices for both overlapping and random neurons. These matrices are mapped to a 2D semantic space using UMAP (McInnes et al., 2018), with each point representing a neuron activated by m sentences from a language. Colored circles of red, yellow, and blue denote overlapping neurons and colored $\times$ symbols of red, yellow, and blue denote random neurons.

As presented in Figure 12, the overlapping neurons, whether identified within the same family or across language families, cluster closely together. This suggests that our approach can effectively identify neurons that encode genuine shared semantics across languages. In addition, random neurons tend to align with linguistic relationships, clustering when feeding LLMs tokens from topologically related languages, such as Portuguese (Pt), Spanish (Es), and Italian (It), while remaining dispersed when feeding LLMs tokens from distant languages, like Hebrew (He), Tagalog (Tl), and Swahili (Sw).

(a) Neuron semantic similarity in Pt-Es-It



(b) Neuron semantic similarity in He-Tl-Sw

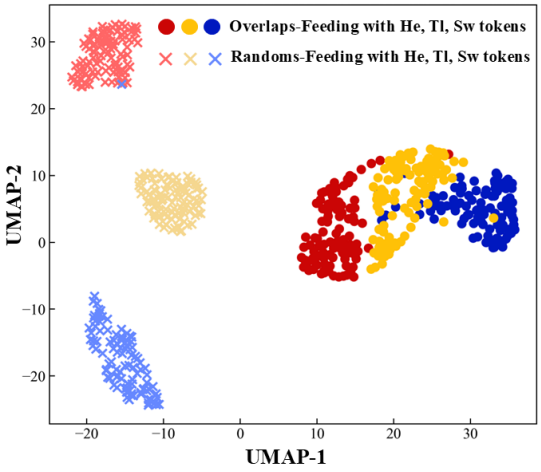


Figure 12. **Visualization of semantic similarity** comparing language-overlapping neurons and randomly sampled neurons.

F Appendix: Prompt Templates

Prompt for zero-shot in BLI Task

Template:

src_lang: "src_word" → trg_lang:

Examples in Indonesian-Tagalog pair:

Indonesian: "matahari" → Tagalog:

Indonesian: "bunga" → Tagalog:

Prompt for zero-shot with bridge in BLI Task

Template:

step1: src_lang: "src_word" → aid_lang:

step2: src_lang: "src_word" → aid_lang: "aid_word" → trg_lang:

Examples in Indonesian-Tagalog pair using English:

step1: Indonesian: "matahari" → English:

step2: Indonesian: "matahari" → English: "sun" → Tagalog:

Prompt for zero-shot in MRC Cross-lingual Task

Template:

Answer the following question based on the passage. Respond with A, B, C, or D.

Passage: <source-language passage>

Question: <target-language question>

Choices:

A: <target-language choice 1>

B: <target-language choice 2>

C: <target-language choice 3>

D: <target-language choice 4>

Answer:

Examples in Swahili-Indonesian pair:

Answer the following question based on the passage. Respond with A, B, C, or D.

Passage: Ndiyo! Mfalme Tutankhamuni, ambaye ...

Question: Kapan Raja Tutankhamun mendapatkan ketenaran?

Choices:

A: Setelah pencurian makamnya

B: Selama masa kekuasaannya

C: Setelah penemuan makamnya

D: Setelah disebutkan dalam daftar raja kuno

Answer:

Prompt for zero-shot with bridge in MRC Cross-lingual Task

Template:

Step1:Translate the following text from source-language to bridge-language,Translation:

Step2:Answer the following question based on the passage. Respond with A, B, C, or D.

Passage: <bridge-language passage>

Question: <target-language question>

Choices:

A: <target-language choice 1>

B: <target-language choice 2>

C: <target-language choice 3>

D: <target-language choice 4>

Answer:

Examples in Swahili-Indonesian pair using English:

Step1: Translate the following text from Swahili to English, Translation: Yes! King Tutankhamun, who is sometimes known as "King Tut" or "Boy King" ...

Step2: Answer the following question based on the passage. Respond with A, B, C, or D.

Passage: Yes! King Tutankhamun, who is sometimes known as "King Tut" or "Boy King" , is ...

Question: Kapan Raja Tutankhamun mendapatkan ketenaran?

Choices:

A: Setelah pencurian makamnya

B: Selama masa kekuasaannya

C: Setelah penemuan makamnya

D: Setelah disebutkan dalam daftar raja kuno

Answer: