From Neurons to Semantics: Evaluating Cross-Linguistic Alignment Capabilities of Large Language Models via Neurons Alignment

Chongxuan Huang[1,3], **Yeyong Shi**[2,3], **Biao Fu**[2,3] ✉, **Qifeng Su**[1,3], **Xiaodong Shi**[1,3] ✉
[1] School of Informatics, Xiamen University
[2] Institute of Artificial Intelligence, Xiamen University
[3] Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism
{huangchongxuan,yeyongshi,biaofu,suqifeng}@stu.xmu.edu.com
mandel@xmu.edu.com

Abstract

Large language models (LLMs) have demonstrated remarkable multilingual capabilities, however, how to evaluate cross-lingual alignment remains underexplored. Existing alignment benchmarks primarily focus on sentence embeddings, but prior research has shown that neural models tend to induce a non-smooth representation space, which impact of semantic alignment evaluation on low-resource languages. Inspired by neuroscientific findings that similar information activates overlapping neuronal regions, we propose a novel *Neuron State-Based Cross-Lingual Alignment* (**NeuronXA**) to assess the cross-lingual a lignment capabilities of LLMs, which offers a more semantically grounded approach to assess cross-lingual alignment. We evaluate NeuronXA on several prominent multilingual LLMs (LLaMA, Qwen, Mistral, GLM, and OLMo) across two transfer tasks and three multilingual benchmarks. The results demonstrate that with only 100 parallel sentence pairs, NeuronXA achieves a Pearson correlation of 0.9556 with downstream tasks performance and 0.8524 with transferability. These findings demonstrate NeuronXA's effectiveness in assessing both cross-lingual alignment and transferability, even with a small dataset. This highlights its potential to advance cross-lingual alignment research and to improve the semantic understanding of multilingual LLMs.

1 Introduction

> *The brain has its own language for testing the structure and consistency of the world.*
>
> Carl Sagan

Recent advancements in autoregressive Large language models (LLMs) have demonstrated remarkable multilingual capabilities in understanding, reasoning, and language generation (OpenAI et al., 2023; Dubey et al., 2024a; Yang et al., 2024). This has spurred growing interest in evaluating their performance across diverse languages (Hendrycks et al., 2021a,b; Ahuja et al., 2023; Zhang et al., 2025; Ye et al., 2025). However, the mechanisms underlying cross-lingual alignment in LLMs remain insufficiently understood.

Research on cross-lingual alignment has focused on linguistic isomorphism in representation spaces and its impact on cross-lingual transfer (Ye et al., 2023). Studies have explored the emergence of latent languages in multilingual processing (Zhao et al., 2024a; Wendler et al., 2024), alignment dynamics during pre-training (Wang et al., 2024a), as well as the morphological and syntactic structures of model embeddings (Papadimitriou et al., 2021). Various strategies have been proposed to enhance alignment, including interventions at different stages of model training (Yang et al., 2020; Zhu et al., 2024; Li et al., 2024).

Additionally, some research has been dedicated to evaluating cross-lingual alignment, particularly through the alignment of embedding spaces. Many studies adopt unsupervised methods to assess conceptual alignment across languages (Mousi et al., 2024), utilizing metrics such as cosine similarity (Li et al., 2025; Kargaran et al., 2024) to compute representational similarity. However, prior work has shown that neural architectures such as BERT and GPT tend to induce anisotropic representation spaces (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020). The collapse of representations in the semantic space diminishes the semantic expressiveness of low-resource languages (Li et al., 2025), thereby limiting the reliability of embedding-based evaluations of cross-lingual semantic alignment.

Prior studies have shown that neurons within feedforward network (FFN) modules encode diverse forms of knowledge (Dai et al., 2022; Voita et al., 2024; Gurnee et al., 2024). Drawing inspiration from neurobiological findings—where similar stimuli activate overlapping neural circuits—we hy-

✉ Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28956–28974

pothesize that neuron activations can serve as intrinsic representations of multilingual queries. These activations may provide a more structured and robust means of capturing cross-lingual knowledge, offering new insights into multilingual alignment.

In this study, we introduce a novel evaluation framework called *Neuron State-Based Cross-Lingual Alignment* (**NeuronXA**) to assess the cross-lingual alignment capabilities of LLMs. The proposed method quantifies the activation likelihood of individual neurons in response to parallel corpora across multiple languages. Using neuron states as intrinsic representations, NeuronXA calculates alignment scores by evaluating the consistency of parallel sentences within the representation space, thus offering a robust method for alignment evaluation.

Based on NeuronXA, we systematically evaluate the alignment of several popular open-source LLMs, yielding several key findings:

- First, the neuron state-based representation method more effectively encodes cross-lingual knowledge. Using this intrinsic representation improves the model's accuracy in semantic retrieval, particularly in bidirectional retrieval tasks.
- Second, our experimental results demonstrate that the proposed NeuronXA method provides a reliable evaluation approach, exhibiting a strong correlation with both the model's transferability and its performance on multilingual benchmarks. NeuronXA offers a robust framework for assessing the cross-lingual alignment capabilities of large language models.
- Third, an analysis of alignment scores across different model layers reveals that the highest scores occur in the middle layers, while the lowest scores are observed in the lower and upper layers. This pattern suggests that lower layers primarily map inputs from various languages into a shared semantic space centered around high-resource languages, whereas upper layers project semantic content onto language-specific vocabulary tokens.

2 Methods

2.1 Background

Currently, LLMs are predominantly developed using the autoregressive Transformer architecture (Vaswani et al., 2017), where the core components include multi-head self-attention (MHA) and feedforward networks (FFNs). Previous research has demonstrated that the feedforward layers in Transformers can be conceptualized as key-value memory networks (Geva et al., 2021), which store world knowledge to aid in sequence understanding. Consequently, our study focuses primarily on the analysis of FFNs.

In the current LLMs architectures, FFNs typically employ gated projections for each token within a sequence. The computation for this process is defined as:

$$\mathrm{FFN}^{I}(\boldsymbol{x}) = \sigma\left(\boldsymbol{W_G}\boldsymbol{x} + \boldsymbol{b_G}\right) \odot \left(\boldsymbol{W^I}\boldsymbol{x} + \boldsymbol{b^I}\right), \tag{1}$$

where $\boldsymbol{W_G}, \boldsymbol{W^I} \in \mathbb{R}^{d_{\text{ff}} \times d}$ and $\boldsymbol{b^I}, \boldsymbol{b_G} \in \mathbb{R}^{d_{\text{ff}}}$ represent the weight matrices and bias vectors for the input linear layer $\mathrm{FFN}^{I}(\cdot)$ and the gate linear layer $\mathrm{FFN}_{G}(\cdot)$, respectively. Following prior work (Zhang et al., 2023; Wang et al., 2022), we can decompose the FFN layer into d_{ff} neurons, each of which corresponds to a row in the input and gate layers, as well as a column in the output layer. The outputs of the FFN layers can thus be rewritten as the sum of the individual neuron outputs:

$$\mathrm{FFN}(\boldsymbol{x}) = \sum_{i}^{d_{\text{ff}}} \mathrm{FFN}^{I}(\boldsymbol{x})_i \boldsymbol{W}^{O}_{:,i} + \boldsymbol{b}^{O}_{i}, \tag{2}$$

where the intermediate value $\mathrm{FFN}^{I}(\boldsymbol{x})_i$ denotes the activation of the i-th neuron.

2.2 NeuronXA

Previous studies have demonstrated that neurons within the FFN modules can store factual knowledge (Dai et al., 2022), encode positional information (Voita et al., 2024), and respond to specific syntactic triggers (Gurnee et al., 2024), among other functions. Building on these insights, we propose treating neuron states as intrinsic representations of the input query, with these representations potentially reflecting the various types of knowledge that underlie the query.

To capture alignment across different levels of linguistic knowledge more effectively, we leverage these neuron states as representations of the input query. Subsequently, we evaluate the alignment between queries from different languages and a high-resource-centered representation space, using this measure to define the corresponding language's alignment score.

Neuron States Detection. Neuron states can be detected in two distinct ways, each providing valuable insights into the model's behavior. The first method examines the neuron's activation states, which reflect the model's response to the input. Specifically, the j-th neuron in the i-th FFN layer is considered *activated* if its activation value, $\alpha(\tilde{\boldsymbol{h}}^i \boldsymbol{W}_1^i)_j$, exceeds zero (Nair and Hinton, 2010; Tang et al., 2024). This approach highlights the neuron's immediate reaction to the input features.

The second method for detecting neuron partitions relies on the neuron's absolute activation value , which indicates the contribution of the neuron to the output of the FFN layer. This approach is commonly used as a functional indicator (Zhang et al., 2023; Wang et al., 2022), where the absolute activation value of the j-th neuron in the i-th layer serves as the representation of that neuron's role in processing a given input sentence pair.

Sentence Representation. To compute the NeuronXA score, it is first necessary to obtain the sentence representation. Unlike encoder-only models, which utilize a bidirectional attention mechanism (Devlin et al., 2019), decoder-only LLMs rely on causal attention. Thus, directly averaging the representations of all tokens, as is typically done in encoder-only models, would result in an overrepresentation of early tokens, which disproportionately influences the overall sentence representation. A common approach to mitigate this issue is to use the representation of the final token (Neelakantan et al., 2022; Wang et al., 2024b; Ma et al., 2024). However, this method does not fully capture the entire sentence. To address this limitation, Muennighoff (2022) proposed a position-weighted average representation, which is defined as:

$$N_l = \sum_{t=1}^{T} w_t n_{lt} \quad \text{with} \quad w_t = \frac{t}{\sum_{k=1}^{T} k}, \tag{3}$$

where T denotes the token count of the sentence, n_{lt} represents the neuron state of the t-th token at layer l, and N_l signifies the sentence neuron states at layer l.

NeuronXA Score. Cross-lingual alignment refers to the tendency of semantically similar words or sentences to be closely aligned within a shared representation space (Hämmerl et al., 2024; Kargaran et al., 2024). When the alignment between languages L_1 and L_2 is strong, semantically similar sentences l_1 and l_2 should have their closest neighbors in the representation space of the opposite language. We evaluate the proportion of sentence pairs that satisfy this alignment to assess cross-lingual alignment.

We generate a square matrix $C(l)$ representing cosine similarities of sentence representation at the output of layer l for all parallel sentences in languages L_1 and L_2. Let c_{ij} denote the element at the i-th row and j-th column of $C(l)$, corresponding to the cosine similarity between the i-th sentence of L_1 and the j-th sentence of L_2 at layer l of LLMs. Then we define the NeuronXA alignment score as:

$$\mu_{C(l)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(c_{ii} > \{c_{ij}, c_{ji}\}_{j \neq i}\right), \tag{4}$$

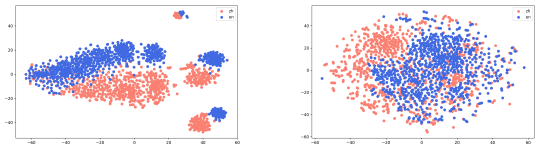
where n is the the dimension of the matrix, and $\mathbf{1}(\cdot)$ is the indicator function, which equals 1 if its argument condition evaluates to true and 0 otherwise. The calculation of this alignment score can be regarded as calculating the proportion of parallel sentences that satisfy weak alignment in the representation space.

The NeuronXA alignment score $\mu_{C(l)}$ is computed for language L_1 with respect to pivot language L_2 at each layer l of the language model. To obtain a single NeuronXA alignment score for a given the language model and language pair (L_1, L_2), we use mean pooling over multiple layers.

3 Experimental Setup

Models. We conduct experiments on several models with approximately 7B parameters, a widely recognized baseline size in the LLM community. The models selected for evaluation include LLaMA-2, LLaMA-3, LLaMA-3.1 (Touvron et al., 2023; Dubey et al., 2024b), Qwen 2.5 (Yang et al., 2024), Mistral 0.3 (Jiang et al., 2023), Olmo 2 (OLMo et al., 2024), and GLM 4 (Zeng et al., 2024). To assess the scalability of our findings, we additionally evaluate the larger Qwen 2.5 14B model, as well as smaller models such as LLaMA-3.2 3B. These models have demonstrated strong multilingual performance and are widely adopted in the research community, making them suitable candidates for our evaluation.

Dataset. We utilize two cross-lingual parallel datasets, FLORES-200 (Costa-jussà et al., 2022) and Tatoeba (Artetxe and Schwenk, 2019), to evaluate the effectiveness of the neuron state-based representation method in bridging the semantic gap



(a) Sentence Embedding. (b) Neuron State.

Figure 1: Visualization of sentence representations for 100 Tatoeba sentence pairs in Chinese and English, projected into 2D using t-SNE. The results compare two representation methods from Llama3.1-8B: sentence embeddings (Figure 1a), which show significant misalignment, and the proposed NeuronXA method (Figure 1b), which mitigates this misalignment.

between semantically similar sentences. To provide a more comprehensive assessment of cross-lingual alignment across a diverse set of languages, we select FLORES-200 for comparative experiments on downstream tasks, due to its extensive language coverage. A detailed discussion of the dataset can be found in Appendix A.

3.1 Parallel Sentence Retrieval

Problem Formulation. Cross-lingual parallel sentence retrieval aims to identify semantically equivalent sentences across languages, facilitating applications such as machine translation, multilingual retrieval, and cross-lingual question answering. The primary challenge is to learn sentence representations that capture meaning within a shared semantic space. The effectiveness of retrieval relies on these representations accurately preserving semantic content across different languages.

Neuron Activation-Based Representations. We propose *Neuron Activation State (NAS)* and *Neuron Activation Value (NAV)* as novel representations derived from neuron activation patterns in pre-trained language models. Unlike conventional embeddings, which suffer from issues such as non-smoothness, as illustrated in Figure 1, neuron-state-based representations offer a smoother representation space, providing a more structured and interpretable approach for cross-lingual alignment.

Setup. We evaluate our method on the FLORES-200 and Tatoeba datasets, covering both head and long-tail languages (see Appendix A for details). Sentence representations are constructed using a weighted token averaging strategy with the Llama 3.1-8B model. Given the model's depth, we apply max-pooling to enhance retrieval accuracy. The primary evaluation metric is the bidirectional retrieval accuracy, which quantifies the proportion of correctly retrieved parallel sentence pairs, providing a robust assessment of representation effectiveness.

3.2 Alignment Evaluate methods

For comparison, we evaluate the model's cross-lingual alignment capabilities using the following methods, with assessment conducted on 100 parallel sentence pairs from the FLORES-200 dataset. The robustness of the NeuronXA method is discussed in detail in Appendix E.

(a) *Multilingual Evaluation via Cross-Linguistic Alignment (MEXA)* (Kargaran et al., 2024): MEXA measures alignment by computing the similarity between English and non-English sentence embeddings using parallel sentences. To mitigate centralization bias, it employs relative cosine similarity for cross-lingual alignment score calculation.

(b) *Neural Activation State-based Cross-Lingual Alignment (NASCA, ours)*: NASCA represents sentences based on neuron activation states (binary 0 or 1). The alignment score is derived from the proportion of parallel sentences exhibiting weak alignment in the representation space.

(c) *Neural Activation Value-based Cross-Lingual Alignment (NAVCA, ours)*: NAVCA follows a similar approach to NASCA but uses the absolute magnitude of neuron activations instead of binary states. Further details are provided in Section 2.2.

3.3 Cross-lingual Transfer Evaluation

Following prior work (Li et al., 2024; Wang et al., 2024a), we assess the zero-shot cross-lingual transfer capability of models through two downstream tasks. To investigate the relationship between alignment scores and transferability, we compute the Pearson correlation coefficient between the alignment score and task performance. A higher correlation indicates that the alignment score effectively predicts the model's cross-lingual transfer ability.

Zero-shot Cross-lingual Transfer (ZS-CLT). This is a standard approach for evaluating a model's cross-lingual generalization. In this setting, a model is fine-tuned on a given task in a source language and tested on the same task in target languages without additional training. We use the widely adopted XNLI dataset (Conneau et al., 2018) for evaluation, which assesses sentence understanding in multiple languages by determining the relationship between sentence pairs.

Representation	Direction	FLORES-200		Tatoeba	
		Head	Long-tail	Head	Long-tail
Embedding	En → xx	90.12	**50.71**	23.93	13.28
	xx → En	88.60	**47.74**	54.66	**41.95**
	En ⇔ xx	83.78	40.95	16.86	10.12
NAS	En → xx	**93.10**	50.49	**67.77**	**42.12**
	xx → En	**89.82**	46.95	**65.05**	38.60
	En ⇔ xx	**87.07**	**42.20**	**57.78**	**32.47**

Table 1: Retrieval results on FLORES-200 and Tatoeba in xx → En and En → xx direction, along with En ⇔ xx direction. The **bold** font denotes the best results.

Cross-lingual Knowledge Application (CLKA). LLMs acquire extensive world knowledge from multilingual corpora. An essential capability of these models is the ability to learn knowledge in one language and apply it across others. To evaluate this ability, we use the BMLAMA-53 dataset (Qi et al., 2023), a benchmark designed to assess cross-lingual knowledge consistency in multilingual LLMs.

All fine-tuning experiments were conducted using the LLaMA Factory framework (Zheng et al., 2024), with prompt templates corresponding to the specific task requirements. Due to computational resource constraints, we applied 4-bit quantized LoRA (Hu et al., 2022) for fine-tuning.

3.4 Multilingual Benchmarks Evaluation

We evaluate model alignment by measuring how different languages are mapped into a shared representation space, which is inherently biased toward high-resource languages. As a result, alignment scores between high-resource languages and others can serve as an indirect indicator of performance in lower-resource languages.

To evaluate this alignment, we utilize three benchmarks—Belebele (Bandarkar et al., 2024), m-ARC (Lai et al., 2023), and m-MMLU (Lai et al., 2023)—which collectively encompass a diverse range of high-, medium-, and low-resource languages. A detailed description of these datasets is provided in Appendix A.

All experiments utilize 5-shot in-context learning via the lm-evaluation-harness framework[1], with default prompt templates for comparability.

4 Results and Analysis

4.1 Enhanced Semantic Alignment in Parallel Sentence Retrieval

Table 1 shows cross-lingual semantic retrieval accuracy for different representations in the LLaMA3.1-8B model. The results highlight a performance gap between head and long-tail languages, with head languages consistently outperforming long-tail ones due to richer training data for the former, leading to stronger semantic alignment. Additionally, the impact of selecting other high-resource languages as query languages on semantic retrieval is discussed in Appendix C.

Directional Asymmetry. On the Tatoeba dataset, sentence embedding-based retrieval exhibits a 30.73% accuracy drop in the En → xx direction compared to xx → En denotes a Head language. This is because English provides richer semantic representations, aiding retrieval from other languages. Conversely, when querying in English, the representation of other languages is less robust, hindering retrieval. The NAS representation, however, achieves nearly symmetric accuracy in both directions, indicating that it better captures cross-lingual semantics and mitigates representation imbalances.

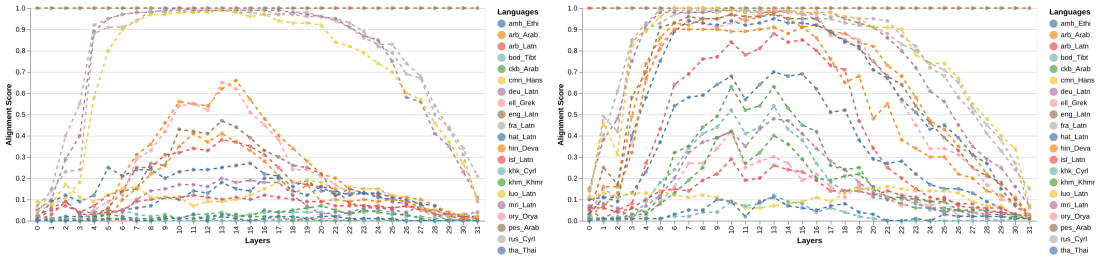
Dataset Impact. Retrieval accuracy is higher on FLORES-200 than on Tatoeba due to Tatoeba's lower sentence diversity, especially in low-resource languages, where semantically similar but distinct sentences complicate retrieval. In contrast, FLORES-200, sourced from Wikimedia and manually validated, offers greater diversity, enabling clearer semantic distinctions.

Representation Comparison. NAS consistently outperforms sentence embeddings in bidirectional retrieval accuracy, demonstrating its superior ability to encode cross-lingual semantics as an intrinsic representation. A key advantage of NAS is its robustness in handling long-tail languages, where it achieves better alignment between high- and low-resource languages. Moreover, NAS reduces directional asymmetry, yielding nearly symmetric performance in both En → xx and xx → En retrieval tasks. This suggests that NAS provides a more balanced cross-lingual representation.

4.2 The Dynamics of Alignment

Alignment Score Across Layers. Figure 2 shows how alignment varies across layers, calculated using NASCA. As model depth increases, alignment ability initially improves and then declines, with the lowest alignment observed in both the bottom and top layers. This suggests that in generative models, neurons in the lower and upper layers are primarily language-specific, while

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



(a) Llama 2 7B. (b) Llama 3.1 8B.

Figure 2: NASCA scores across all layers for different languages.

High→High		**High→Low**		**Low→Low**	
Language Pair	NASCA Score	Language Pair	NASCA Score	Language Pair	NASCA Score
English→German	0.7300	German→Silesian	0.5222	Azerbaijani→Turkmen	0.3862
Italian→French	0.8372	French→Panjabi	0.2872	Hungarian→Yiddish	0.3821
German→French	0.7628	Italian→Banjar	0.4353	Gujarati→Banjar	0.2191
French→Chinese	0.6922	Italian→Uighur	0.1831	Kazakh→Tatar	0.5291

Table 2: NASCA score of different language pairs of Llama-3.1-8B.

the intermediate layers contain shared multilingual neurons, a pattern found in previous studies (Zeng et al., 2025; Del and Fishel, 2022). These findings indicate that in the lower layers, LLMs rely on language-specific neurons to map aligned text from different languages into a shared representational space for semantic transformation. In contrast, the upper layers, responsible for token generation, require a higher concentration of language-specific neurons to handle vocabulary mapping.

Analysis of the selection of baseline languages. English is selected as the pivo language for evaluating cross-lingual alignment, as LLMs often align multilingual inputs around high-resource languages. To mitigate potential biases introduced by using English as the reference, we categorize evaluation into three groups: high-resource to high-resource, high-resource to low-resource, and low-resource to low-resource. Using the FLORES-200 dataset, we select four representative language pairs for each category, with results shown in Table 2. Our analysis shows that high-resource languages exhibit relatively stable distributions, while low-resource languages show significant variability. Although English serves as a natural reference point, other high-resource languages such as German and French can also be considered as baselines.

4.3 Downstream tasks Correlation

In this section, we empirically evaluate the effectiveness of our proposed representation method based on neuron states. We calculate NeuronXA scores between English and other languages, and investigate their correlation with both model cross-lingual transferability and performance on multilingual tasks.

NeuronXA is more closely related to model transferability. As shown in Table 3, both NASCA and NAVCA—our NeuronXA-based methods—outperform the sentence embedding-based baseline MEXA, which achieves an average Pearson correlation of 0.7731. In contrast, NASCA and NAVCA yield average Pearson correlations of 0.8293 and 0.8306, respectively, demonstrating a stronger correlation with the model's transferability. Notably, correlations with the CLKA task is significantly lower than that with the ZS-CLT task. We hypothesize that this gap arises from the limited size of the BMLAMA-53 dataset, which contains only 3,012 samples, potentially restricting its ability to capture real-world factual knowledge transfer. Nevertheless, both NASCA and NAVCA consistently exhibit high correlation coefficients overall.

NeuronXA is more closely associated with the model's multilingual capabilities. Similar to the results discussed in the transferability task, Table 4 presents the Pearson correlation coefficients between cross-lingual alignment scores and three multilingual benchmarks. The MEXA, NASCA, and NAVCA methods achieve average Pearson correlations of 0.8725, 0.9489, and 0.9341, respectively. Notably, both NASCA and NAVCA show substan-

			Llama 3.1 8B	Llama 3 8B	Llama 2 7B	Llama 3.2 3B	Qwen 2.5 14B	Qwen 2.5 7B	Mistral 0.3 7B	OLMo2 7B	GLM 4 9B	AVG
XNLI	weighted	MEXA	0.9259	0.9211	0.7519	0.9182	0.6212	0.6898	0.9446	**0.9500**	0.8107	0.8370
		NASCA	**0.9309**	0.9271	**0.9639**	0.9401	**0.8647**	**0.9209**	**0.9583**	0.9411	**0.9467**	**0.9326**
		NAVCA	0.9227	**0.9283**	0.9063	**0.9430**	0.8038	0.8415	0.8982	0.8777	0.9213	0.8937
	average	MEXA	0.7948	0.8379	**0.7486**	0.8506	0.7426	0.8039	0.9175	**0.9258**	0.8402	0.8291
		NASCA	**0.8332**	**0.8506**	0.6420	0.8589	**0.7574**	**0.8281**	**0.9373**	0.9161	**0.8575**	**0.8312**
		NAVCA	0.8243	0.8490	0.6253	**0.8709**	0.7270	0.8278	0.9100	0.8810	0.8570	0.8191
	last	MEXA	0.7737	0.8023	0.7155	0.8163	0.5539	0.6010	0.7642	**0.9377**	0.7288	0.7437
		NASCA	**0.9250**	**0.9248**	**0.9585**	0.9387	**0.8497**	**0.9143**	**0.9547**	0.9305	**0.9422**	**0.9265**
		NAVCA	0.9143	0.9248	0.8959	**0.9408**	0.7964	0.8384	0.8915	0.8635	0.9181	0.8871
BMLAMA-53	weighted	MEXA	0.6567	0.6739	0.8426	0.6223	0.7522	0.8473	0.8795	0.7500	**0.6922**	0.7463
		NASCA	0.6825	0.7187	0.8707	0.6748	0.7785	0.8575	0.8750	0.7975	0.6761	0.7701
		NAVCA	**0.7285**	**0.7415**	**0.8924**	**0.7361**	**0.8046**	**0.8773**	**0.9062**	**0.8871**	0.6850	**0.8065**
	average	MEXA	0.6618	0.6619	**0.8653**	0.6905	0.7602	0.8312	0.8436	0.8352	0.7270	0.7641
		NASCA	**0.7028**	0.6529	0.7739	0.6712	0.7794	0.8196	0.8462	0.8441	0.6794	0.7522
		NAVCA	0.6792	**0.6797**	0.7570	**0.7099**	**0.7909**	**0.8584**	**0.8684**	**0.8962**	**0.7290**	**0.7743**
	last	MEXA	0.6731	0.6987	0.7826	0.6864	0.6943	0.6789	0.8378	0.7689	0.6423	0.7181
		NASCA	0.6731	0.6999	0.8651	0.6657	0.7786	0.8537	0.8690	0.8068	0.6565	0.7632
		NAVCA	**0.7202**	**0.7304**	**0.8906**	**0.7342**	**0.8074**	**0.8688**	**0.9076**	**0.8946**	**0.6737**	**0.8031**

Table 3: Pearson correlation of MEXA and NeuronXA on the FLORES dataset across ZS-CLT and CLKA tasks. The values in the table represent the pearson correlation of NeuronXA and benchmark settings. The highest average correlations for each task are highlighted in **bold**, and the second highest are underlined.

tial improvements in their average Pearson correlations with downstream tasks compared to MEXA.

Analysis of different sentence representation calculation methods. Token-position-based weighted sentence representation methods are generally considered to capture more contextual information, a trend reflected in both Table 3 and Table 4. For both transferability tasks and multilingual benchmarks, the highest correlation coefficients are observed with the weighted method (except for the m-ARC task). The second-best performance is achieved by the average method, while the last-token method demonstrates relatively lower correlation coefficients.

Across all settings, the best overall results (higher correlation) were achieved when embeddings were computed using a weighted average and alignment scores were computed using NASCA, so we adopted this configuration as the default for NeuronXA.

Furthermore, Appendix B discusses the correlation coefficient between alignment scores and generative tasks. Additionally, the robustness of NeuronXA scores when other languages serve as base languages is explored in Appendix C.

5 Related Work

The remarkable progress in autoregressive LLMs has highlighted their exceptional multilingual competencies across comprehension, reasoning, and generative tasks (OpenAI et al., 2023; Dubey et al., 2024a; Yang et al., 2024; Fu et al., 2025a,b); however, the fundamental mechanisms governing these cross-linguistic capabilities remain inadequately elucidated. A systematic investigation of cross-lingual alignment through rigorous empirical evaluation could not only unravel the operational principles underlying linguistic generalization in LLMs but also inform the design of optimized methodologies for enhancing cross-lingual alignment efficiency in LLMs.

Multilingual mechanism. Prior studies have demonstrated that layers closer to the model's input or output exhibit more language-specific behavior than intermediate layers (Bhattacharya and Bojar, 2023). Zhao et al. (2024b) transformed queries into English for comprehension, conducted inference in intermediate layers using English while integrating multilingual knowledge, and generated responses consistent with the original language in the final layer. Additionally, Wendler et al. (2024) defined intermediate layers as the concept space and revealed that, for Llama models, this concept space is closer to English. Some researchers have explored the multilingual mechanisms of large models at the neuron level. Zhang et al. (2024) found regions in large models corresponding to multilingual and monolingual capabilities. Kojima et al. (2024) and Bhattacharya and Bojar (2023) analyzed language-specific neurons in large mod-

			Llama 3.1 8B	Llama 3 8B	Llama 2 7B	Llama 3.2 3B	Qwen 2.5 14B	Qwen 2.5 7B	Mistral 0.3 7B	OLMo 2 7B	GLM 4 9B	AVG
m-ARC	weighted	MEXA	0.9551	0.9464	0.9124	0.9142	0.8709	0.9589	0.9575	0.8925	0.9225	0.9256
		NASCA	0.9570	0.9522	0.9369	0.9186	0.9696	0.9479	0.9539	0.9177	0.9713	0.9472
		NAVCA	**0.9756**	**0.9725**	**0.9649**	**0.9522**	**0.9786**	**0.9820**	**0.9847**	**0.9569**	**0.9731**	**0.9712**
	average	MEXA	0.9657	**0.9624**	**0.9426**	0.9319	**0.9773**	0.9664	0.9310	0.8800	**0.9688**	0.9473
		NASCA	0.9650	0.9575	0.9277	0.9308	0.9692	0.9438	0.9470	0.8853	0.9592	0.9428
		NAVCA	**0.9678**	0.9616	0.9241	**0.9412**	0.9686	**0.9692**	**0.9540**	**0.9133**	0.9638	**0.9515**
	last	MEXA	0.8833	0.8979	0.8853	0.8925	0.7729	0.7938	0.9279	0.9187	0.8543	0.8696
		NASCA	0.9591	0.9535	0.9400	0.9212	0.9687	0.9510	0.9565	0.9261	0.9705	0.9496
		NAVCA	**0.9751**	**0.9728**	**0.9682**	**0.9545**	**0.9756**	**0.9804**	**0.9867**	**0.9624**	**0.9727**	**0.9720**
m-MMLU	weighted	MEXA	**0.9720**	0.9678	0.9232	0.9543	0.7293	0.8560	**0.9855**	0.8797	0.8873	0.9061
		NASCA	0.9704	0.9693	0.9541	0.9678	**0.9683**	**0.9849**	0.9846	**0.8871**	**0.9717**	**0.9620**
		NAVCA	0.9702	**0.9700**	**0.9762**	**0.9787**	0.9322	0.9499	0.9842	0.8663	0.9673	0.9550
	average	MEXA	0.9638	0.9622	**0.9300**	0.9708	**0.9170**	**0.9705**	0.9698	**0.9076**	**0.9599**	**0.9502**
		NASCA	**0.9700**	**0.9663**	0.8347	**0.9711**	0.9086	0.9696	**0.9757**	0.9035	0.9578	0.9397
		NAVCA	0.9504	0.9433	0.8086	0.9652	0.8539	0.9557	0.9679	0.8802	0.9277	0.9170
	last	MEXA	0.8443	0.8471	0.8861	0.8448	0.6226	0.6312	0.8614	0.8772	0.8156	0.8034
		NASCA	**0.9675**	**0.9669**	0.9574	0.9697	**0.9597**	**0.9792**	**0.9859**	**0.8860**	**0.9661**	**0.9598**
		NAVCA	0.9611	0.9612	**0.9790**	**0.9719**	0.9170	0.9351	0.9783	0.8504	0.9578	0.9458
Belebele	weighted	MEXA	0.9483	0.9583	0.8108	0.9562	0.6076	0.7422	0.9745	0.9654	0.7229	0.8540
		NASCA	**0.9588**	**0.9614**	**0.9658**	**0.9633**	**0.9444**	**0.9494**	**0.9774**	**0.9699**	**0.9283**	**0.9576**
		NAVCA	0.9087	0.9214	0.9420	0.9339	0.8671	0.8501	0.9301	0.8951	0.8612	0.9011
	average	MEXA	0.9452	0.9525	**0.8498**	0.9580	0.8572	0.8996	0.9685	0.9640	0.8888	0.9204
		NASCA	**0.9526**	**0.9555**	0.7626	**0.9590**	**0.8877**	**0.9416**	**0.9744**	**0.9648**	**0.9334**	**0.9257**
		NAVCA	0.9343	0.9387	0.7438	0.9444	0.8298	0.9016	0.9610	0.9330	0.9104	0.8997
	last	MEXA	0.6675	0.6907	0.7507	0.7202	0.4955	0.4924	0.7448	0.9629	0.5611	0.6762
		NASCA	**0.9600**	**0.9647**	**0.9621**	**0.9686**	**0.9335**	**0.9446**	**0.9796**	**0.9683**	**0.9183**	**0.9555**
		NAVCA	0.9089	0.9190	0.9356	0.9334	0.8569	0.8473	0.9207	0.8745	0.8508	0.8941

Table 4: Pearson correlation of NeuronXA on the FLORES dataset across there multilingual benchmarks. The values in the table represent the correlation of NeuronXA and benchmark settings. The highest average correlations for each task are highlighted in **bold**, and the second highest are underlined.

els and discovered that these neurons are predominantly concentrated in the top and bottom layers of the model. Furthermore, certain studies have focused on dynamic changes. Wang et al. (2024a) and Bhaskar et al. (2024) analyzed the dynamic alignment capabilities of multilingual large models during pretraining.

Cross-lingual Alignment. Cross-lingual alignment can be evaluated by the similarity of representations. Several research has focused on embedding-based approaches. Papadimitriou et al. (2021) investigated morphological and syntactic alignment within embedding spaces, while Wen-Yi and Mimno (2023) studied token-level embedding similarity across models with respect to language-specific encoding patterns. Xu et al. (2023b) and Mousi et al. (2024) explored concept representation alignment in the semantic space. To evaluate cross-lingual alignment through semantic similarity, Li et al. (2025) computed cosine similarity between embeddings of parallel sentences to assess multilingual model performance. Building on this, Kargaran et al. (2024) introduced relative cosine similarity to predict alignment scores and analyzed its correlation with downstream task performance.

Despite these advancements, the representation collapse phenomenon prevalent in neural models compromises semantic expressivity, particularly for low-resource languages (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020), thereby the effectiveness of embedding-based methods for cross-lingual semantic alignment is inherently limited. This limitation is also reflected in the restricted correlation with zero-shot transfer performance observed in earlier methods. Various techniques, such as Canonical Correlation Analysis (Kornblith et al., 2019) and Centered Kernel Alignment (Conneau et al., 2020), have been employed to measure the similarity of intrinsic representations for parallel inputs. The work most closely related to ours is that of SADS (Zeng et al., 2025), who computed cosine similarity based on neuron activation values from parallel sentences as the cross-lingual alignment score. In contrast, our study goes further by analyzing why neuron-based approaches are effective. Furthermore, given the anisotropy issue in neural representations Kargaran et al. (2024), rather than

relying solely on cosine similarity values, we adopt a binary perspective. This approach ensures more reliable assessments of alignment.

6 Conclusion

In this paper, we propose a novel cross-lingual alignment evaluation method, Neuron State Similarity-Based Cross-Lingual Alignment (*NeuronXA*), which offers a more semantically grounded approach compared to traditional methods. By leveraging NeuronXA, we assess a model's alignment ability based on the consistency of parallel sentences. Through extensive experiments, we analyze the Pearson correlation between the NeuronXA score and three downstream tasks, as well as a zero-shot cross-lingual transfer task. Our results demonstrate that the NeuronXA score is strongly correlated with both the model's transferability and its performance on multilingual tasks.

While NeuronXA demonstrates robust performance across a variety of settings, it achieves the highest alignment scores when combined with token-weighted average methods and the NASCA score evaluation approach. Notably, in the transfer task, the average Pearson correlation reaches 0.9556, while the correlation with multilingual tasks is 0.8524, highlighting the effectiveness of NeuronXA in capturing cross-lingual alignment.

Overall, NeuronXA demonstrates significant potential as a robust method for evaluating the multilingual capabilities of LLMs, paving the way for future efforts to expand these models to a wider range of underrepresented languages.

Limitations

In this study, we employ neuron states as intrinsic representations to evaluate alignment by examining the consistency of parallel sentences within the representation space. Therefore, a limitation of our evaluation method is its requirement for access to the model's intrinsic representations. Consequently, developers of closed-source models may be unable to directly apply NeuronXA. Nevertheless, they could utilize NeuronXA internally and report their results, which would provide valuable insights into their model's cross-lingual capabilities.

Moreover, various perspectives on the capabilities of large models offer alignment across different abilities. However, NeuronXA cannot encompass all of these aspects. Our goal is to provide a simple yet effective evaluation method for multilingual alignment in large models, contributing insights for future research on cross-lingual alignment and multilingual mechanisms.

Acknowledgement

We would like to thank all the anonymous reviewers for the insightful and helpful comments. This work is supported by National Science and Technology Major Project (Grant No. 2022ZD0116101), the Major Scientific Research Project of the State Language Commission in the 13th Five-Year Plan (Grant No. WT135-38), and the public technology service platform project of Xiamen City (No. 3502Z20231043).

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

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

Adithya Bhaskar, Dan Friedman, and Danqi Chen. 2024. The heuristic core: Understanding subnetwork generalization in pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14351–14368.

Sunit Bhattacharya and Ondřej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 120–126, Singapore. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind

Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

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

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.

Maksym Del and Mark Fishel. 2022. Cross-lingual similarity of multilingual representations revisited. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 185–195, Online only. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024a. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Biao Fu, Minpeng Liao, Kai Fan, Chengxi Li, Liang Zhang, Yidong Chen, and Xiaodong Shi. 2025a. Llms can achieve high-quality simultaneous machine translation as efficiently as offline. *Preprint*, arXiv:2504.09570.

Biao Fu, Donglei Yu, Minpeng Liao, Chengxi Li, Yidong Chen, Kai Fan, and Xiaodong Shi. 2025b. Efficient and adaptive simultaneous speech translation with fully unidirectional architecture. *Preprint*, arXiv:2504.11809.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

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.

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.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *CoRR*.

Katharina Hämmerl, Jindřich Libovickỳ, and Alexander Fraser. 2024. Understanding cross-lingual alignment—a survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10922–10943.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2024. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. *arXiv preprint arXiv:2410.05873*.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. 2024. Prealign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10257.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. 2024. Exploring alignment in shared cross-lingual spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6326–6348, Bangkok, Thailand. Association for Computational Linguistics.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 2522–2532, Online. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1288–1301.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024a. Probing the emergence of cross-lingual alignment during llm training. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12159–12173.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

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.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2023b. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13951–13976, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9386–9393.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. 2025. How well do large reasoning models translate? a comprehensive evaluation for multi-domain machine translation. *Preprint*, arXiv:2505.19987.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *CoRR*.

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.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *Preprint*, arXiv:2411.09116.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. 2023. Emergent modularity in pre-trained transformers. In *Proceedings of ACL: Findings*, pages 4066–4083.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuan-Jing 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.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8411–8423.

A Dataset

A.1 Parallel Datasets

FLORES-200. This multilingual parallel corpus consists of English sentences sampled in equal proportions from Wikinews, Wikijunior, and Wikivoyage. Each sentence has been translated into more than 200 languages, with data quality ensured through a combination of automated validation and human review. Since the test set is not publicly available, our experiments are conducted on the dev-test set, which consists of 1,012 sentences covering 213 languages. We set the 68 languages: "bel_Cyrl, bos_Latn, hun_Latn, epo_Latn, khm_Khmr, urd_Arab, srp_Cyrl, jav_Latn, hye_Armn, gla_Latn, por_Latn, lit_Latn, bul_Cyrl, slk_Latn, mal_Mlym, ita_Latn, nno_Latn, mar_Deva, hrv_Latn, hin_Deva, kat_Geor, ben_Beng, fin_Latn, cym_Latn, oci_Latn, cat_Latn, fao_Latn, xho_Latn, spa_Latn, ron_Latn, amh_Ethi, ces_Latn, swe_Latn, nld_Latn, tat_Cyrl, kor_Hang, glg_Latn, fra_Latn, eus_Latn, ind_Latn, dan_Latn, tha_Thai, deu_Latn, tel_Telu, afr_Latn, pol_Latn, est_Latn, uig_Arab, ukr_Cyrl, uzn_Latn, heb_Hebr, kaz_Cyrl, nob_Latn, rus_Cyrl, vie_Latn, arb_Arab, zho_Hans, tuk_Latn, khk_Cyrl, jpn_Jpan, ell_Grek, isl_Latn, tam_Taml, slv_Latn, tur_Latn, mkd_Cyrl, tgl_Latn, gle_Latn" as "Head" languages, and the remaining 135 languages (excluded English data) as "Long-tail" ones.

Tatoeba. The Tatoeba dataset(Artetxe and Schwenk, 2019) serves as a benchmark for evaluating multilingual sentence embeddings in similarity search tasks. It covers 112 languages and provides up to 1,000 English-aligned sentence pairs for each language. The evaluation is performed by computing cosine similarity to retrieve the nearest neighbors of each sentence in other languages, followed by calculating the error rate. We treat the 36 languages contained in XTREME(Hu et al., 2020) as head languages, which are: "ar, he, vi, id, jv, tl, eu, ml, ta, te, af, nl, en, de, el, bn, hi, mr, ur, fa, fr, it, pt, es, bg, ru, ja, ka, ko, th, sw, zh, kk, tr, et, fi, hu, az, lt, pl, uk, ro". The remaining 76 languages in Tatoeba are treated as long-tail ones.

A.2 Multilingual Benchmarks

Belebele. A multilingual multiple-choice machine reading comprehension dataset spanning 122 language variants. It evaluates both monolingual and multilingual models across resource-rich and resource-scarce languages. Each item consists of a question, four answer choices, and a passage sourced from FLORES-200. The dataset is meticulously annotated to distinguish proficiency levels, with rigorous quality control measures. Since five languages in Belebele are not present in FLORES-200, our analysis focuses on the 117 overlapping languages.

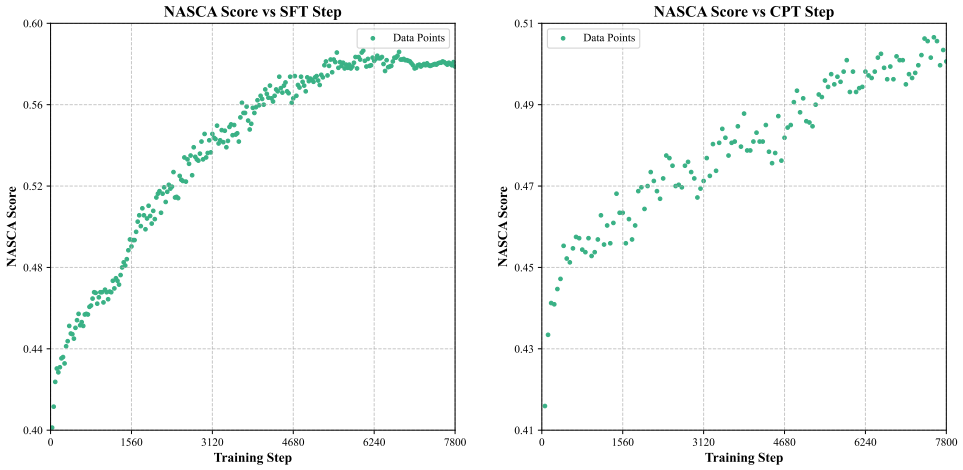
m-ARC. The Multilingual AI2 Reasoning Challenge extends the original English ARC benchmark(Clark et al., 2018) to assess cross-lingual scientific reasoning. It consists of systematically translated multiple-choice questions in 31 languages, generated using GPT-3.5-Turbo. The dataset includes 1,116 training items, 1,169 test items, and 298 validation items, all aligned with scientific reasoning objectives and grade-school science curricula.

m-MMLU. A multilingual extension of the MMLU benchmark (Hendrycks et al., 2021a), covering 34 languages. The dataset was initially translated into 31 languages using GPT-3.5-Turbo, with expert translations for Icelandic and Norwegian. It contains 277 training items, 13,258 test items, and 1,433 validation items, spanning four domains: humanities, social sciences, STEM disciplines, and professional subjects. As the most comprehensive multilingual knowledge benchmark, m-MMLU provides a robust evaluation of cross-lingual understanding.

B Generative Tasks Evaluation

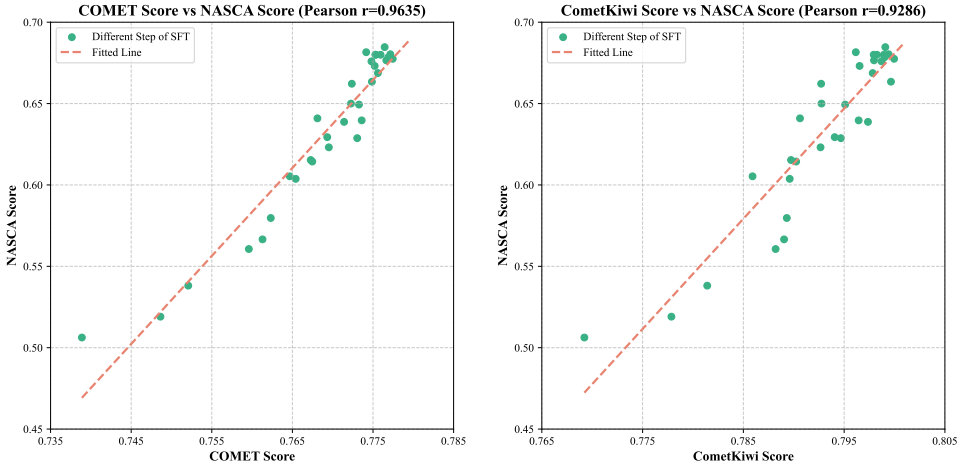
Certain generation tasks are strongly correlated with a model's cross-lingual alignment capabilities. In the context of machine translation, several training paradigms have been proposed to enhance a model's ability to map low-resource languages into a unified representation space with high-resource languages (Xu et al., 2023a; Guo et al., 2024). These approaches aim to improve the model's understanding of low-resource languages, fostering emergent multilingual alignment during fine-tuning.

Given this, we hypothesize that a model's translation performance is closely related to its alignment ability. To test this, we selected the NLLB (Costa-jussà et al., 2022) dataset, specifically 1 million sentence pairs of English and Icelandic (with a 1:1 ratio), and fine-tuned the model with 4-bit



(a) Supervised fine-tuning. (b) Continue Pre-training.

Figure 3: Alignment Score Trends During Supervised Fine-Tuning and Continued Pre-Training of LLaMA-3.1 8B.



(a) COMET. (b) CometKiwi.

Figure 4: Correlation coefficients between Alignment Scores and COMET/CometKiwi Scores during Supervised Fine-Tuning of LLaMA-3.1 8B.

quantized LoRA for supervised training. We used the NAS to assess alignment at each fine-tuning step and calculated the Pearson correlation between the alignment scores and the COMET/CometKiwi scores at each step.

Fine-tuning facilitates the alignment of the model to a unified representation space. As shown in Figure 3, the alignment scores increase with fluctuations during the fine-tuning process, indicating that fine-tuning promotes the alignment of languages in the training data into a shared space.

NeuronXA is closely related to machine translation performance. In our analysis, we computed the Pearson correlation between COMET/CometKiwi scores and alignment scores at each step, resulting in a correlation coefficient of 0.9635 and 0.9286, respectively. This strong correlation indicates that alignment scores are highly indicative of translation performance. Furthermore, alignment serves as a valuable metric for evaluating the model's translation capabilities.

C Other Baselines

C.1 Exploring Other Languages as Base Languages

In the domain of multilingual modeling, English was selected as the primary base language for this study due to its predominant role in mainstream multilingual models. Nevertheless, we acknowledge the importance of evaluating the generalizabil-

Base Language	Representation	Head			Long-tail		
		src → xx	xx → src	src ⇔ xx	src → xx	xx → src	src ⇔ xx
French	Embedding	86.31	81.16	78.53	49.69	41.14	38.74
	NAS	**93.35**	**88.29**	**86.03**	**51.57**	**45.91**	**42.02**
Italian	Embedding	85.96	81.99	78.77	49.30	42.27	39.15
	NAS	**92.48**	**88.60**	**85.84**	**50.92**	**46.71**	**41.89**
German	Embedding	86.69	82.51	79.48	50.16	43.15	40.05
	NAS	**94.40**	**89.00**	**87.01**	**52.53**	**46.54**	**42.48**

Table 5: Retrieval results on FLORES-200 in xx → src and src → xx direction, along with src ⇔ xx direction. The bold font denotes the best results.

Baselines	Languages			
	English	German	French	Italian
MEXA	0.9585	0.9250	0.9389	0.9356
NASCA	**0.9621**	**0.9287**	**0.9460**	**0.9428**

Table 6: Average Pearson correlation of MEXA and NeuronXA across marc, mmlu and belebele tasks.

Base Language	Baselines	XNLI	Bmlama	Average
English	MEXA	0.9259	0.6567	0.7913
	NASCA	**0.9309**	**0.6825**	**0.8067**
German	MEXA	**0.8993**	0.5613	0.7303
	NASCA	0.8986	**0.6656**	**0.7821**

Table 7: Average Pearson correlation of MEXA and NeuronXA across XNLI, Bmlama tasks.

ity of our method to other high-resource languages. To this end, we conducted a series of experiments, including semantic retrieval, downstream task performance correlation, and cross-lingual transferability correlation.

Specifically, for cross-lingual retrieval and downstream task correlation experiments, we employed German, French, and Italian as base languages on the LLaMA-3.1 8B model. The transferability correlation experiments were conducted using German as the base language.

As shown in Table 5, when using these high-resource languages for semantic retrieval, the NAS-based method consistently outperformed the embedding-based approach in retrieval accuracy across all three languages. These results align with our English-based findings, suggesting that the NAS-based method generalizes well to other base languages.

We further computed NASCA scores using German, French, and Italian as base languages and evaluated their correlation with downstream task performance. As presented in Table 6, the NASCA scores maintained strong correlations even with non-English base languages.

Finally, we assessed the relationship between alignment scores and cross-lingual transferability using German. The results, reported in Table 7, further confirm the robustness and cross-linguistic applicability of our approach.

C.2 Other Baselines

To verify the advantages of NeuronXA in interpreting model downstream task performance and transferability, we conducted comparisons with three representation similarity-based evaluation methods. Table 8 and Table 9 present the correlation coefficients between alignment scores and downstream task performance, and between alignment scores and model transferability, respectively. Table 10 shows the average correlation coefficients of all baselines with downstream task performance and model transferability. The results indicate that NeuronXA outperforms others in interpreting both multilingual capabilities and cross-lingual transferability, confirming the effectiveness and robustness of our method.

Centered Kernel Alignment (CKA). CKA (Kornblith et al., 2019) is a similarity measure rooted in the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), a non-parametric method designed to assess the independence among random variables. CKA serves as a second-order similarity index, functioning by comparing the subspaces spanned by neurons, which endows it with robust power for comparing representations across different networks. Its theoretical foundation lies in identifying dominant correlation directions within distinct datasets and conducting comparisons based on these directions. Furthermore, CKA can be adjusted to a weighted version by incorporating eigenvalues, thereby giving rise to Linear CKA. By design, CKA is intended to exhibit invariance with

		Llama 3.1 8B	Llama 2 7B	Llama 3.2 3B	Qwen 2.5 7B	Mistral 0.3 7B	OLMo 2 7B	GLM 4 9B	AVG
m-ARC	CKA	0.8333	0.8328	0.7711	0.7989	0.8774	0.8109	0.8955	0.8314
	SVCCA	0.9303	0.9189	0.8865	0.9172	0.9329	0.8519	0.9370	0.9107
	ANC	0.9683	0.9385	0.9305	0.9633	0.9659	0.9025	0.9690	0.9483
	MEXA	0.9551	0.9124	0.9142	0.9589	0.9575	0.8925	0.9225	0.9304
	NASCA	0.9570	0.9369	0.9186	0.9479	0.9539	0.9177	0.9713	0.9433
	NAVCA	**0.9756**	**0.9649**	**0.9522**	**0.9820**	**0.9847**	**0.9569**	**0.9731**	**0.9699**
m-MMLU	CKA	0.8779	0.8642	0.8468	0.8836	0.9256	0.8745	0.9451	0.8882
	SVCCA	0.9478	0.9691	0.9454	0.9202	0.9693	**0.8936**	0.9474	0.9418
	ANC	0.9522	0.9760	0.9619	0.9051	0.9770	0.8907	0.9618	0.9464
	MEXA	**0.9720**	0.9232	0.9543	0.8560	**0.9855**	0.8797	0.8873	0.9226
	NASCA	0.9704	0.9541	0.9678	**0.9849**	0.9846	0.8871	**0.9717**	**0.9601**
	NAVCA	0.9702	**0.9762**	**0.9787**	0.9499	0.9842	0.8663	0.9673	0.9561
Belebele	CKA	0.4751	0.5824	0.5239	0.3463	0.6317	0.9260	0.4293	0.5592
	SVCCA	0.9157	0.9530	0.9354	0.8732	0.9431	0.9439	0.8858	0.9214
	ANC	0.9022	0.9378	0.9331	0.8324	0.9313	0.9257	0.8319	0.8992
	MEXA	0.9483	0.8108	0.9562	0.7422	0.9745	0.9654	0.7229	0.8743
	NASCA	**0.9588**	**0.9658**	**0.9633**	**0.9494**	**0.9774**	**0.9699**	**0.9283**	**0.9590**
	NAVCA	0.9087	0.9420	0.9339	0.8501	0.9301	0.8951	0.8612	0.9030

Table 8: Pearson correlation of NeuronXA on the FLORES dataset across there multilingual benchmarks. The values in the table represent the correlation of NeuronXA and benchmark. The highest average correlations for each task are highlighted in **bold**, and the second highest are underlined.

respect to data scaling, centering, and orthogonal transformations, and it maintains its stability even under any invertible linear transformations of the data.

Singular Value Canonical Correlation Analysis (SVCCA). SVCCA is a method introduced by Raghu et al. (2017) for comparing learned representations in neural networks. It combines Singular Value Decomposition (SVD) and Canonical Correlation Analysis (CCA) to provide an efficient and invariant way to compare representations. The approach first applies SVD to each set of neurons to identify the most significant directions that explain the majority of the variance in the data. Then, CCA is used to find linear transformations that maximize the correlation between these subspaces from different layers or networks. SVCCA is designed to be invariant to affine transformations, making it suitable for comparisons across different architectures and training stages.

Averaged Neuron-Wise Correlation (ANC). The ANC method, introduced by Del and Fishel (2022), offers a novel approach to analyzing cross-lingual similarity in multilingual language models. It is based on the assumption that neurons in the representations of different languages are aligned one-to-one a priori. ANC calculates the correlations between pairs of neurons from different languages and then averages these correlations to generate a similarity score. Compared to other methods, ANC provides improved interpretability by enabling the identification of specific neurons that contribute the most or the least to the similarity.

D NeuronXA Score for Other Datasets

We examine the model's evaluation results on other datasets, specifically using the Tatoeba dataset. Additionally, we explore the Pearson correlation coefficients between alignment scores and three multilingual benchmarks, as well as the correlation coefficients with zero-shot cross-lingual transfer performance.

As shown in Table 11, NeuronXA achieves relatively high correlation coefficients compared to sentence embeddings, suggesting that NeuronXA is a more generalizable method that can be applied across different datasets.

It is important to note that the quality of the bilingual datasets used for NeuronXA evaluation—particularly their distribution and diversity—can influence the alignment scores. Ideally,

		Llama 3.1 8B	Llama 2 7B	Llama 3.2 3B	Qwen 2.5 7B	Mistral 0.3 7B	OLMo 2 7B	GLM 4 9B	AVG
XNLI	CKA	0.6694	0.6612	0.6879	0.4022	0.6604	0.8944	0.6601	0.6622
	SVCCA	0.8800	0.8846	0.9144	0.7610	0.8897	0.8714	0.8797	0.8687
	ANC	0.8645	0.8815	0.9082	0.7751	0.8784	0.8534	0.8775	0.8627
	MEXA	0.9259	0.7519	0.9182	0.6898	0.9446	**0.9500**	0.8107	0.8559
	NASCA	**0.9309**	**0.9639**	0.9401	**0.9209**	**0.9583**	0.9411	**0.9467**	**0.9431**
	NAVCA	0.9227	0.9063	**0.9430**	0.8415	0.8982	0.8777	0.9213	0.9015
BMLAMA	CKA	0.6694	0.8280	0.5313	0.7855	0.6893	0.7196	0.6662	0.6985
	SVCCA	**0.8800**	0.8516	0.7039	0.8608	0.8247	0.8234	0.7441	0.8126
	ANC	0.8645	0.8895	0.7239	0.8663	0.8646	**0.9156**	**0.7891**	**0.8448**
	MEXA	0.6567	0.8426	0.6223	0.8473	0.8795	0.7500	0.6922	0.7558
	NASCA	0.6825	0.8707	0.6748	0.8575	0.8750	0.7975	0.6761	0.7763
	NAVCA	0.7285	**0.8924**	**0.7361**	**0.8773**	**0.9062**	0.8871	0.6850	0.8161

Table 9: Pearson correlation of NeuronXA on the FLORES dataset across ZS-CLT and CLKA tasks. The values in the table represent the correlation of benchmarks. The highest average correlations for each task are highlighted in **bold**, and the second highest are underlined.

Baselines	Multilingual performance	Cross-lingual transferability
CKA	0.7596	0.6804
SVCCA	0.9246	0.8407
ANC	0.9313	0.8537
MEXA	0.9091	0.8058
ours		
NASCA	**0.9541**	**0.8597**
NAVCA	0.9430	0.8588

Table 10: Average Pearson correlation of several baselines across Multilingual performance and Cross-lingual transferability tasks.

the greater the diversity of the dataset, the more accurately NeuronXA reflects the alignment of semantic knowledge across languages. Despite the relatively lower diversity of the Tatoeba dataset, as evidenced in Table 11, NeuronXA still achieves a reasonably high correlation coefficients, further validating the robustness of our approach.

E Robustness of NeuronXA

Similar to the discussion of MEXA (Kargaran et al., 2024), NeuronXA scores are highly robust, with a very low probability of achieving randomly high values. Our matrix $\mu_{C(l)}$ measures the alignment scores of matrix $C(l)$, specifically the proportion of diagonal elements that attain the maximum value within their respective rows and columns. We assume the existence of an n-dimensional matrix $C(l)$, with k elements satisfying this condition. For an $N \times N$ matrix, the probability of diagonal elements being the maximum value in both their row and column is given by $p = \frac{1}{2n-1}$.

$$P(X \geq \frac{k}{n}) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-i} \quad (5)$$

Assuming the diagonal elements are the maximum in both their row and column, the probability that at least k of the n independent variables satisfy this condition can be computed using the binomial distribution formula in 5. This formula suggests that, given a sufficient number of parallel sentences (n), the likelihood of achieving a high score by chance is very low. For example, with $n = 100$, the probability of obtaining a NeuronXA alignment score greater than 0.05 (with $k = 5$) from a random 100×100 matrix is $p(x \geq 0.05) = 0.00016$.

			Llama 3.1 8B	Llama 3 8B	Qwen 2.5 14B	Mistral 0.3 7B	OLMo 2 7B	GLM 4 9B	AVG
m-ARC	weighted	MEXA	**0.8274**	0.8264	0.8140	0.8961	0.9046	0.8043	0.8455
		NASCA	0.7197	0.9134	0.9011	0.8825	0.9046	0.9382	0.8766
		NAVCA	0.8102	**0.9364**	**0.9419**	**0.9685**	**0.9446**	**0.9447**	**0.9244**
	average	MEXA	**0.7405**	0.7086	0.9036	**0.9042**	0.8783	**0.9261**	**0.8436**
		NASCA	0.6869	0.8335	**0.9347**	0.7991	0.8795	0.8800	0.8356
		NAVCA	0.7075	**0.8464**	0.9152	0.8118	**0.8956**	0.8838	0.8434
	last	MEXA	0.8119	0.8139	0.8172	0.8487	0.9114	0.8369	0.8400
		NASCA	0.7200	**0.9152**	0.9121	0.9261	0.9157	**0.9451**	0.8890
		NAVCA	**0.8392**	0.9134	**0.9347**	**0.9563**	**0.9548**	0.9413	**0.9233**
m-MMLU	weighted	MEXA	0.7644	0.7627	0.5269	0.7119	0.8272	0.6717	0.7108
		NASCA	**0.9155**	**0.9168**	**0.9201**	0.7813	**0.8597**	**0.9143**	**0.8846**
		NAVCA	0.9069	0.9086	0.8609	**0.8706**	0.7997	0.9029	0.8749
	average	MEXA	0.7357	0.7241	**0.8738**	**0.9295**	**0.8652**	**0.8833**	**0.8353**
		NASCA	0.8404	**0.8398**	0.8654	0.7134	0.8645	0.8508	0.8291
		NAVCA	**0.8421**	0.8345	0.8601	0.6814	0.8434	0.8260	0.8146
	last	MEXA	0.7267	0.7246	0.4953	0.6359	0.8354	0.7223	0.6900
		NASCA	**0.9131**	**0.9128**	**0.8938**	**0.8608**	**0.8549**	**0.9180**	**0.8922**
		NAVCA	0.8742	0.8719	0.8313	0.8066	0.7684	0.8899	0.8404
Belebele	weighted	MEXA	0.6424	0.6553	0.4100	0.6180	0.9159	0.5104	0.6253
		NASCA	**0.8952**	**0.9039**	**0.9179**	0.7659	**0.9246**	**0.8360**	**0.8739**
		NAVCA	0.8282	0.8440	0.7941	**0.8180**	0.8120	0.7867	0.8139
	average	MEXA	0.7977	0.7950	0.8553	**0.9351**	0.8948	0.8091	**0.8478**
		NASCA	**0.8571**	**0.8510**	**0.8880**	0.7294	**0.9030**	**0.8209**	0.8416
		NAVCA	0.8340	0.8277	0.8406	0.6927	0.8627	0.7943	0.8087
	last	MEXA	0.5997	0.6127	0.4006	0.5301	**0.9147**	0.5591	0.6028
		NASCA	**0.8903**	**0.9000**	**0.8763**	**0.8408**	0.9107	**0.8208**	**0.8732**
		NAVCA	0.7979	0.8101	0.7662	0.7471	0.7639	0.7689	0.7757
BMLAMA-53	weighted	MEXA	**0.7394**	0.7402	0.6949	0.7553	0.7980	0.7283	0.7427
		NASCA	0.7223	0.7314	0.8074	0.6982	0.8489	0.7304	0.7564
		NAVCA	0.7377	**0.7434**	**0.8359**	**0.8488**	**0.9158**	**0.7463**	**0.8047**
	average	MEXA	0.7378	0.7238	0.8177	**0.8390**	0.8926	0.7612	**0.7954**
		NASCA	0.7266	0.7442	0.8085	0.6698	0.8809	0.7263	0.7594
		NAVCA	**0.7450**	**0.7530**	**0.8383**	0.6683	**0.9025**	**0.7815**	0.7814
	last	MEXA	0.7232	0.7259	0.6669	0.6794	0.8115	0.7513	0.7264
		NASCA	0.7135	0.7106	0.8238	0.8035	0.8633	0.7349	0.7749
		NAVCA	**0.7480**	**0.7479**	**0.8333**	**0.8061**	**0.9108**	**0.7531**	**0.7999**

Table 11: Pearson correlation of NeuronXA on the Tatoeba dataset across there multilingual benchmarks and one Cross-language transfer task. The values in the table represent the correlation of NeuronXA and benchmark settings. The highest average correlations for each task are highlighted in **bold**, and the second highest are underlined.