High-Dimensional Interlingual Representations of Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) trained on massive multilingual datasets hint at the formation of interlingual constructs–a shared subspace in the representation space. However, evidence regarding this phenomenon is mixed, leaving it unclear whether these models truly develop unified interlingual representations, or present a partially aligned constructs. We explore 31 diverse languages varying on their resource-levels, typologies, and geographical regions; and find that multilingual LLMs exhibit inconsistent cross-lingual alignments. To address this, we propose an interlingual representation framework identifying both the shared interlingual semantic subspace and fragmented components, existed due to representational limitations. We introduce Interlingual Local Overlap (ILO) score to quantify interlingual alignment by comparing the local neighborhood structures of high-dimensional representations. We utilize ILO to investigate the impact of single-language fine-tuning on the interlingual representations in multilingual LLMs. Our results indicate that training exclusively on a single language disrupts the alignment in early layers, while freezing these layers preserves the alignment of interlingual representations, leading to improved cross-lingual generalization. These results validate our framework and metric for evaluating interlingual representation, and further underscore that interlingual alignment is crucial for scalable multilingual learning.

1 Introduction

Interlingua, a universal language-neutral representation, is pivotal for cross-lingual generalization. Grounded in both linguistic theories and computational practice, this concept aims to treat languages equitably and capture universal semantic structures independent of any specific language (Richens, 1958; Vauquois, 1968; Schubert, 1989; Rayner et al., 2010a; Johnson et al., 2017). The advent

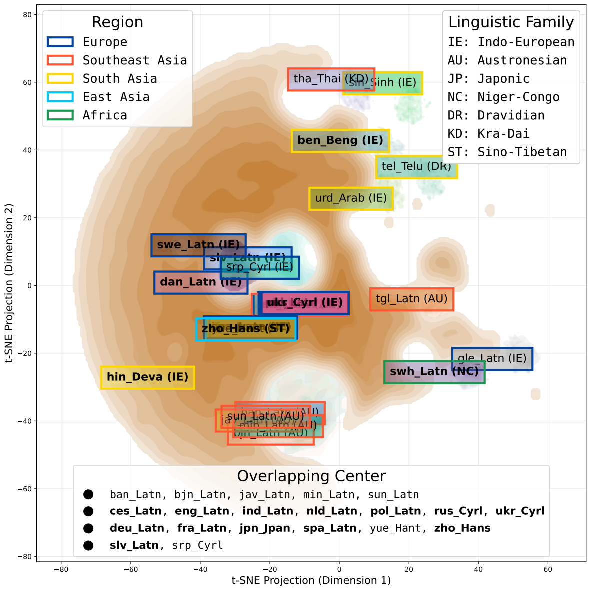


Figure 1: Interlingual overlaps transcending familial and regional boundaries in the intermediate layer of multilingual LLMs, observed in a t-SNE visual of Aya Expanse (8B) hidden states (layer 16). HRLs are **bold**.

of LLMs trained on extensive multilingual corpora suggests the potential of interlingual constructs naturally emerging without any explicit objectives (Conneau et al., 2020a; Chang et al., 2022; Moschella et al., 2023; Wendler et al., 2024). This is attributed to their ability to map representations from different languages into a shared multilingual embedding space (Pires et al., 2019; Libovickỳ et al., 2020; Conneau et al., 2020b; Muller et al., 2021; Zhao et al., 2024; Zeng et al., 2025).

However, evidence remains mixed on whether they converge all language-specific representations into a unified single interlingua representation space, and raising questions about whether LLMs can retain the interlingua representations in diverse linguistic typology, geographical distribution, and resource-level settings. It is unclear whether LLMs form a unified interlingual construct or if fragmentation occurs across different language groups. A critical question persists: Do LLMs develop a universal interlingua representation, or present a partially aligned construct toward certain languages?

Our preliminary experiments reveal that LLMs represent parallel semantic input differently across languages. Notably, their neuron activations align better within high-resource pairs and the same familial or regional roots, suggesting that LLMs exhibit varying alignment consistencies across differing language groups. Building upon these insights, we introduce a novel interlingual representation framework aimed at enhancing the understanding of how LLMs encapsulate interlingual semantics. Our framework identifies both the core subspace that captures shared semantics across languages, and address fragmented components due to representational limitations underscoring the importance of interlingual alignment across diverse linguistic contexts. With the framework, we introduce a novel metric, Interlingual Local Overlap (ILO), which quantifies intrinsic interlingual alignment consistencies by comparing the local neighborhood structures of high-dimensional representations. ILO is calculated as the harmonic mean of two measurements, inspired by graph theory (Guimera and Amaral, 2005; Freeman et al., 2002; Borgatti and Everett, 2006), on **bridge**: the extent to which representations of a given language neighboring diverse other languages, and **reachability**: the degree of connectivity these representations have with other languages within the multilingual space.

We demonstrate the effectiveness our framework and metric through an in-depth analysis of LLMs' internal states on a multilingual mathematical reasoning task, chosen for its language-agnostic properties. We first observe that training multilingual LLMs on a single-language causes catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999; Biesialska et al., 2020) degrading their cross-lingual generalization (Liu et al., 2021; Winata et al., 2023). These degradations are correlated with the disruption of interlingual alignment that originate in the early layers of LLMs. To ensure the preservation of interlingual alignments, we adopt a strategy of selectively-freezing parameters during the single-language fine-tuning. Evaluations using ILO highlight that this approach effectively safeguards the interlingual alignments across all layers and maintains the levels observed prior to training, which results in significant improvements in cross-lingual generalization. Our findings underscore the pivotal role of interlingual semantic alignment in the pursuit of scalable multilingual learning.

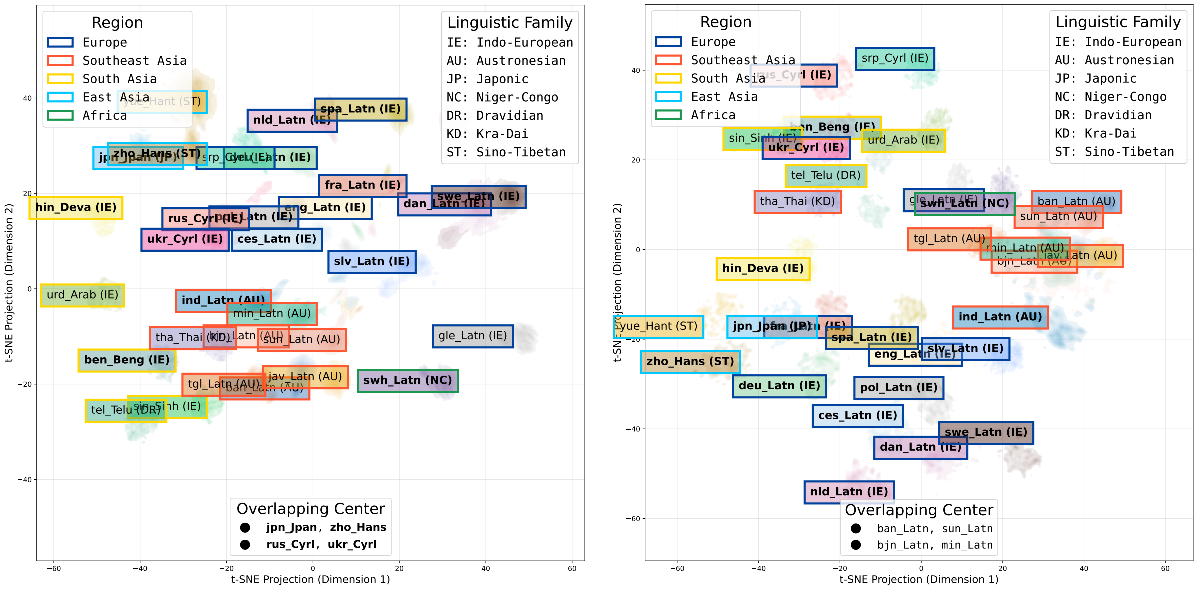
Properties	Details
Resources	High: 18 / Low: 13
Families	Indo-European: 18 / Austronesian: 7 / Sino-Tibetan: 2 / Japonic: 1 / Niger-Congo: 1 / Dravidian: 1 / Kra-Dai: 1
Regions	Europe: 14 / Southeast Asia: 8 / South Asia: 5 / East Asia: 3 / Africa: 1

Table 1: Distribution of the 31 languages across families, regions, and resource-levels in our analysis, sampled from Flores-200 (see Table A1 for complete details).

2 Related Works

Syntactical Interlingua Representations Interlingua has played a huge role throughout the development of NLP. Various representations of interlingua have been developed along with the advancement of NLP. In the early years, a logically formalized interlingua representation for mechanical translation has been proposed (Richens, 1958; Vauquois, 1968). In the early days, interlingua is presented as delexicalized grammar extracted from the original text that can be mapped to other language interlingua delexicalized grammar. In this case, each language has its own interlingua form which can then be mapped into other language with a dictionary lookup (Richens, 1958; Rayner et al., 2010b). A more sophisticated method involves interlingua representation as a common abstract syntax that are shared across all languages (Rayner et al., 2008; Kanzaki et al., 2008). This method has been applied in various systems such as Spoken Languge Translator (Rayner, 2000), PARC's XLE (Riezler et al., 2002), and Verbmobil (Wahlster, 2013). Despite its advancement, this method tends to be incomplete and difficult to scale to new languages (Ranta et al., 2020).

Semantic Interlingua Representations With the rise of statistical machine translation (Brown et al., 1990; Och et al., 1999; Lopez, 2008) and cross-lingual alignment (Brown et al., 1991; Och and Ney, 2003; Mikolov et al., 2013; Miceli Barone, 2016; Artetxe and Schwenk, 2019), methods for representing interlingua using latent semantic vectors become more prominent (Fung and Chen, 2004; Fung and Mckeown, 1994; Fung and Church, 1994; Seneff, 2006). Methods involving specialized objectives to construct better semantic interlingua representations have also been proposed (Lu et al., 2018; Al-Shedivat and Parikh, 2019; Zhu et al., 2020; Wei et al., 2021; Feng et al.,



(a) Early (Layer 0)

(b) Late (Layer 32)

Figure 2: Embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. In these early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.

2022; Cahyawijaya et al., 2023, 2024b). In recent years, various studies have showcased that current LLMs inherit such interlingua representation (Muller et al., 2021; Chang et al., 2022; Moschella et al., 2023; Zhao et al., 2024; Wendler et al., 2024) which enables LLMs to process sentences with a single shared representation across different languages. However, the characteristics of this representation in LLMs remain unexplored. This research aims to explore the extent of this interlingua representation offering a novel perspective on interlingual representation in LLMs.

3 Interlingual Representations in Multilingual LLMs

To explore the emergence of interlingual representation in LLMs, we assess the semantic alignment of their hidden states to understand whether the latent structures capture universal semantics across languages. We presume that multilingual LLMs adhere to a "first align, then predict" pattern (Muller et al., 2021) and that their aligned states represent semantically similar features across languages. Ideally, these features map parallel semantic inputs from many languages to similar vector representations that overlaps in the high-dimensional space.

Consider the high-dimensional representation space $\mathcal{H} \subseteq \mathbb{R}^d$ learned by LLMs, where d is the model's embedding dimension. For an input $\mathbf{x}$ in language ℓ, the model uses language-specific encoding functions $f_\ell(\mathbf{x}) \in \mathcal{H}$. Here, $\mathcal{H}$ serves as a shared multilingual space where different encoding functions $f_\ell(\mathbf{x})$ align semantic and syntactic patterns across languages. Building on this, we define semantic alignment α of representations from parallel inputs $\mathbf{x}$ and $\mathbf{x}'$ in languages ℓ and ℓ' as:

$$\alpha(\ell, \ell') = \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{D}_{\ell,\ell'}}\left[\phi\left(f_\ell(\mathbf{x}), f_{\ell'}(\mathbf{x}')\right)\right].$$

Here, ϕ denotes a similarity function and $\mathcal{D}_{\ell,\ell'}$ is the distribution of semantically equivalent input pairs. A higher $\alpha(\ell, \ell')$ indicates better alignment.

3.1 Multilingual Shared Representation Space

We introduce a novel conceptual framework that reevaluates the representation space $\mathcal{H}$ within multilingual LLMs. Specifically, we posit an interlingual representation framework that incorporates an intricate internal structure influenced by inherent model representational limitations. This framework highlights that the quality of alignment among representations may vary, leading to latent discrepancies that may stem from differences in resource availability or language-specific properties. Formally, we conceptualize $\mathcal{H}$ as a shared space that includes both core and fragmented components in

$$\mathcal{H} \supset \mathcal{M}_c \cup \bigcup_{i \in N} \mathcal{M}_{f_i}.$$

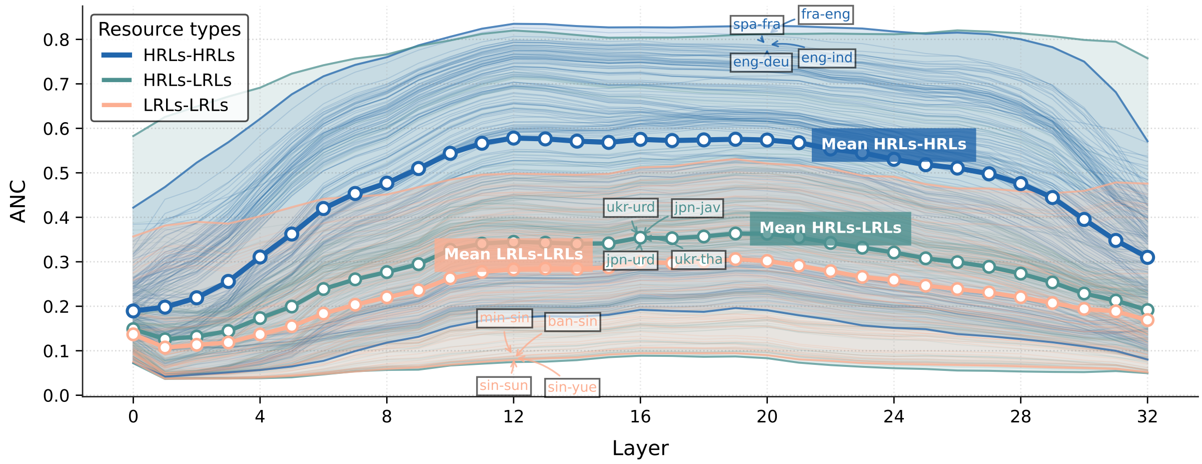


Figure 3: Comparisons of per-layer ANC scores on Aya Expanse (8B) with highlights on pairs w.r.t their resource-levels. Pairs of HRLs demonstrate strong correlations, while pairs involving LRLs exhibit lower ANC scores.

The core component $\mathcal{M}_c$ is an interlingual subspace that encapsulates shared semantics across languages. In contrast, the fragmented $\mathcal{M}_{f_i}$ represent subspaces where alignment with $\mathcal{M}_c$ is challenging. This framework refines the "first align, then predict" paradigm, noting that while LLMs converge inputs from some languages to a shared interlingual subspace, others remain partially aligned.

3.2 Core Interlingual Subspace

Conceptually, we define $\mathcal{M}_c$ as a subspace that encodes universal semantic structures and syntactic abstractions. By positioning multilingual embeddings in this shared space, LLMs effectively learns interlingual semantic representations that facilitate multilingual performance. This is where key interlingual alignments form, enabling LLMs to leverage universal patterns for multilingual tasks.

3.3 Fragmented Subspaces

While some languages enjoy substantial overlaps in $\mathcal{M}_c$, the less-aligned others occupy fragmented subspaces $\mathcal{M}_{f_i}$ as they reflect model's representational limitation to embed the languages into $\mathcal{M}_c$. Factors such as sparse training data, typological distance, and morphological complexity might lead to partial alignment of these representations. Consequently, embeddings in $\mathcal{M}_{f_i}$ tend to be more weakly aligned to the universal semantics encoded by $\mathcal{M}_c$. This misalignment can degrade multilingual performances: tasks that rely on inputs from the less-aligned languages may exhibit lower performance since they draw from representations that loosely intersects with the core subspace.

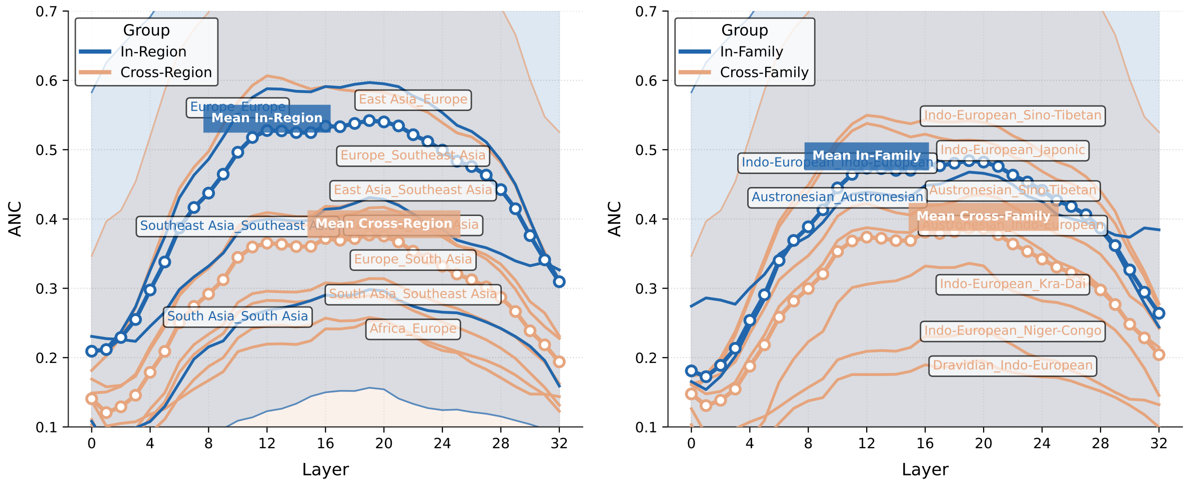
4 Semantic Alignment of Multilingual LLMs Representations

We explore the presence and characteristics of the components $\mathcal{M}_c$ and $\mathcal{M}_{f_i}$ within multilingual LLMs through assessing the semantic alignment between its representations. Initially, we project LLMs' internal hidden-state embeddings into a 2D space to broadly assess the proximities of model latent representations and observe whether parallel input pairs in different languages clusters or overlaps. We then measure the cross-lingual alignment across language representations through neuron activation consistency and w.r.t their resource-level, linguistic features, and geographical region.

We sample 31 diverse language subsets of Flores-200 (Team, 2024) varied on its resource-level, region, and family (Eberhard et al., 2024) (see Tables 1 and A1) as proxies to typological and morphological features (Georgi et al., 2010). Over experiments, we assess several multilingual LLMs: Aya Expanse (8B) (Dang et al., 2024), Llama-3.1 (8B) (Dubey et al., 2024), Gemma-2 (9B) (Team et al., 2024), Qwen-2.5 (7B) (Yang et al., 2024).

4.1 Inherent Regional Clustering with Mid-Layers High-Resource Alignment

We employ t-SNE (Van der Maaten and Hinton, 2008) to project LLMs' hidden-state embeddings into a 2D space and assess the proximities across language clusters. As t-SNE retains local neighborhood structures, overlaps in this 2D space imply closeness in the original high-dimensional space. In scenarios where embeddings are interlingually aligned, their nearest neighbors should comprise of



(a) Highlights on pairs w.r.t their linguistic region (b) Highlights on pairs w.r.t their linguistic family

Figure 4: Comparisons of per-layer ANC scores on Aya Expanse (8B) with highlights on pairs w.r.t their linguistic region and family. Consistently stronger alignments are observed between within-group mean correlations.

multiple languages. We analyze, the early, middle, and late layers, to visualize the changes in hidden states that demonstrate inflection, peak, and decline in correlations observed in Figure 3. We visualize the cross-lingual comparisons on Aya Expanse (8B) in Figures 1 and 2, and others in Appendix D.

The t-SNE visualizations reveal distinct structural patterns across early (layer 0), intermediate (layer 16), and late (layer 32) layers (see Figures 2a, 1, 2b, respectively). In early and late layers, language representations cluster according to resource levels and linguistic features, with minimal overlap. In contrast, the intermediate layer shows interlingual overlaps that transcend familial and regional boundaries, such as English and Russian overlapping with Indonesian, and Chinese with French. While overlaps mainly involve high-resource languages (HRLs), low-resource languages (LRLs) also exhibit overlaps, often due to regional factors. Meanwhile, some representations remain fragmented outside these overlaps, suggesting that the shared space $\mathcal{H}$ holds both the core and fragmented subspaces. We further investigate these interactions in high-dimensional space to understand the impact and properties of the alignment, in order to complement the analysis in t-SNE.

4.2 Cross-lingual Alignments Depend on Resource-level and Linguistic Properties

Measurement. We further quantify the alignment characteristics by measuring neuron activation alignment for semantically identical inputs across different ℓ through *Average Neuron-wise Correlation* (ANC) (Del and Fishel, 2022). The ANC score in a certain LLM layer is defined as:

$$\mathrm{ANC}(\ell, \ell') = \frac{1}{d} \sum_{i \in d} corr\left(f_\ell^i(\mathbf{x}), f_{\ell'}^i(\mathbf{x}')\right),$$

with $f_\ell^i(\mathbf{x})$ as the activation of i-th neuron for language ℓ and $corr$ denotes Pearson correlation between corresponding activations in ℓ and ℓ'. We visualize layer-wise ANCs from Aya Expanse in Figure 3 and 4, and others in Appendix B.

Findings. We find the "first align, then predict" patterns varies across language pairs. Notably, pairs of HRLs demonstrate strong correlations, while pairs involving LRLs exhibit lower scores (see Figure 3). Similarly, a consistent gap persists between within- and cross-group mean correlations, indicating stronger alignment within familial and regional language groups. Detailed analysis in Table A2 illustrates that most correlated pairs among LLMs are similar on their HRLs. Despite differing rankings, instruction-tuned LLMs exhibit similar sets of top language pairs with its pre-trained counterparts. These significant alignment gaps in cross-lingual correlations indicates latent discrepancies between semantically identical representations. This finding hints that interlingual representational limitations in multilingual LLMs may stem from sparse data, typological distance, and the morphological complexity of languages.

5 Intrinsic Interlinguality of LLMs

Building on our framework and the preliminary findings, we measure the interlingual alignment consistencies of the multilingual LLMs representations. We inspect models' high-dimensional encoding of diverse linguistic inputs and analyze the local neighborhoods of the models' hidden states to quantify their intrinsic interlingual alignment.

5.1 Interlingual Local Overlap Score

Given N input samples from set of languages in $\mathcal{L}$, $\{\mathbf{x}_i^\ell\}_{\ell \in \mathcal{L}, i \in N}$, each sample $\mathbf{x}_i^\ell$ is embedded in model space $\mathcal{H}$ via $f_\ell(\mathbf{x})$. Let's denote $\mathcal{N}(\mathbf{x}_i^\ell)$ as the set of k-nearest neighboring languages of $\mathbf{x}_i^\ell$, defined as $\mathcal{N}(\mathbf{x}_i^\ell) = \{\ell' \neq \ell : \mathbf{x}_j^{\ell'} \in \mathrm{NN}_k(\mathbf{x}_i^\ell)\}$.

Bridge. The bridge score B_ℓ determines the degree of local interlingual mixing, analogous to the participation coefficient in graph theory, which assesses a node's link distribution across modules (Guimera and Amaral, 2005; Mijalkov et al., 2017). Bridge score measures the proportion of samples whose k-nearest neighbors in $\mathcal{H}$ include at least τ unique other languages, formally:

$$B_\ell = \frac{1}{N} \sum_{i \in N} \mathbf{1}\big(|\mathcal{N}(\mathbf{x}_i^\ell)| - 1 \geq \tau\big)$$

A score of ≈ 1 indicates that samples from ℓ consistently neighboring with diverse other languages.

Reachability. Inspired by classical degree of centrality in network analysis (Freeman et al., 2002; Borgatti and Everett, 2006), which quantifies a node's connections, we define the reachability score to measure cross-lingual connectivity of language embeddings. We view the multilingual space $\mathcal{H}$ as an undirected graph with embeddings as nodes linked to their k-nearest neighbors. The reachability score R_ℓ quantifies the connectivity degree of the ℓ embeddings within $\mathcal{H}$, defined as:

$$R_\ell = \frac{1}{|\mathcal{L}| - 1} \left| \bigcup_{i \in N} \mathcal{N}(\mathbf{x}_i^\ell) \right|$$

R_ℓ enumerates the fraction of unique languages encountered across all samples of ℓ in $\mathcal{L}$, excluding itself. A high R_ℓ suggests that ℓ embeddings connect extensively within the multilingual space.

Interlingual Local Overlap (ILO). We then define an interlingual local overlap score ILO_ℓ to quantify the holistic interlingual alignment of language ℓ within $\mathcal{H}$, formally:

$$\mathrm{ILO}_\ell = 2 \cdot \frac{B \cdot R}{B + R}$$

Dataset	Usage	# Lang	# Sample
Flores-200	Analysis	31	30,907
GSM8KInstruct	Training	10	73,559
MGSM	Evaluation	11	2,750

Table 2: Dataset statistics. "# Lang" indicates the number of languages represented in the dataset, and "# Sample" signifies the total sample size included.

with the harmonic mean emphasizes the requirement of strong assessments in both the mixing and connectivity for the representations of ℓ to be considered as locally overlapping with other languages. Consequently, aggregated $\overline{\mathrm{ILO}}_\mathcal{L}$ of high ILO_ℓ in,

$$\overline{\mathrm{ILO}}_\mathcal{L} = \frac{1}{|\mathcal{L}|} \sum_{\mathcal{L}} \mathrm{ILO}_\ell \,.$$

signals that multilingual LLMs effectively encode all of the diverse language inputs as aligned interlingual semantics within those in $\mathcal{L}$.

Preserving Interlinguality of LLMs. We demonstrate how ILO illuminate the performance variations in cross-lingual transfer and concurrently underscore the critical role of semantic interlingual alignment in multilingual LLMs. Cross-lingual transfer capitalizes on shared features to enhance multilingual capabilities (Philippy et al., 2023), typically involving single-language fine-tuning on a source language and directly applying it to target languages without further tuning. Despite its success, LLMs can suffer from catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999; Biesialska et al., 2020), where their cross-lingual generalization may degrade (Liu et al., 2021; Winata et al., 2023). Research suggests LLMs align multilingual inputs into language-independent representations, then revert them back to the query's original language (Muller et al., 2021; Zhao et al., 2024). Building on these insights, we conduct an experiment to preserve interlingual alignments by employing a **selective freezing** strategy, where we partially freeze parameters critical to language alignment. Our aim is to assess the potential mitigation of cross-lingual disruption, evaluated through ILO scores.

5.2 Experiment Design

To preserve the aligned semantics within multilingual model space, we freeze the parameters of its first eight layers. Additionally, we keep the token

Method	Training languages	Accuracy											Average	
		ben	tha*	swh	tel*	jpn	zho	deu	fra	rus	spa	eng	All	XL
Pre-trained	mixed	11.6%	12.0%	7.2%	0.0%	10.4%	8.8%	16.0%	12.4%	14.0%	11.6%	17.6%	10.3%	-
Fine-tuning	ben	**23.2%**	4.8%	1.2%	3.2%	10.0%	9.6%	10.8%	13.6%	11.6%	14.8%	12.8%	9.8%	9.6%
	tha*	1.6%	**32.8%**	4.4%	1.6%	14.4%	14.8%	17.2%	19.2%	18.0%	20.4%	25.6%	13.8%	13.3%
	swh	3.2%	6.4%	**30.8%**	2.8%	11.2%	12.4%	20.4%	19.6%	14.8%	22.4%	26.8%	13.5%	13.4%
	jpn	3.6%	7.2%	2.8%	1.2%	**32.8%**	21.6%	19.6%	18.0%	18.4%	22.4%	28.8%	13.9%	14.2%
	zho	0.8%	7.2%	2.4%	1.6%	22.0%	**34.8%**	19.6%	19.6%	21.6%	21.2%	27.6%	14.4%	15.1%
	deu	8.0%	16.4%	8.0%	**4.0%**	19.2%	19.6%	**37.6%**	**34.4%**	23.6%	28.8%	36.4%	**19.0%**	**20.5%**
	fra	4.8%	11.6%	4.0%	3.2%	16.0%	16.8%	31.6%	**34.4%**	25.6%	34.4%	35.6%	16.4%	15.2%
	rus	4.0%	14.0%	4.0%	1.2%	17.2%	16.4%	29.6%	28.4%	**34.0%**	30.0%	26.4%	16.5%	14.1%
	spa	4.8%	16.0%	2.8%	2.4%	14.4%	19.6%	28.4%	30.8%	31.2%	**38.4%**	38.4%	16.7%	13.0%
	eng	6.4%	14.4%	6.0%	2.4%	18.8%	24.4%	37.2%	27.2%	33.6%	33.2%	**43.2%**	18.9%	15.7%
Selective Freezing	ben	**23.2%**	9.2%	8.8%	10.0%	17.6%	11.6%	18.0%	16.4%	17.6%	18.4%	20.8%	14.7%	14.4%
	tha*	14.0%	**35.2%**	12.4%	12.4%	16.4%	20.8%	24.8%	20.8%	16.8%	18.0%	28.0%	19.3%	19.6%
	swh	8.4%	13.6%	**30.0%**	8.4%	15.2%	12.8%	20.8%	19.2%	16.8%	24.8%	29.2%	16.1%	16.1%
	jpn	15.6%	15.2%	12.0%	14.0%	**30.0%**	27.2%	24.8%	22.8%	23.2%	24.0%	28.0%	20.5%	20.9%
	zho	15.6%	21.2%	10.4%	10.4%	22.0%	**40.8%**	23.6%	20.4%	21.6%	25.2%	34.8%	20.7%	22.0%
	deu	18.0%	18.4%	8.4%	16.0%	22.4%	24.0%	34.0%	31.2%	27.6%	32.0%	38.4%	22.2%	22.3%
	fra	**23.2%**	19.2%	13.2%	14.0%	18.8%	20.0%	30.4%	**35.2%**	30.8%	33.2%	37.6%	22.8%	22.2%
	rus	17.2%	18.4%	10.8%	14.4%	15.2%	18.0%	29.6%	24.4%	**38.0%**	29.6%	36.8%	20.7%	18.6%
	spa	17.2%	18.4%	11.6%	14.0%	20.4%	22.8%	31.6%	31.6%	28.8%	38.0%	36.4%	21.8%	19.4%
	eng	18.8%	23.2%	19.6%	**17.6%**	26.4%	29.6%	**36.8%**	32.4%	36.4%	**40.0%**	**42.0%**	**26.8%**	**24.6%**

Table 3: Cross-lingual transfer performance on MGSM for Llama-3.1 (8B) without and with selective freezing. "XL" denotes average on languages that were not fine-tuned. Diagonal entries in blue highlights correspond to source language performances. Red highlights indicate decrease from pre-trained baseline. **Bold** and underline respectively denote the best within group and within column. The (*) marks languages classified as low-resource in Flores-200.

embedding, final layer normalization, and language modeling head (output projection layer) fixed. We identify these parameters as the language aligners.

Datasets. We attend specifically to mathematical reasoning task, as it is inherently language-independent. We utilize the multilingual dataset `GSM8KInstruct` (Chen et al., 2024), which extends the English mathematical reasoning dataset `GSM8K` (Cobbe et al., 2021). This extension involves translating English instructions and chain-of-thought responses into 9 non-English languages via automatic translation and native-speaker human verification. To evaluate the model performance on multilingual mathematical reasoning tasks, we utilize the `MGSM` benchmark dataset (Shi et al., 2022). We attach the complete dataset statistics in Table 2.

Evaluation. We evaluate the accuracy of LLM generated responses under zero-shot conditions using a greedy decoding strategy. Specifically, we employ the evaluation scripts of Zhu et al. and determine answer accuracy by verifying that the final numerical value produced in the LLM's output exactly matches the corresponding ground-truth. To compute the ILO scores, we set $k = 10$ and $\tau = 5$, to define a neighborhood size that is large enough to be informative but small enough to respect the local structures, while requiring each local neighborhood to be rich in interlingual mixing, that at least 5 out of 10 neighbors be from other languages.

Models. We employ two multilingual LLMs: Llama-3.1 (8B) and Gemma-2 (9B). We train both LLMs using the same hyperparameters with learning rate $8e-5$, batch size 8, and gradient accumulation of 16 for 3 epochs using 4 A800 GPUs.

5.3 Results and Analysis

Cross-lingual transfer. We present findings from our cross-lingual transfer experiments, detailed in the Tables 3 and A3 within the "**fine-tuning**" rows, where we evaluated the performance of the fine-tuned Llama-3.1 and Gemma-2 respectively. Consistent with the expectations, we observed substantial cross-lingual transfer signified by improved performance in both source and target languages, even without direct training in those languages. The transfer is notably more pronounced in HRLs and languages within the same families and regions, such as the Indo-European languages in Europe: English, Spanish, Russian, French, and German. Remarkably, in some instances, performances on the target languages paralleled the accuracies in the source language – as exemplified by

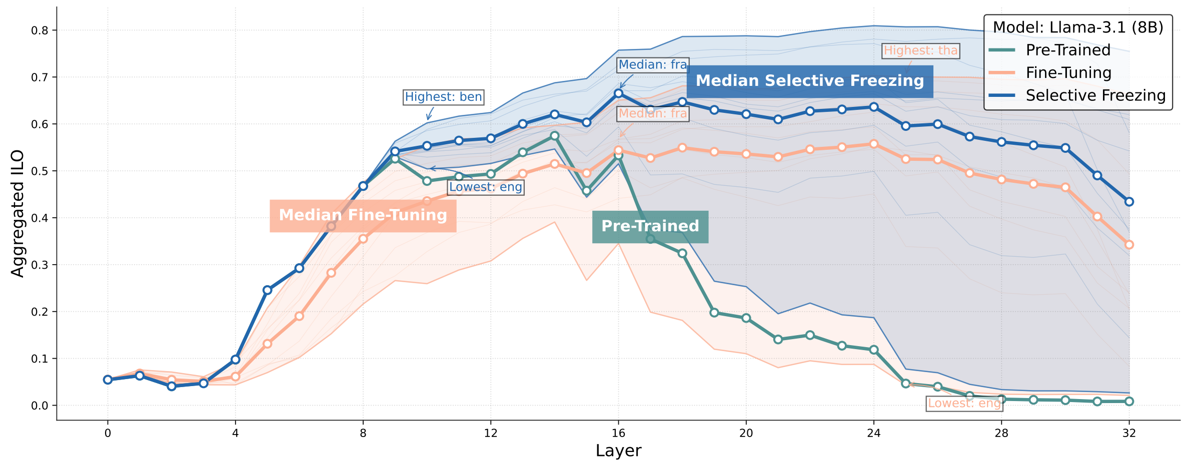


Figure 5: Layer-wise $\overline{\mathrm{ILO}}_{\mathcal{L}}$ scores for all of the source-languages in the single-language training on Llama-3.1 (8B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes. Notable decrease in alignment from single-language training is seen in the early layers on **fine-tuning**, whereas the **selective freezing** mechanism allows the model to sustain its **pre-trained** semantic alignment across layers.

Spanish-to-English achieving 38.4%, which is on par with the Spanish-to-Spanish performance.

Despite the transfer, performance degradations are also observed on some of target languages. We conjecture that this issue stems from disruptions in the functionality of the aligner module. To investigate this hypothesis, we compute per-layer aggregated $\overline{\mathrm{ILO}}_{\mathcal{L}}$ scores, and visualize them in Figures 5 and A13, for all of the source-languages trained on each the Llama-3.1 (8B) and Gemma-2 (9B) models. Both figures show a notable decrease in interlingual semantic alignment post fine-tuning that appears as early as in the 4^{th} layer for Llama and the 6^{th} layer for Gemma. Critically, the degree of alignment does not recover to the height of its pre-trained levels even after additional computational stages in subsequent layers. To address this misalignment, we employ the **selective-freezing** strategy, to preserve the level of interlingual alignment by safeguarding the functionalities of aligning linguistic representations.

Preservation of LLMs' interlinguality. The quantitative analysis through the lens of the aggregated $\overline{\mathrm{ILO}}_{\mathcal{L}}$ reveals that multilingual LLMs trained with **selective-freezing** mechanism sustain their prior semantic alignment levels in the early layers, and across all layers, as demonstrated in Figures 5 and A13. Empirical findings in Tables 3 and A3 further corroborate these insights, highlighting the substantial impact of maintaining interlingual semantic alignments on enhancing multilingual performances. Through keeping the aligner parameters unchanged, both LLMs understudy gain improved cross-lingual generalization compared to their post fine-tuning performances on source languages. Enhanced transfers can be observed on languages within-families and within-regions, with improvements, and nearly no degradation towards the low-resource, cross-family, and cross-regional languages. These findings strongly indicate that preserving the alignment of interlingual representations in LLMs is essential for scalable multilingual learning. They emphasize the critical role of the models' interlingual representation alignments in enhancing the multilingual capabilities of LLMs.

6 Conclusion

The emergence of multilingual LLMs demonstrates that interlingual constructs naturally arise, even in the absence of explicit objectives. We introduce a conceptual framework to understand interlingual representations, identifying both the core interlingual subspace that captures shared semantics, and fragmented components that reveal representational limitations in aligning with this core subspace. To advance understanding of interlingual semantic alignment, we propose the Interlingual Local Overlap (ILO) score which quantifies alignment in the local neighborhood structures of interlingual high-dimensional representations. Our proposed framework and metric illuminates the critical role of semantic alignment, offering a quantitative view into the high-dimensional alignment of multilingual representations. This study emphasizes interlingual semantic alignment and provides critical insights to optimize multilingual LLMs in context of diverse linguistic tasks.

Limitations and Future Works

Expanding the core interlingual subspace. Our works assumes the existence of the core interlingual subspace where semantically aligned representations shared across languages, and others that only partially aligned to this core. Future works could explore on expanding this core interlingual subspace to encompass a broader range of languages, i.e. to introduce learning techniques that explicitly encourage deeper and more diverse interlingual mixing. Incorporating a larger, more heterogeneous multilingual datasets and leveraging linguistic priors might further strengthen the core subspace, and in turn, enhancing the universality of the core interlingual representations.

Bridging fragmented subspaces. A significant limitation of existing multilingual LLMs is that certain languages, particularly the underrepresented or typologically distant ones, most likely form fragmented subspaces rather than being integrated fully with the core cluster. To address this, future work could aim to develop targeted strategies to encourage the integration of these subspaces and to narrow these gaps, i.e. under conditions of extremely limited data. Such interventions could facilitate the alignments of interlingual representation, thereby improving overall inclusivity and richness in linguistic diversity of the multilingual LLMs.

Predicting cross-lingual transfer. Although our work provides valuable insights into the local alignment of multilingual embeddings, it does not predict downstream cross-lingual transfer performance. One key limitation, for example, is that our proposals captures generic interlingual mixing of hidden-states representations and not the alignments of task vectors (Ilharco et al., 2022) that might be integral for effective transfer. This disconnect may arise when models achieve strong interlingual alignment while simultaneously losing critical nuances required for task performance. Future work could explore the integration of our proposals with task-aware signals, to develop quantifiers that are more designed to predict cross-lingual transfer.

Towards pure semantic representations. While our current work focuses solely on textual embeddings, a major frontier for future research lies in extending the framework of quantifying alignment via the local neighborhood structures of high-dimensional representations, to multimodal settings. Considering information from another modalities, it may be beneficial to disentangle and measure pure semantic content from modality-specific biases effectively. Exploring this direction not only hints promises to elucidate and improve modality-transfer but also potentially advance our understanding of how different forms of information interact to shape a universal semantic space. We envision our work, upon many others (e.g. Cahyawijaya et al. (2024a); Engels et al. (2025); Ji et al. (2024); Liu et al. (2024); Grosse et al. (2023)), to foster explorations towards the study of LLMs' semantic space.

Ethical Considerations

The exploration of interlingual representation in multilingual LLMs presents a unique opportunity to foster diversity and inclusivity in the field of NLP. Our work introduces framework and metrics to inspect interlingual representations in multilingual LLMs. They enable the analysis of interlingual alignment of different languages in the naturally emerging interlingual constructs within LLMs. We use publicly available parallel corpora and adhere to best practices in data handling, ensuring that no sensitive or personally identifiable information is involved. While our proposals help reveal disparities in representation, through this work, we instead leverage these insights to drive proactive interventions—ensuring future multilingual LLMs are not only more inclusive but also more reflective of the rich linguistic diversity they aim to serve. We hope our results contributes to more equitable model development and encourages further investigation into mitigating potential representational gaps across underrepresented languages.

Embracing Language Diversity Our work aims to create a universal representation that respects and preserves the unique characteristics of each language. Our findings highlight the importance of consistent interlingual alignments. By recognizing and capturing shared semantic structures through interlingua representations, LLMs can contribute to the preservation of linguistic diversity, ensuring that no single language or language group dominates the representation space. We envision LLMs to effectively represent and understand diverse languages, to be truly inclusive in language technology (e.g. Cahyawijaya (2024)). This is particularly crucial for underrepresented languages and communities, enabling them to have their voices heard and enabling them equal access of information, for example to their language-agnostic applications.

Addressing Bias and Fairness The study's observation of varying alignment consistencies across language groups underscores the need for careful consideration of bias. By identifying and addressing fragmented components due to representational limitations, we can work towards creating fairer representations. This is essential to prevent the reinforcement of existing biases and ensure equitable treatment of all languages. When LLMs effectively bridge the gap between languages, they enable seamless communication and understanding, benefiting diverse communities and fostering a more inclusive digital information systems.

References

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. 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 1184–1197.

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.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541.

Stephen P Borgatti and Martin G Everett. 2006. A graph-theoretic perspective on centrality. *Social networks*, 28(4):466–484.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

Samuel Cahyawijaya. 2024. *Llm for everyone: Representing the underrepresented in large language models*. Ph.D. thesis, Hong Kong University of Science and Technology (Hong Kong).

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024a. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024b. 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, Mexico City, Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. Instructalign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

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.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are linear. In *The Thirteenth International Conference on Learning Representations*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Linton C Freeman et al. 2002. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Pascale Fung and Benfeng Chen. 2004. Biframenet: bilingual frame semantics resource construction by cross-lingual induction. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 931–937.

Pascale Fung and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Pascale Fung and Kathleen Mckeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 385–393.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Roger Guimera and Luís A Nunes Amaral. 2005. Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(02):P02001.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kyoko Kanzaki, Yukie Nakao, Manny Rayner, Marianne Santaholma, Marianne Starlander, and Nikos Tsourakis. 2008. Many-to-many multilingual medical speech translation on a PDA. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Government and Commercial Uses of MT*, Waikiki, USA. Association for Machine Translation in the Americas.

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

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024. On the universal truthfulness hyperplane inside llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18199–18224.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71.

Adam Lopez. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3).

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. *arXiv preprint arXiv:1804.08198*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Mite Mijalkov, Ehsan Kakaei, Joana B Pereira, Eric Westman, Giovanni Volpe, and Alzheimer's Disease Neuroimaging Initiative. 2017. Braph: a graph theory software for the analysis of brain connectivity. *PloS one*, 12(8):e0178798.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

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 (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

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

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2):425–486.

Manny Rayner. 2000. *The spoken language translator*. Cambridge University Press.

Manny Rayner, Pierrette Bouillon, Beth Ann Hockey, and Yukie Nakao. 2008. Almost flat functional semantics for speech translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 713–720, Manchester, UK. Coling 2008 Organizing Committee.

Manny Rayner, Paula Estrella, and Pierrette Bouillon. 2010a. A bootstrapped interlingua-based smt architecture. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Manny Rayner, Paula Estrella, and Pierrette Bouillon. 2010b. A bootstrapped interlingua-based SMT architecture. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

R. H. Richens. 1958. Interlingual machine translation. *The Computer Journal*, 1(3):144–147.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Klaus Schubert. 1989. Interlinguistics–its aims, its achievements, and its place in language science. *Interlinguistics: Aspects of the Science of Planned Languages. Trends in Linguistics*, 42:7–44.

Stephanie Seneff. 2006. Combining interlingua with SMT. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Panel on hybrid machine translation: why and how?*, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, pages 1–6.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Bernard Vauquois. 1968. A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress*.

Wolfgang Wahlster. 2013. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

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

Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2023. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 768–777.

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*.

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.

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

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

Appendix

A Detail on Linguistic Properties

We provide the detail of the region and linguistic properties of the language subsets sampled from Flores-200 in A1. Here, while most of them are extracted from Team (2024), we refer to (Eberhard et al., 2024) for the details on linguistic families.

B Further Details on ANC Scores

Here we provide a detailed view on the ANC comparison of the language pairs for all the model understudy. We compute aggregate peak score for each language pair as the mean over the peak layers. We define the peak layer by computing the 75^{th} percentile of ANCs for each layer and select the top 3 layers as the peak layers. We denote all the top correlated language pairs from the layers with peak ANC scores and the unique languages from the top language pairs in Table A2. We find that the top correlated pairs with high ANCs among the LLMs are similar on their HRLs. Instruction-tuned LLMs exhibit similar sets of top language pairs with its pre-trained counterparts, despite the differing rankings of them.

C ANC Comparisons from Other LLMs

We attach the complete visualization on ANC scores derived from the hidden-states of Aya Expanse (8B), Llama-3.1 (8B), Llama-3.1-Instruct (8B), Gemma-2 (9B), Gemma-2-Instruct (9B), and Qwen (9B), respectively in Figures A1, A2, A3, A4, A5, and A6.

D T-SNE Visualizations from Other LLMs

We attach the complete t-SNE visualization projected from the hidden-states of Aya Expanse (8B), Qwen (9B), Llama-3.1 (8B), Llama-3.1-Instruct (8B), Gemma-2 (9B), and Gemma-2-Instruct (9B), respectively in Figures A7, A8, A9 A10, A11, and A12.

E Reports on Cross-Lingual Transfer Experiments for Gemma-2 (9B)

We attach the cross-lingual transfer performance on MGSM and the layer-wise $\overline{\mathrm{ILO}}_{\mathcal{L}}$ scores, for Gemma-2 (9B) in its pre-trained, fine-tuning, and selective-freezing modes, in Table A3 and Figure A13 .

F Observation of Interlingual Alignment Preservation in T-SNE Projections

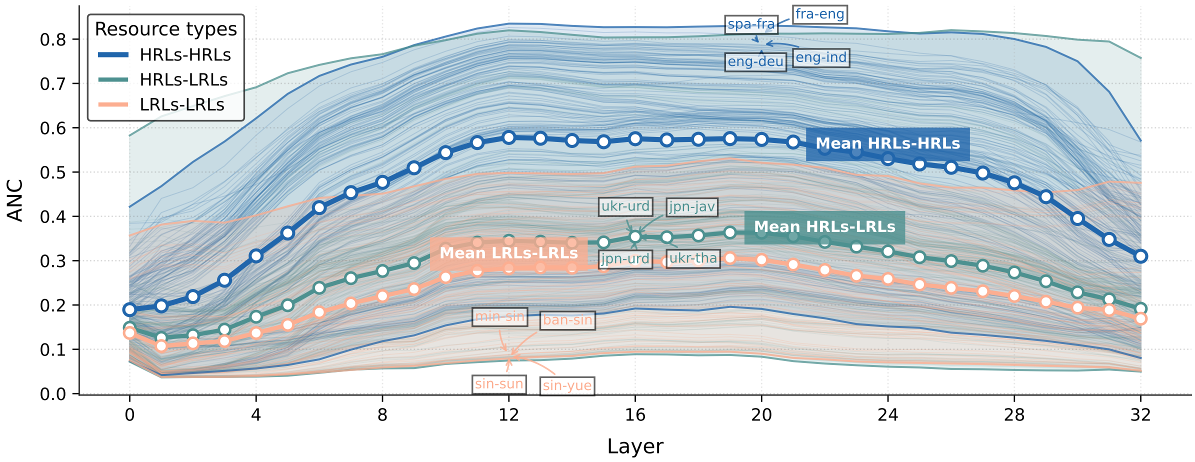
Through our single-language training experiments in the multilingual mathematical reasoning task, we observe that the visual projections using t-SNE, also support that ILO score effectively captures the same interlingual alignment phenomenon, albeit in a projected lower-dimensional dimensions. In other words, layers with high ILO scores consistently exhibits interlingual overlaps in the t-SNE dimensions that hints at strong interlingual alignment, whereas those with lower scores tend to be more fragmented. This correspondence validates ILO as a robust quantitative measure that reflects the local structure of the multilingual shared embedding space. We attach the complete t-SNE visualization projected from the hidden-states of the models underwent single-language training on English in the **fine-tuning** vs **selective freezing** modes of Llama-3.1 (8B) and Gemma-2 (9B), respectively in Figures A14 vs A15, and A16 vs A17.

Code	Language	Script	Region	Family	Res.
ban_Latn	Balinese	Latin	Southeast Asia	Austronesian	Low
ben_Beng	Bengali	Bengali	South Asia	Indo-European	High
bjn_Latn	Banjar	Latin	Southeast Asia	Austronesian	Low
ces_Latn	Czech	Latin	Europe	Indo-European	High
dan_Latn	Danish	Latin	Europe	Indo-European	High
deu_Latn	German	Latin	Europe	Indo-European	High
eng_Latn	English	Latin	Europe	Indo-European	High
fra_Latn	French	Latin	Europe	Indo-European	High
gle_Latn	Irish	Latin	Europe	Indo-European	Low
hin_Deva	Hindi	Devanagari	South Asia	Indo-European	High
ind_Latn	Indonesian	Latin	Southeast Asia	Austronesian	High
jav_Latn	Javanese	Latin	Southeast Asia	Austronesian	Low
jpn_Jpan	Japanese	Japanese	East Asia	Japonic	High
min_Latn	Minangkabau	Latin	Southeast Asia	Austronesian	Low
nld_Latn	Dutch	Latin	Europe	Indo-European	High
pol_Latn	Polish	Latin	Europe	Indo-European	High
rus_Cyrl	Russian	Cyrillic	Europe	Indo-European	High
sin_Sinh	Sinhala	Sinhala	South Asia	Indo-European	Low
slv_Latn	Slovenian	Latin	Europe	Indo-European	High
spa_Latn	Spanish	Latin	Europe	Indo-European	High
srp_Cyrl	Serbian	Cyrillic	Europe	Indo-European	Low
sun_Latn	Sundanese	Latin	Southeast Asia	Austronesian	Low
swe_Latn	Swedish	Latin	Europe	Indo-European	High
swh_Latn	Swahili	Latin	Africa	Niger-Congo	High
tel_Telu	Telugu	Telugu	South Asia	Dravidian	Low
tgl_Latn	Tagalog	Latin	Southeast Asia	Austronesian	Low
tha_Thai	Thai	Thai	Southeast Asia	Kra-Dai	Low
ukr_Cyrl	Ukrainian	Cyrillic	Europe	Indo-European	High
urd_Arab	Urdu	Arabic	South Asia	Indo-European	Low
yue_Hant	Yue Chinese	Han (Traditional)	East Asia	Sino-Tibetan	Low
zho_Hans	Chinese (Simplified)	Han (Simplified)	East Asia	Sino-Tibetan	High

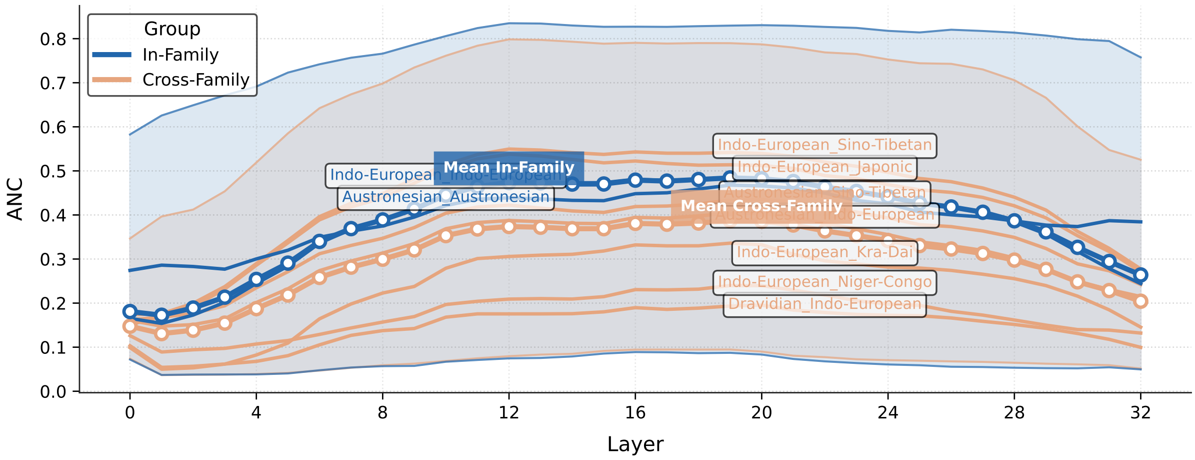
Table A1: Complete distribution of the 31 languages across families, regions, and resource-levels in our analysis, sampled from Flores-200

Models	Gemma-2 (9B)	Gemma-2 It (8B)	Aya Expanse (8B)	Llama-3.1 (8B)	Llama-3.1 It (8B)	Qwen-2.5 (7B)
Top language pairs	dan_Latn - swe_Latn	eng_Latn - fra_Latn	rus_Cyrl - ukr_Cyrl	yue_Hant - zho_Hans	yue_Hant - zho_Hans	yue_Hant - zho_Hans
	eng_Latn - fra_Latn	dan_Latn - swe_Latn	eng_Latn - fra_Latn	rus_Cyrl - ukr_Cyrl	rus_Cyrl - ukr_Cyrl	dan_Latn - swe_Latn
	rus_Cyrl - ukr_Cyrl	rus_Cyrl - ukr_Cyrl	yue_Hant - zho_Hans	dan_Latn - swe_Latn	dan_Latn - swe_Latn	rus_Cyrl - ukr_Cyrl
	yue_Hant - zho_Hans	deu_Latn - eng_Latn	eng_Latn - ind_Latn	eng_Latn - fra_Latn	eng_Latn - fra_Latn	fra_Latn - spa_Latn
	dan_Latn - eng_Latn	yue_Hant - zho_Hans	fra_Latn - spa_Latn	fra_Latn - spa_Latn	fra_Latn - spa_Latn	eng_Latn - fra_Latn
	eng_Latn - swe_Latn	eng_Latn - swe_Latn	deu_Latn - eng_Latn	deu_Latn - swe_Latn	deu_Latn - swe_Latn	fra_Latn - rus_Cyrl
	deu_Latn - eng_Latn	dan_Latn - eng_Latn	ces_Latn - rus_Cyrl	deu_Latn - fra_Latn	deu_Latn - fra_Latn	rus_Cyrl - spa_Latn
	deu_Latn - swe_Latn	deu_Latn - fra_Latn	ces_Latn - ukr_Cyrl	deu_Latn - eng_Latn	deu_Latn - eng_Latn	deu_Latn - fra_Latn
	deu_Latn - fra_Latn	deu_Latn - swe_Latn	deu_Latn - fra_Latn	deu_Latn - nld_Latn	eng_Latn - swe_Latn	ces_Latn - pol_Latn
	dan_Latn - deu_Latn	dan_Latn - deu_Latn	fra_Latn - ind_Latn	ces_Latn - rus_Cyrl	eng_Latn - spa_Latn	deu_Latn - nld_Latn
Unique languages	swe, dan, fra, eng, ukr, rus, zho, yue, deu, spa	fra, eng, swe, dan, rus, ukr, deu, zho, yue, spa	rus, ukr, fra, eng, zho, yue, ind, spa, deu, ces	yue, zho, ukr, rus, swe, dan, fra, eng, spa, deu	zho, yue, rus, ukr, dan, swe, fra, eng, spa, deu	yue, zho, dan, swe, ukr, rus, spa, fra, eng, deu

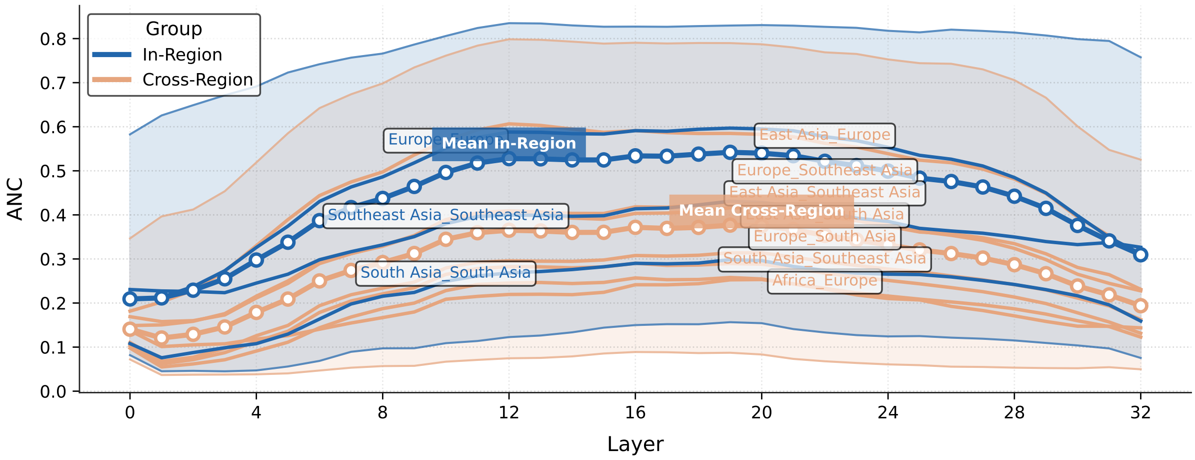
Table A2: Top correlated language pairs from the layers with peak ANC scores and the unique languages from the top language pairs. Most correlated pairs among LLMs are similar on their HRLs. Despite differing rankings, instruction-tuned LLMs exhibit similar sets of top language pairs with its pre-trained counterparts.



(a) Highlights on pairs w.r.t their resource levels

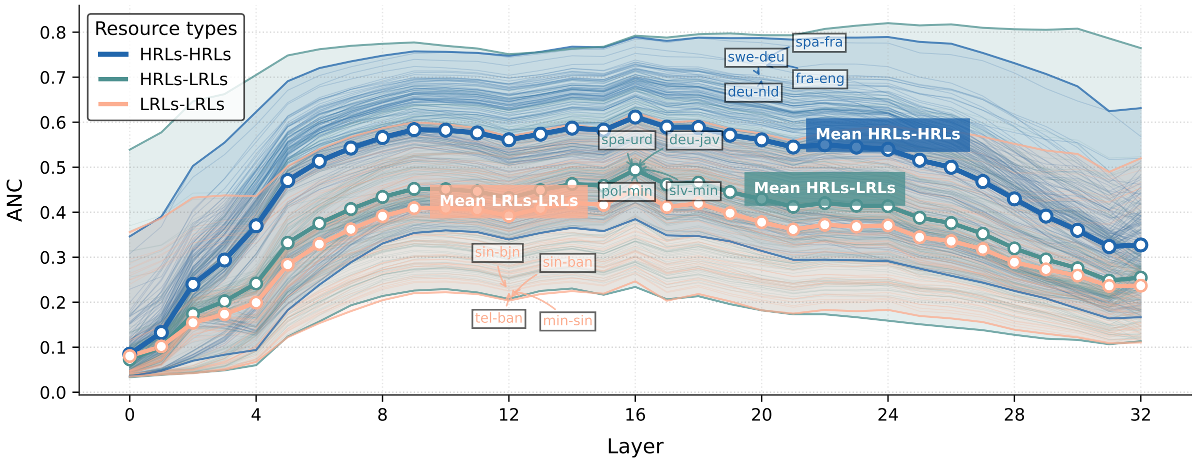


(b) Highlights on pairs w.r.t their linguistic region

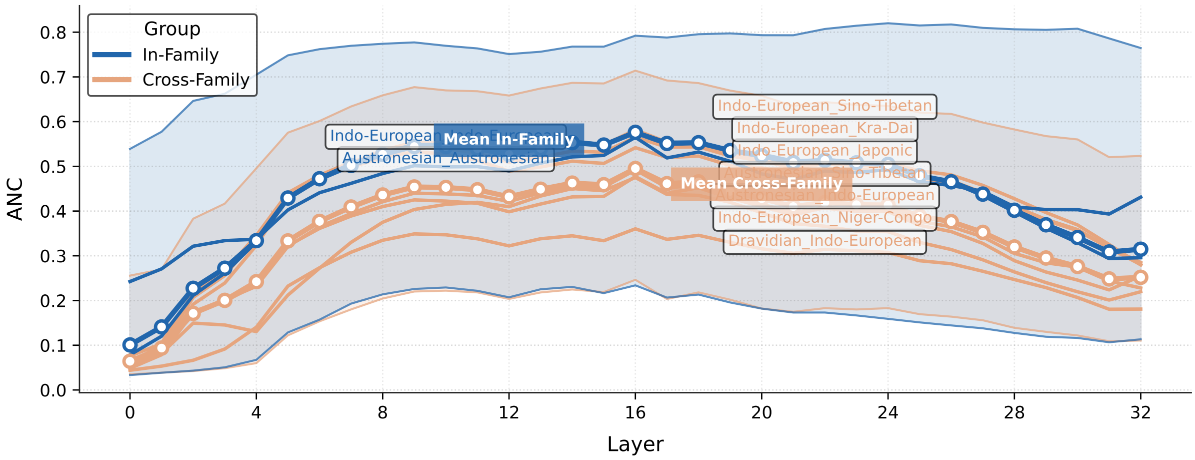


(c) Highlights on pairs w.r.t their linguistic family

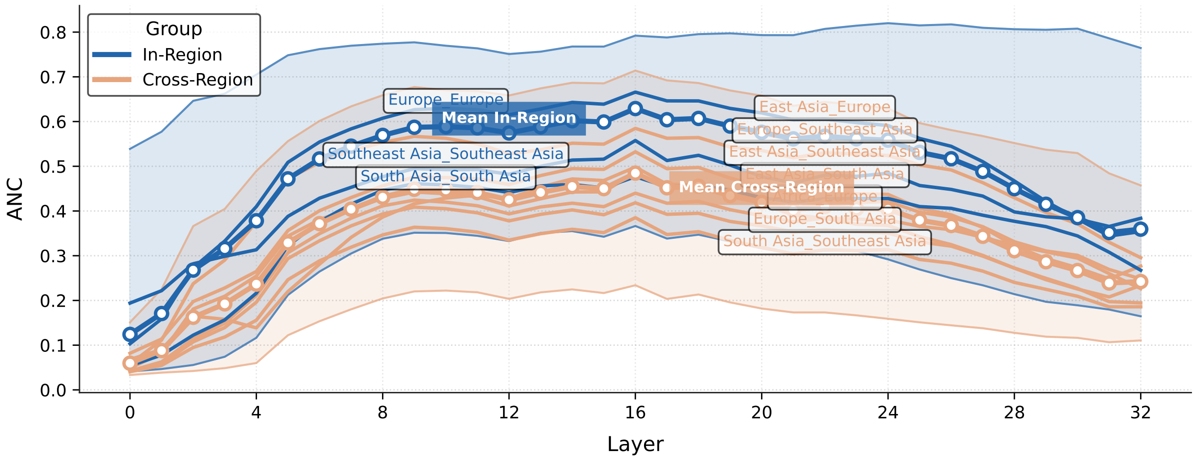
Figure A1: Comparisons of per-layer ANC scores on Aya Expanse (8B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

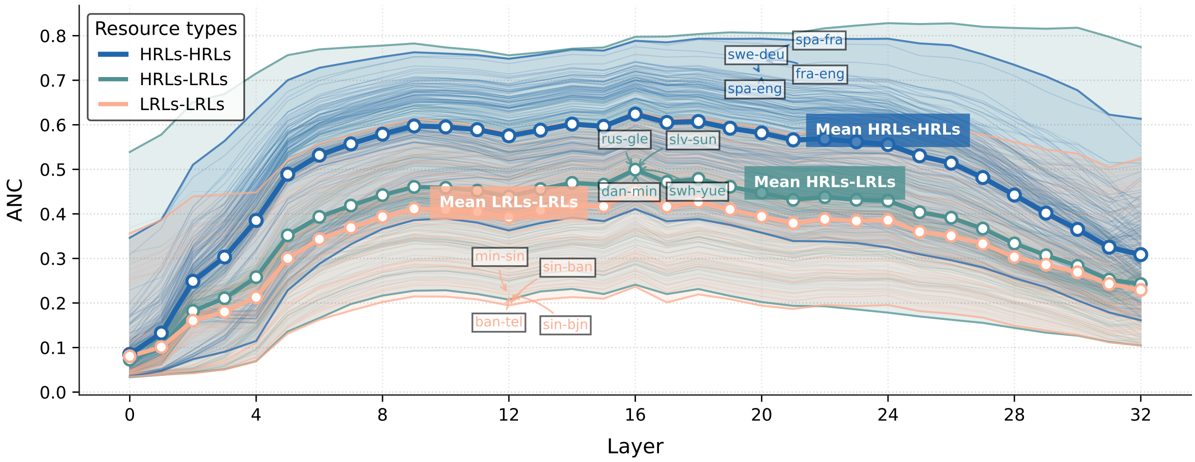


(b) Highlights on pairs w.r.t their linguistic region

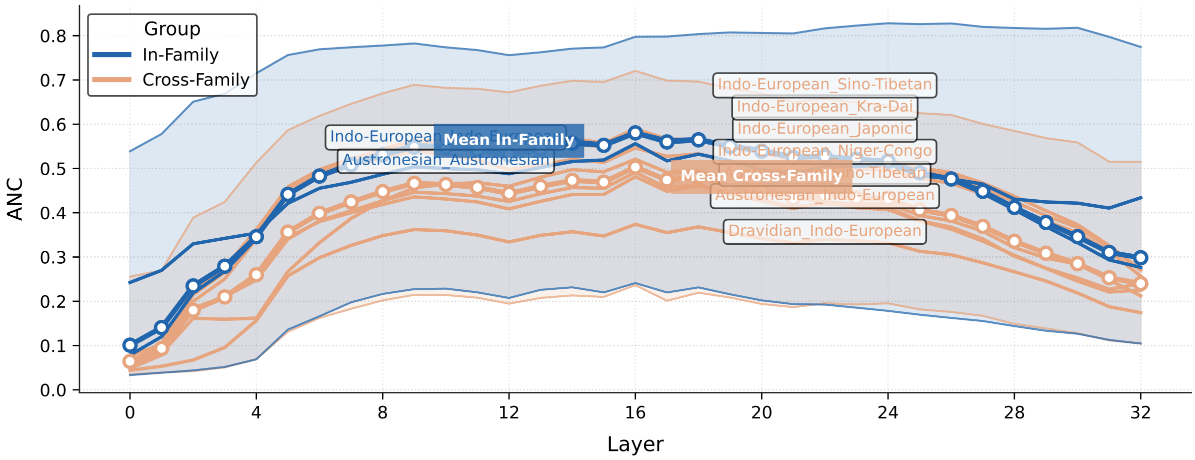


(c) Highlights on pairs w.r.t their linguistic family

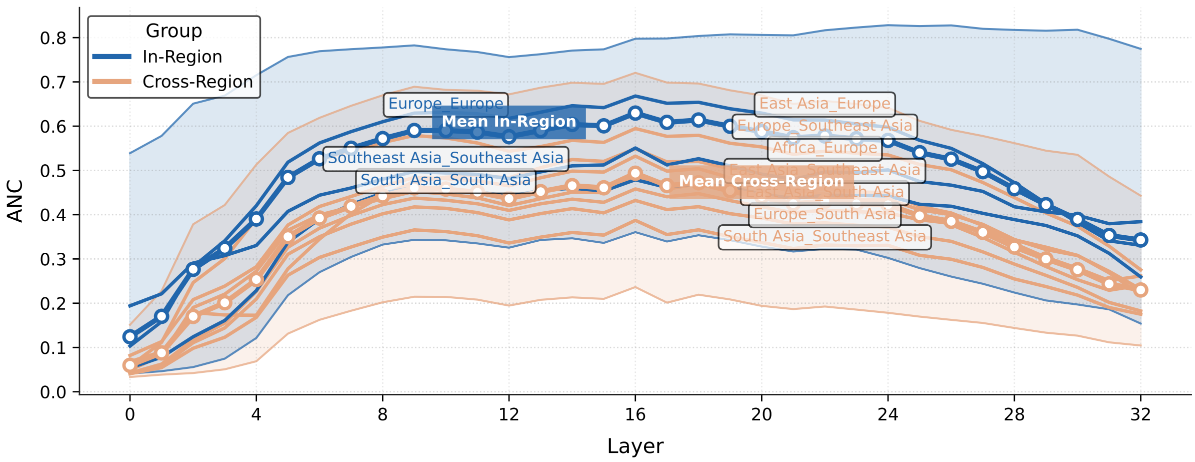
Figure A2: Comparisons of per-layer ANC scores on Llama-3.1 (8B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

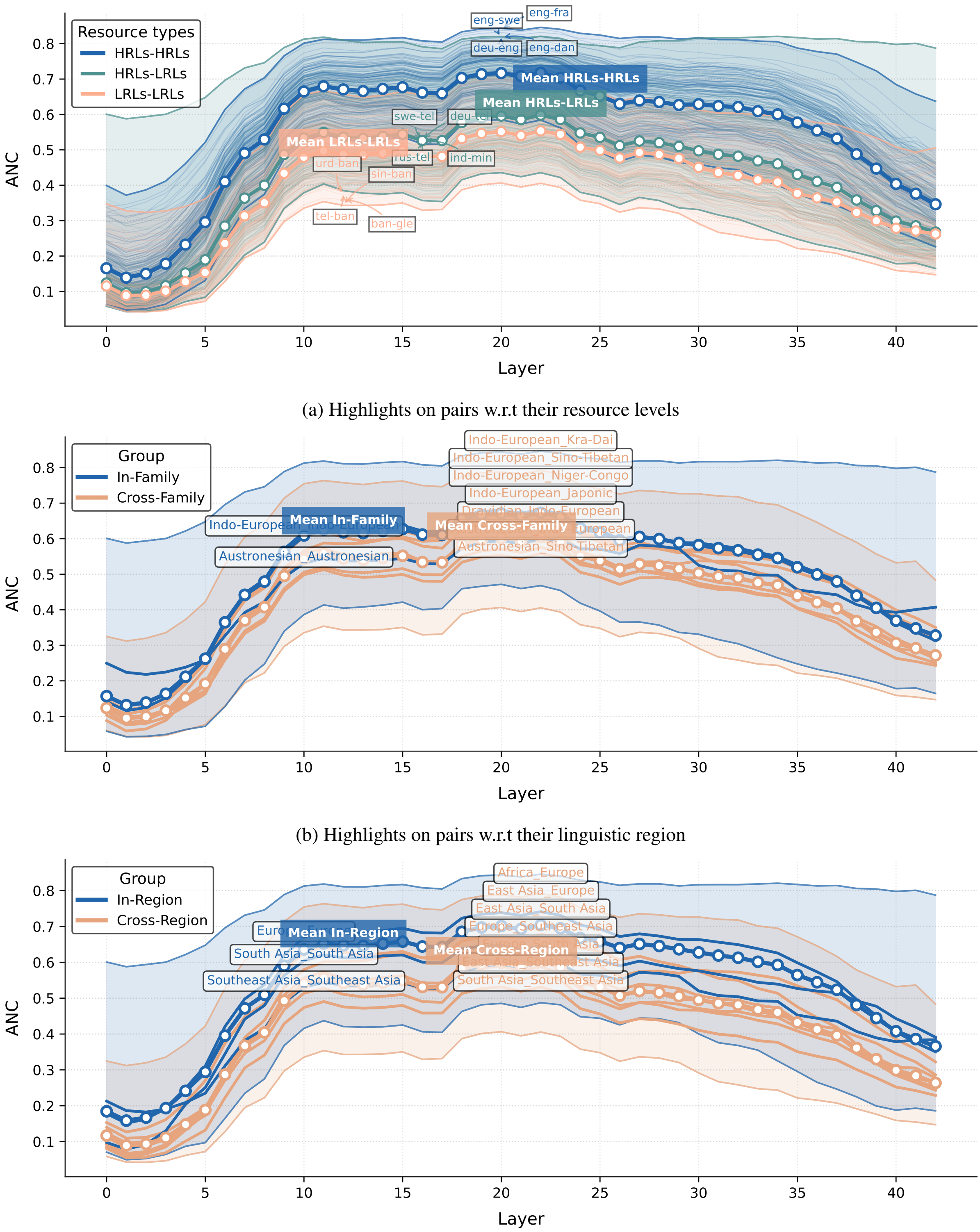


(b) Highlights on pairs w.r.t their linguistic region



(c) Highlights on pairs w.r.t their linguistic family

Figure A3: Comparisons of per-layer ANC scores on Llama-3.1-Instruct (8B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.

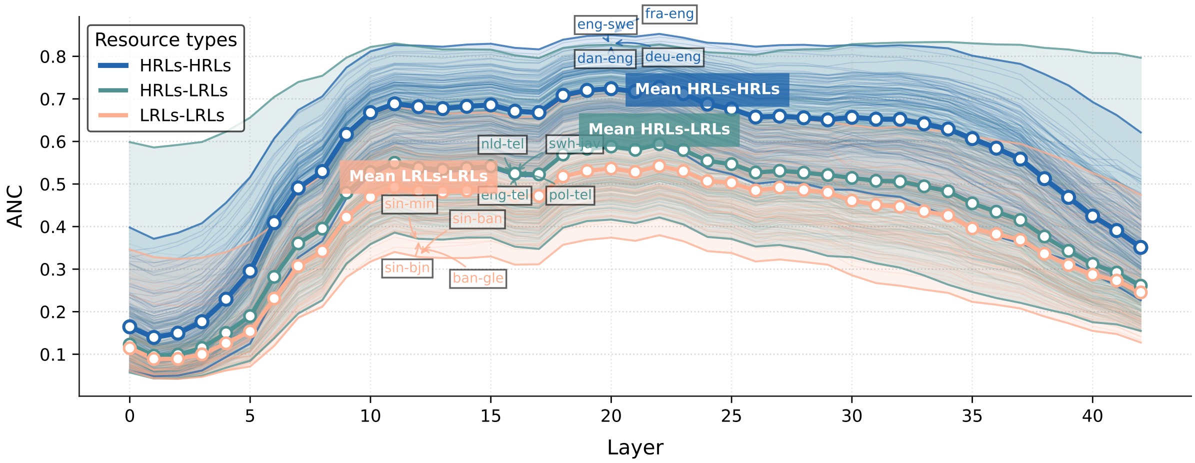


(a) Highlights on pairs w.r.t their resource levels

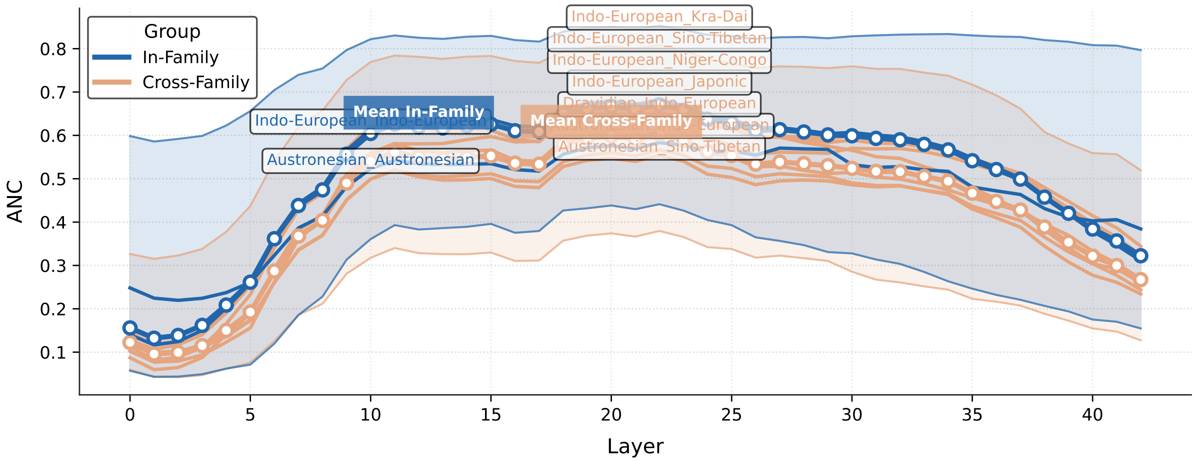
(b) Highlights on pairs w.r.t their linguistic region

(c) Highlights on pairs w.r.t their linguistic family

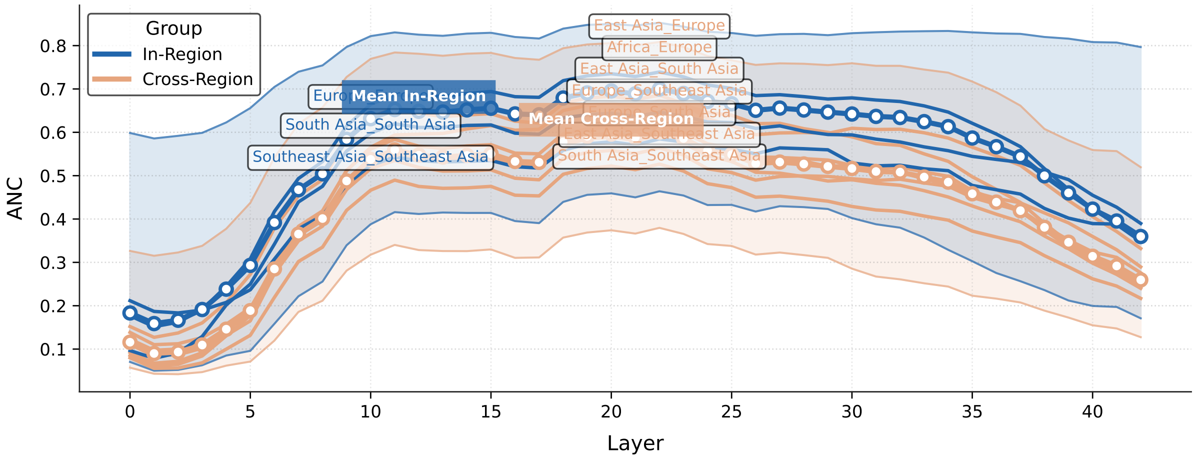
Figure A4: Comparisons of per-layer ANC scores on Gemma-2 (9B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

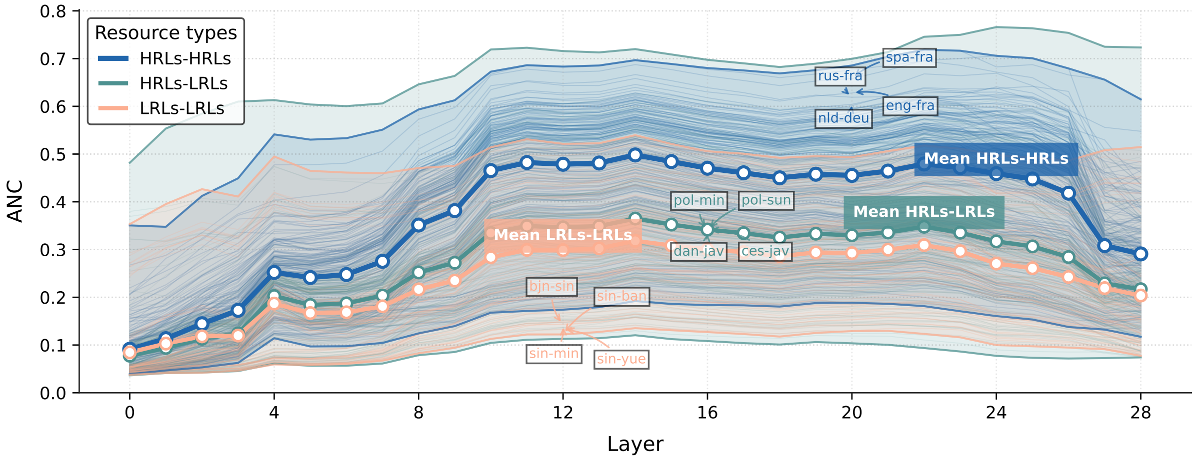


(b) Highlights on pairs w.r.t their linguistic region

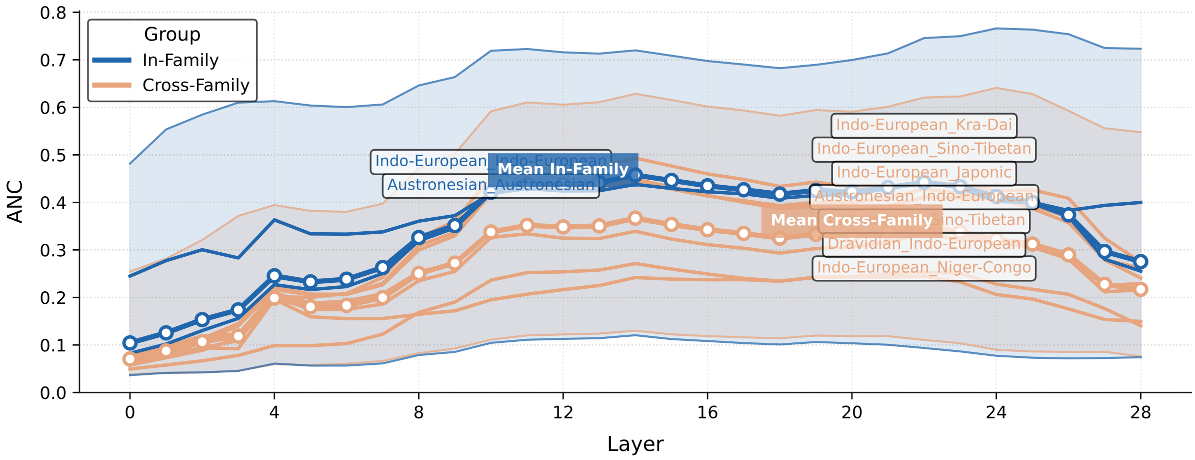


(c) Highlights on pairs w.r.t their linguistic family

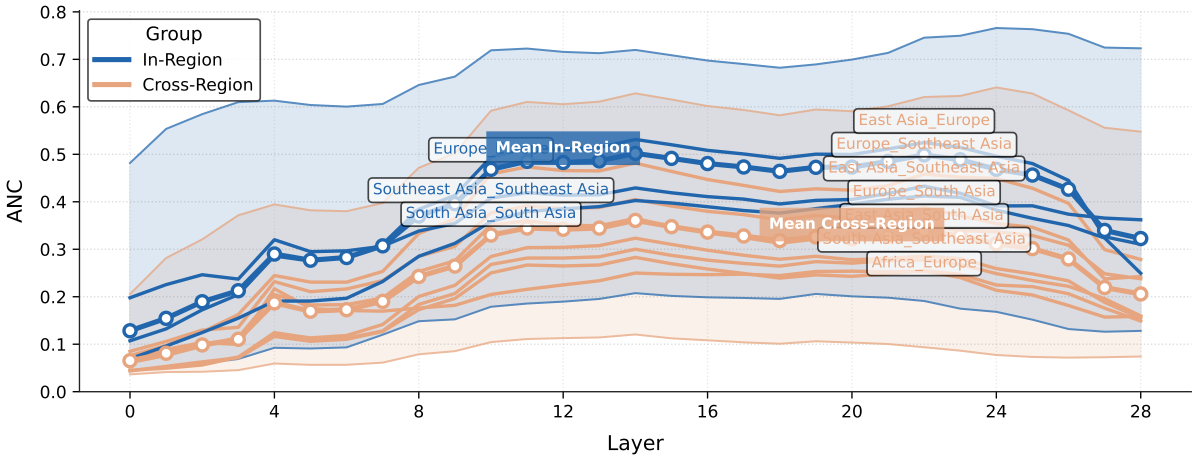
Figure A5: Comparisons of per-layer ANC scores on Gemma-2-Instruct (9B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels



(b) Highlights on pairs w.r.t their linguistic region

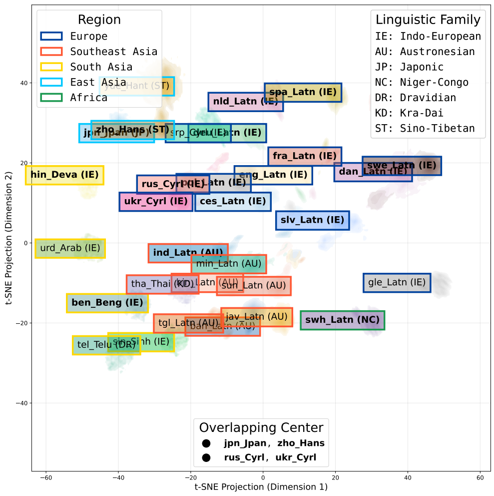


(c) Highlights on pairs w.r.t their linguistic family

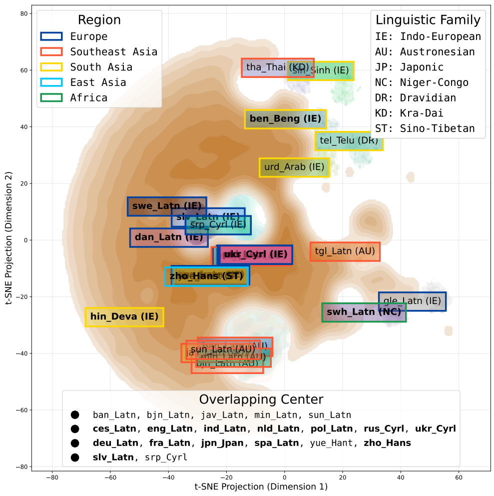
Figure A6: Comparisons of per-layer ANC scores on Qwen-2.5 (7B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.

Method	Training languages	Accuracy											Average	
		ben	tha*	swh	tel*	jpn	zho	deu	fra	rus	spa	eng	All	XL
Pre-trained	mixed	13.2%	12.0%	9.2%	16.0%	10.0%	17.6%	16.8%	16.8%	10.8%	15.2%	17.6%	13.6%	-
Fine-tuning	ben	**27.6%**	4.4%	2.0%	4.4%	11.6%	12.8%	6.8%	10.4%	10.0%	14.4%	18.4%	10.0%	10.0%
	tha*	5.6%	**32.4%**	6.0%	2.8%	10.4%	14.4%	14.8%	16.8%	12.0%	20.0%	26.0%	12.8%	12.9%
	swh	5.6%	5.6%	**32.4%**	0.8%	10.4%	9.6%	15.6%	14.8%	10.8%	21.2%	26.4%	11.7%	11.9%
	jpn	2.4%	6.0%	2.8%	2.4%	**26.8%**	19.6%	13.2%	10.8%	14.4%	18.0%	26.0%	10.9%	11.1%
	zho	2.0%	6.4%	1.6%	0.8%	16.8%	**32.0%**	17.6%	10.4%	16.4%	18.0%	28.0%	11.6%	12.3%
	deu	4.4%	9.2%	5.2%	**6.8%**	16.0%	18.4%	**32.8%**	23.6%	23.2%	26.4%	34.4%	15.5%	**15.7%**
	fra	5.6%	10.8%	6.0%	0.8%	17.6%	18.8%	29.2%	**30.8%**	21.6%	29.6%	31.6%	15.7%	14.6%
	rus	4.8%	4.8%	5.2%	1.2%	13.2%	16.8%	30.0%	24.4%	**32.8%**	29.2%	29.2%	14.8%	11.9%
	spa	7.2%	7.6%	4.8%	4.4%	17.6%	22.0%	26.8%	27.6%	28.4%	**33.2%**	37.6%	16.3%	13.0%
	eng	8.0%	10.4%	8.0%	6.0%	17.6%	20.8%	28.0%	24.4%	25.2%	29.6%	**39.2%**	**16.5%**	14.1%
Selective Freezing	ben	**34.0%**	13.2%	12.8%	**20.8%**	16.4%	16.8%	21.2%	19.2%	16.8%	20.4%	19.2%	19.0%	19.3%
	tha*	14.8%	**33.6%**	17.2%	16.4%	16.4%	19.6%	18.0%	20.8%	21.2%	19.2%	23.6%	19.8%	19.6%
	swh	12.4%	14.4%	**28.8%**	12.4%	12.0%	16.0%	22.0%	26.4%	18.0%	26.4%	34.8%	18.0%	18.1%
	jpn	11.2%	15.2%	11.6%	9.6%	**27.2%**	25.6%	16.8%	18.8%	19.2%	18.8%	24.0%	17.2%	17.3%
	zho	13.2%	19.2%	15.2%	8.0%	22.8%	**36.0%**	20.0%	24.0%	18.4%	23.2%	32.0%	19.6%	20.5%
	deu	13.6%	14.4%	19.6%	14.4%	22.0%	19.6%	31.2%	28.8%	26.4%	31.2%	35.6%	21.1%	21.3%
	fra	20.0%	21.6%	25.2%	13.6%	20.4%	24.0%	29.6%	30.0%	25.6%	32.8%	35.2%	23.3%	23.4%
	rus	14.4%	17.6%	14.0%	9.6%	16.4%	20.8%	24.8%	24.0%	**34.4%**	27.2%	30.0%	19.6%	17.3%
	spa	14.4%	20.4%	22.0%	17.6%	19.6%	21.6%	28.0%	28.0%	28.0%	**40.4%**	35.6%	22.2%	20.5%
	eng	22.8%	20.4%	20.0%	17.2%	26.8%	28.4%	**32.0%**	**35.6%**	**34.4%**	35.2%	**44.0%**	**26.4%**	**23.9%**

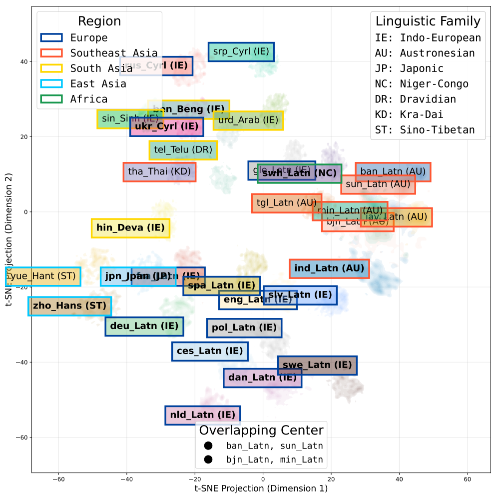
Table A3: Cross-lingual transfer performance on MGSM for Gemma-2 (9B) w/ and w/o selective freezing. "XL" denotes average on languages that were not fine-tuned. Diagonal entries in blue highlights correspond to source language performances. Red highlights indicate decrease from pre-trained baseline. **Bold** and underline respectively denote the best within group and within column. The (*) marks languages classified as low-resource in Flores-200.



(a) Early (layer 0)

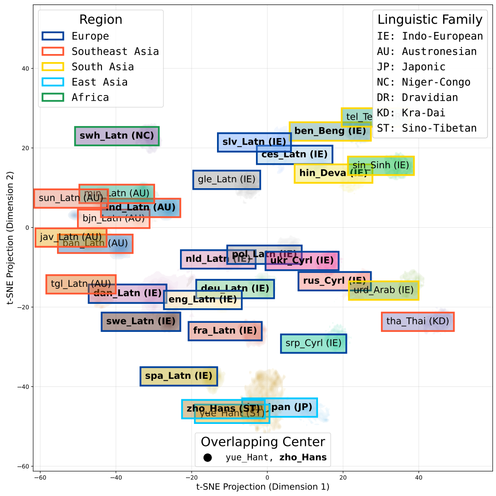


(b) Intermediate (layer 16)

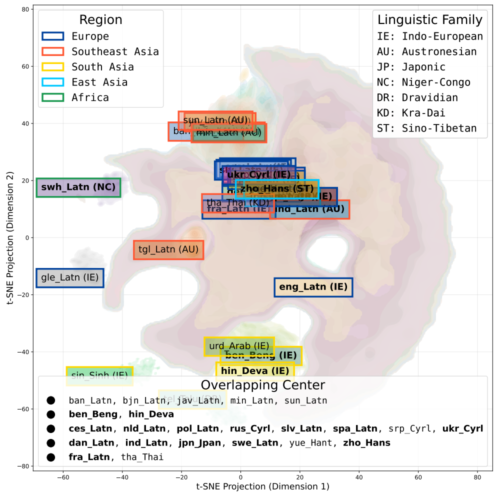


(c) Late (layer 32)

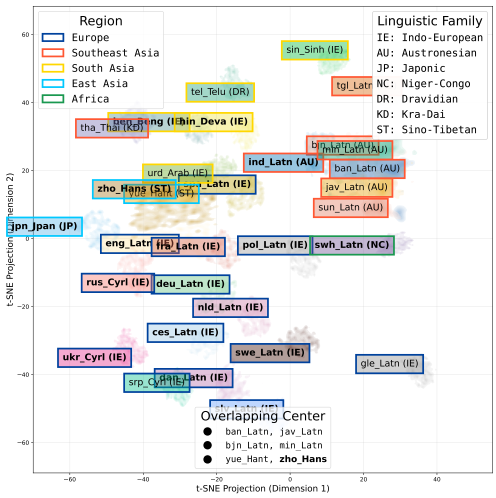
Figure A7: Embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in this hidden states of the intermediate layer of multilingual LLMs. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.



(a) Early (layer 0)

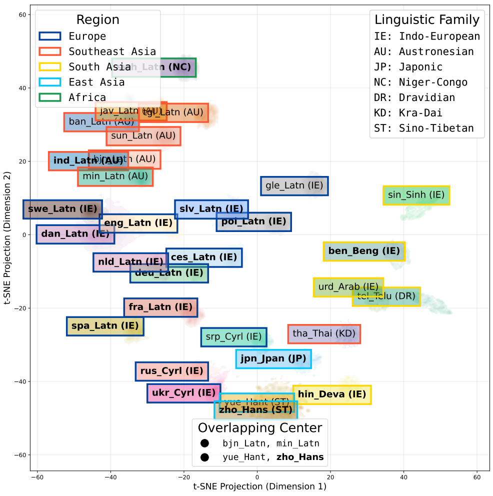


(b) Intermediate (layer 14)

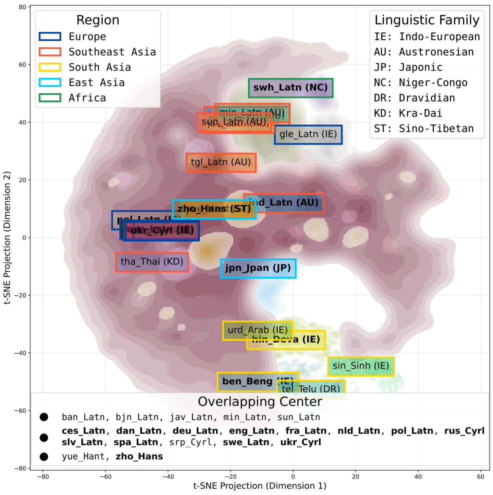


(c) Late (layer 28)

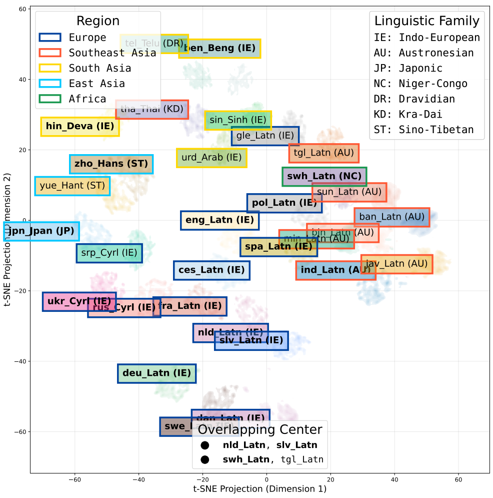
Figure A8: Embeddings of Qwen-2.5 (7B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in this hidden states of the intermediate layer of multilingual LLMs. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.



(a) Early (layer 0)

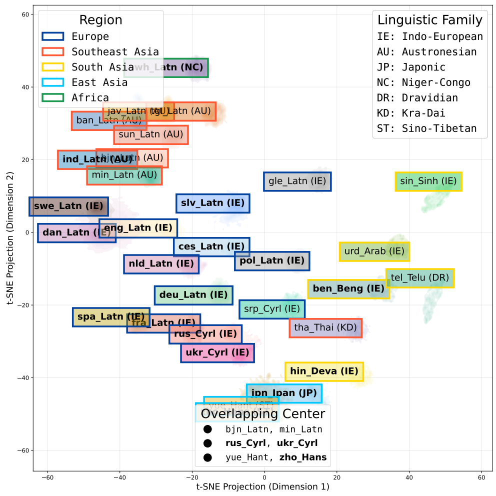


(b) Intermediate (layer 16)

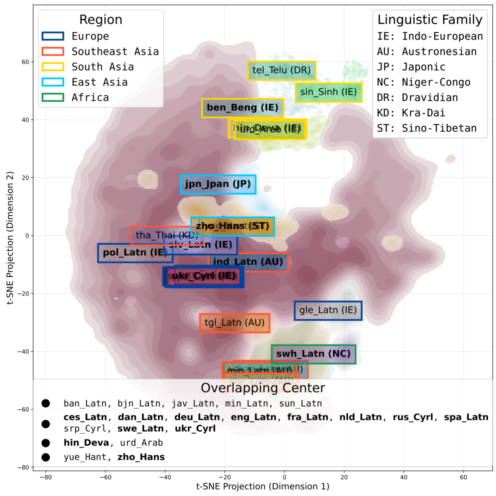


(c) Late (layer 32)

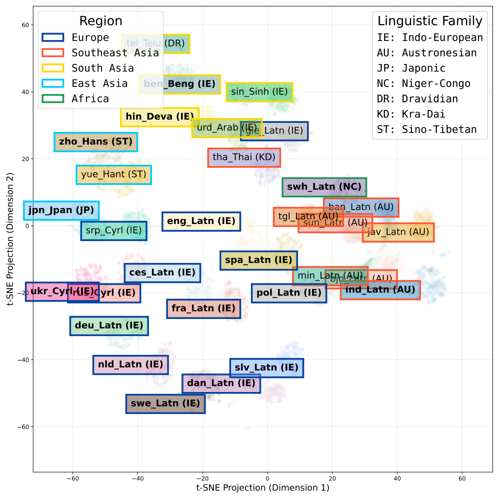
Figure A9: Embeddings of Llama-3.1 (8B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in this hidden states of the intermediate layer of multilingual LLMs. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.



(a) Early (layer 0)

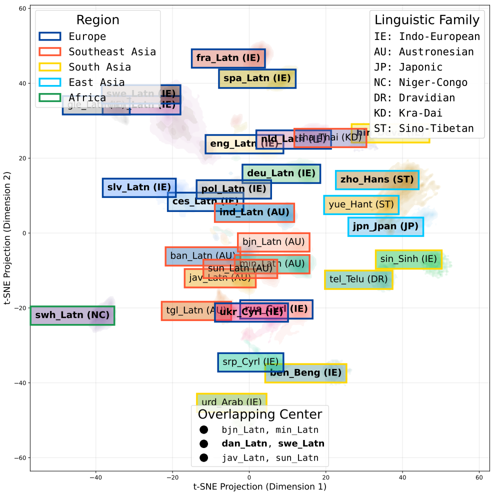


(b) Intermediate (layer 16)

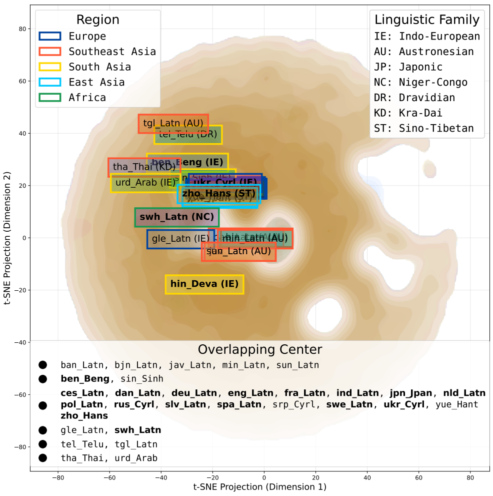


(c) Late (layer 32)

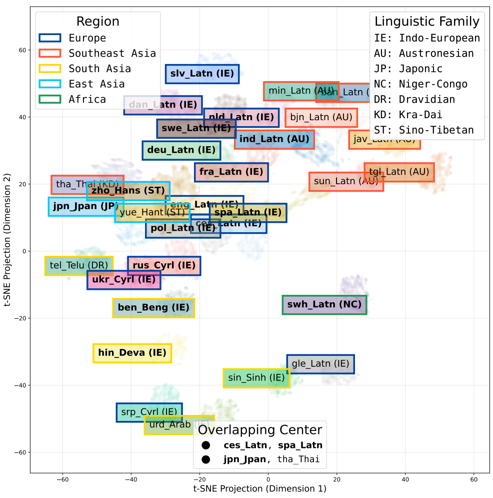
Figure A10: Embeddings of Llama-3.1-Instruct (8B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in this hidden states of the intermediate layer of multilingual LLMs. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.



(a) Early (layer 0)

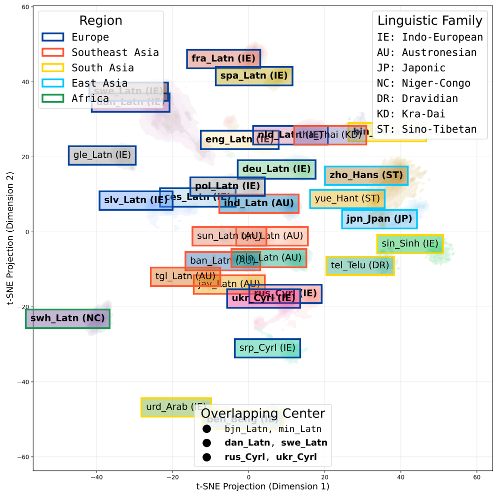


(b) Intermediate (layer 21)

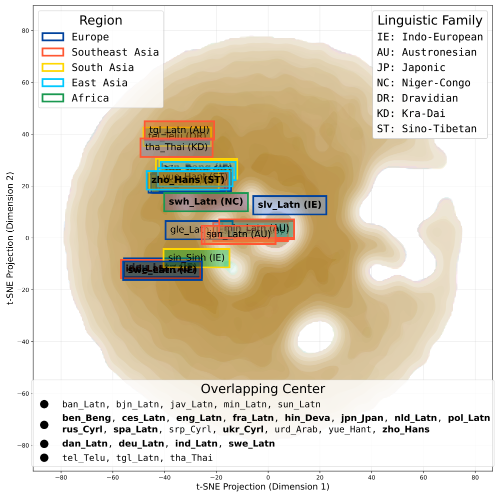


(c) Late (layer 42)

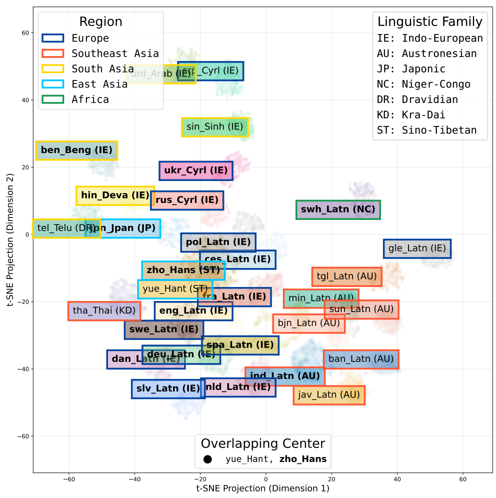
Figure A11: Embeddings of Gemma-2 (9B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in this hidden states of the intermediate layer of multilingual LLMs. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.



(a) Early (layer 0)



(b) Intermediate (layer 21)



(c) Late (layer 42)

Figure A12: Embeddings of Gemma-2-Instruct (9B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in this hidden states of the intermediate layer of multilingual LLMs. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, but with minimal overlap.

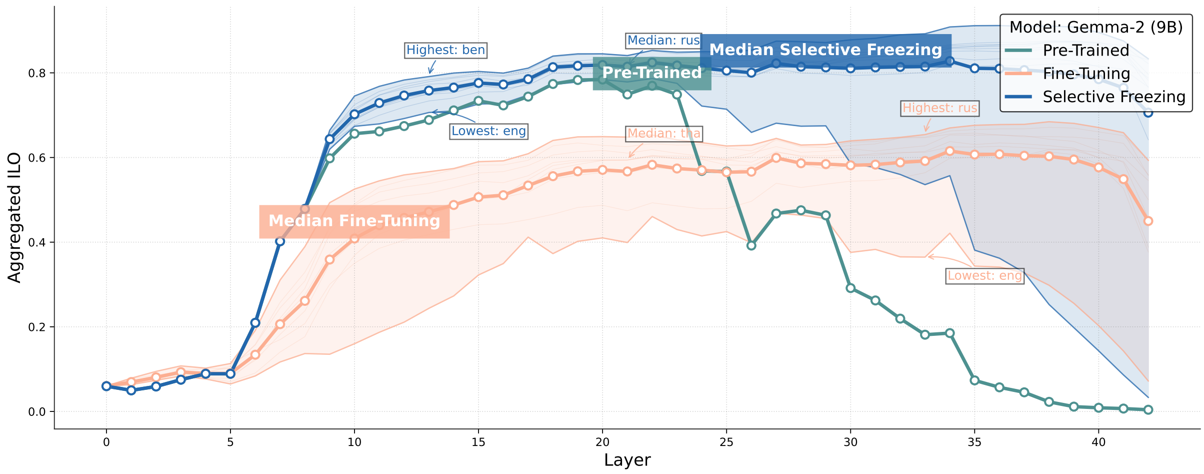
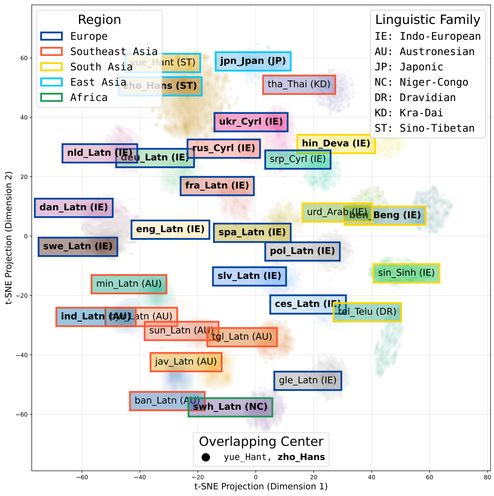
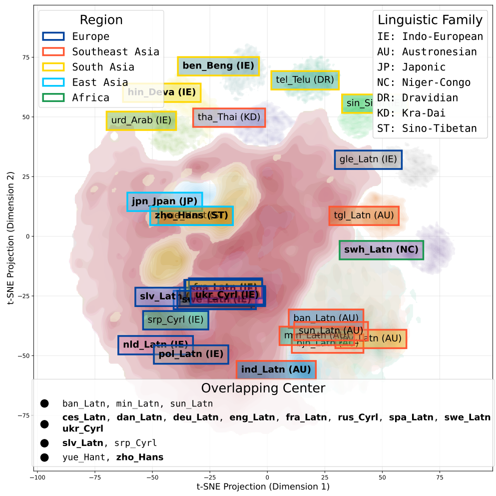


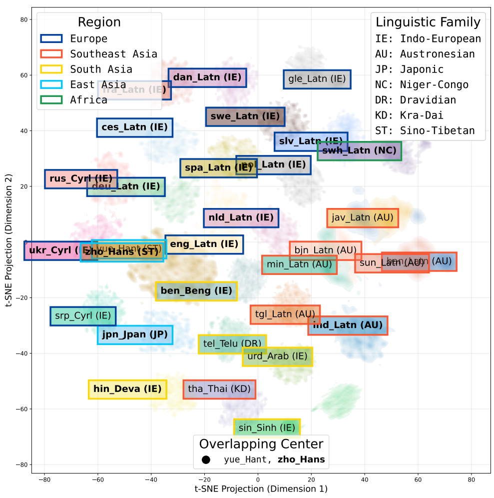
Figure A13: Layer-wise $\bar{\text{ILO}}_{\mathcal{L}}$ scores for Gemma-2 (9B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes. Notable decrease in alignment from single-language training is seen in the early layers on **fine-tuning**, whereas the **selective freezing** mechanism allows the model to sustain its **pre-trained** semantic alignment across layers.



(a) Early (layer 0)

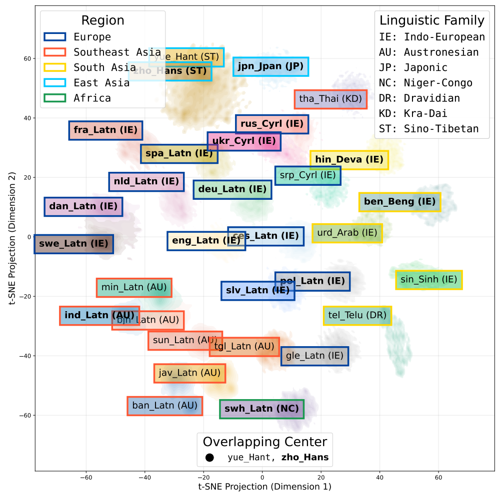


(b) Intermediate (layer 16)

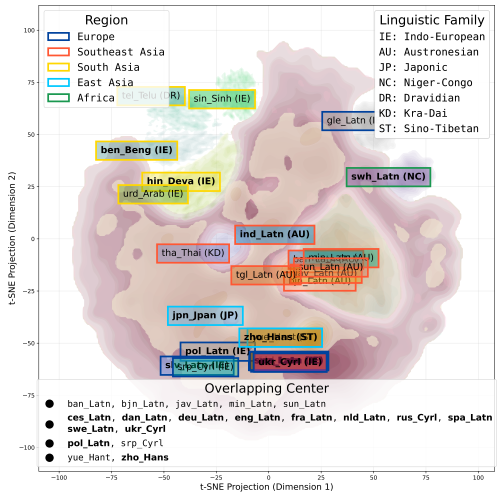


(c) Late (layer 32)

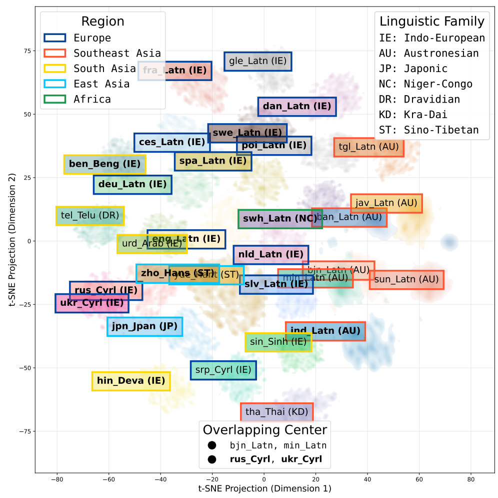
Figure A14: Embeddings of Llama-31 (8B) **fine-tuned** on single-language dataset on English, projected in t-SNE dimensions, with HRLs in **bold**. The decline in interlingual semantic alignment is evident from the reduced interlingual overlaps in the projected embeddings within the model's intermediate layer, compared to the observations in Figure A9.



(a) Early (layer 0)

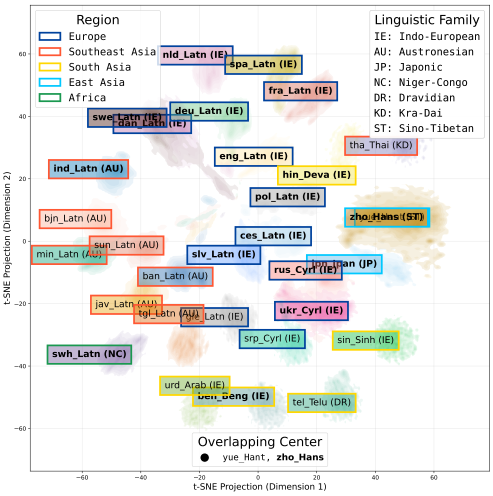


(b) Intermediate (layer 16)

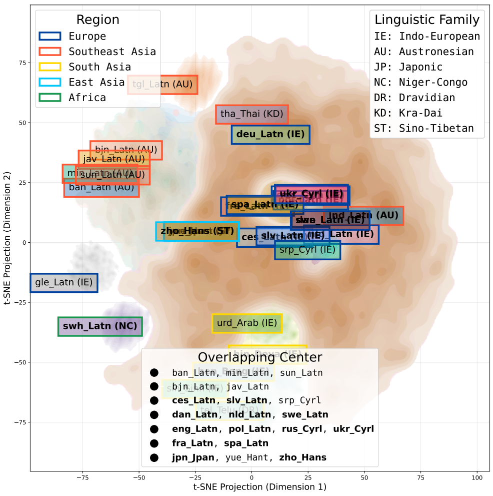


(c) Late (layer 32)

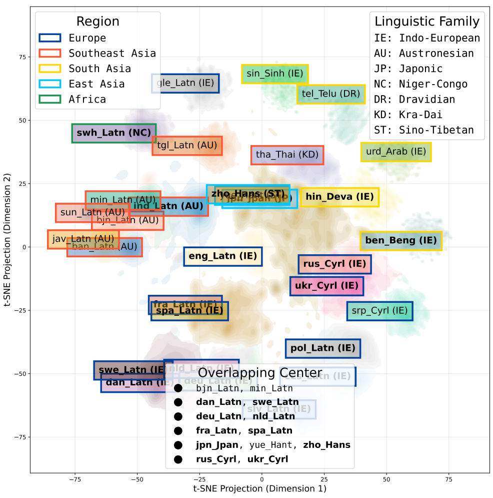
Figure A15: Embeddings of Llama-31 (8B) fine-tuned on single-language dataset on English, with **selective freezing** strategy, projected in t-SNE dimensions, with HRLs in **bold**. This approach preserved interlingual alignment, as indicated by high ILO scores that correlate with observed preservation of interlingual overlaps.



(a) Early (layer 0)

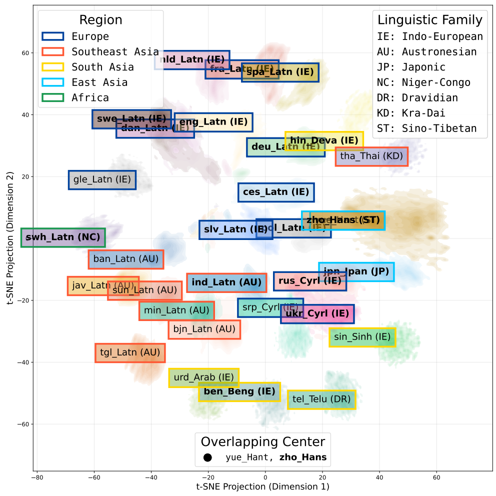


(b) Intermediate (layer 21)

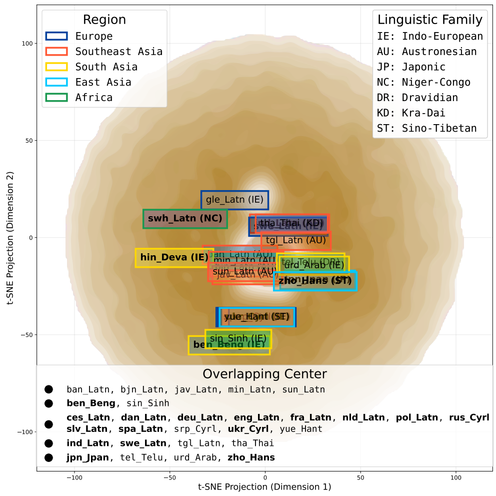


(c) Late (layer 42)

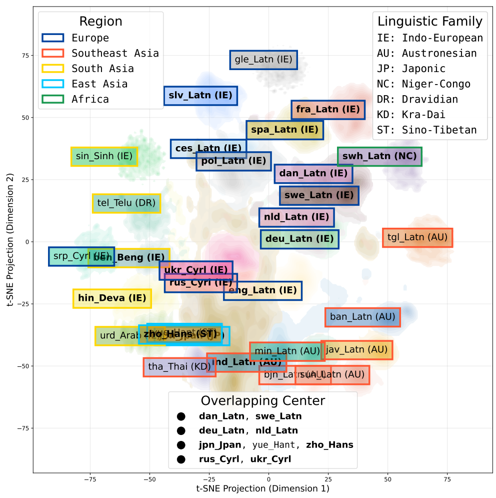
Figure A16: Embeddings of Gemma-2 (9B) **fine-tuned** on single-language dataset on English, projected in t-SNE dimensions, with HRLs in **bold**. The decline in interlingual semantic alignment is evident from the reduced interlingual overlaps in the projected embeddings within the model's intermediate layer, compared to the observations in Figure A11.



(a) Early (layer 0)



(b) Intermediate (layer 21)



(c) Late (layer 42)

Figure A17: Embeddings of Gemma-2 (9B) fine-tuned on single-language dataset on English, with **selective freezing** strategy, projected in t-SNE dimensions, with HRLs in **bold**. This approach preserved interlingual alignment, as indicated by high ILO scores that correlate with observed preservation of interlingual overlaps.