ShifCon: Enhancing Non-Dominant Language Capabilities with a Shift-based Multilingual Contrastive Framework

Hengyuan Zhang[1 * †], **Chenming Shang**[1 †], **Sizhe Wang**[3], **Dongdong Zhang**[2 ✉], **Yiyao Yu**[1], **Feng Yao**[4], **Renliang Sun**[5], **Yujiu Yang**[1✉], **Furu Wei**[2]
[1] Tsinghua University [2] Microsoft [3] University of Southern California
[4] University of California, San Diego [5] University of California, Los Angeles
{zhang-hy22,scm22,yuyy23}@mails.tsinghua.edu.cn
{dozhang,fuwei}@microsoft.com

Abstract

Although fine-tuning Large Language Models (LLMs) with multilingual data can rapidly enhance the multilingual capabilities of LLMs, they still exhibit a performance gap between the dominant language (e.g., English) and non-dominant ones due to the imbalance of training data across languages. To further enhance the performance of non-dominant languages, we propose *ShifCon*, a **Shif**t-based multilingual **Con**trastive framework that aligns the internal forward process of other languages toward that of the dominant one. Specifically, it shifts the representations of non-dominant languages into the dominant language subspace, allowing them to access relatively rich information encoded in the model parameters. The enriched representations are then shifted back into their original language subspace before generation. Moreover, we introduce a subspace distance metric to pinpoint the optimal layer area for shifting representations and employ multilingual contrastive learning to further enhance the alignment of representations within this area. Experiments demonstrate that our *ShifCon* framework significantly enhances the performance of non-dominant languages, particularly for low-resource ones. Further analysis offers extra insights to verify the effectiveness of *ShifCon* and propel future research.

1 Introduction

While LLMs have demonstrated strong multilingual capabilities (Lin et al., 2022; Achiam et al., 2023; Anil et al., 2023), a performance gap remains between the dominant language and non-dominant ones, primarily due to the imbalance in training data across languages (Shi et al., 2022; Huang et al., 2023; Gurgurov et al., 2024). A common strategy to mitigate this issue is translating dominant language data into non-dominant languages and applying Multilingual Supervised Fine-Tuning (MSFT) on the resulting multilingual datasets (Chen et al., 2023a; Zhang et al., 2023b).

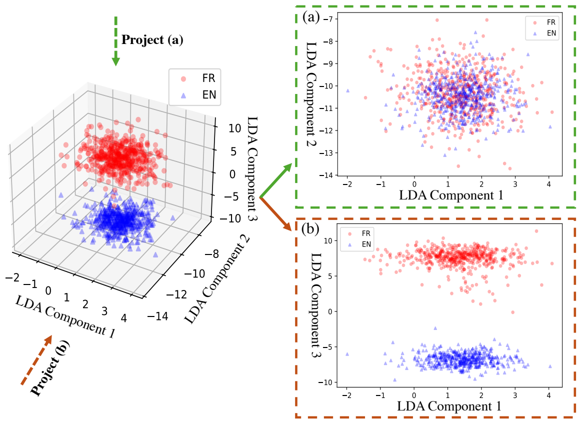

Figure 1: Two different projections on the sentence representations visualized using LDA. Projection (a) shows the representations are mutually aligned, implying a language-agnostic status, whereas projection (b) illustrates separated representations in distinct spaces, suggesting a language-specific status. The sentence representations are obtained through mean-pooling the hidden states from the 15th layer of Llama-2_{7B}.

While MSFT provides initial capabilities for non-dominant languages, two key challenges limit further progress: 1) annotating high-quality data for non-dominant languages is expensive, even for the dominant language that serves as the source for translation (Kholodna et al., 2024); 2) translation errors often lead to error propagation in subsequent procedures (Agrawal et al., 2024), thus requiring extensive verification to ensure data quality. As a result, high-quality data for non-dominant languages is limited in scale, which restricts the effectiveness of MSFT. This raises an important question: *Can we improve the performance of non-dominant languages with limited MSFT data?*

Considering this external limitation, previous work has delved into exploring internal representation alignment to improve performance (Yoon

* This work was done during internship at Microsoft.
† Equal contribution
✉ Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4818–4841
July 27 - August 1, 2025

et al., 2024; Li et al., 2024). A growing consensus indicates that it is the language-agnostic representations, which are exhibited in the middle layer of the model, facilitating this enhancement (Kojima et al., 2024; Tang et al., 2024). Beyond those efforts, we consider that the representations, even in the middle layer, still retain language-specific information. Specifically, by visualizing sentence representations of translation pairs using linear discriminant analysis (LDA) in Fig. 1, we observe representations under projection (a) in the middle layer are mapped closely together (e.g., the 15th layer of Llama-2_{7B} out of 32 layers), suggesting a language-agnostic status, consistent with findings in prior research. However, in projection (b), we find that different languages occupy distinct subspaces across layers, indicating that language-specific information is consistently encoded within the representations (See Appendix A.1 for complete results across all languages, layers, and models). This information enables the model to differentiate between languages. Moreover, we consider the superior performance of dominant languages is due to their representations being able to access more information during the internal forward process. This is because dominant language data predominates during pre-training, so much of the model's knowledge is encoded in the dominant language format, which is more easily accessible through its representations (Kassner et al., 2021; Yin et al., 2022; Zhao et al., 2024).

Based on these findings, we propose a **Shif**t-based multilingual **Con**trastive framework (*ShifCon*) to boost the performance of non-dominant language. It includes shift-toward and shift-backward projections, as well as multilingual contrastive learning (MCL). The shift-toward process maps non-dominant language representations into the dominant language subspace to obtain their dominant-like representations, allowing them to access more information encoded in the model, similar to how the dominant language operates. As language-specific information is crucial for generating outputs in the target language (Li and Murray, 2023; Xu et al., 2023; Tang et al., 2024), a shift-backward process is needed to project the enriched dominant-like representations back into the original non-dominant language subspace before generation. During this process, a subspace distance metric is proposed to pinpoint the optimal layer area for shifting representations. Moreover, our analysis reveals that even after shifting, the alignment between non-dominant language's dominant-like representations and their dominant language counterparts remains insufficient. Therefore, we further apply multilingual contrastive learning to enhance their alignment.

To summarize, our contributions are as follows:

1) We present *ShifCon* framework, designed to boost the performance of non-dominant languages by aligning their internal forward process with that of the dominant language. We also define a subspace distance metric to pinpoint the optimal layer area for implementing shift projection.

2) Extensive experiments validate the efficacy of *ShifCon* across diverse tasks and model scales, e.g., a 18.9% improvement on MGSM for low-resource languages in Llama-2_{7B}. Further analysis confirms the effectiveness of the identified layer area for shift projection using subspace distance metric. The improved alignment between dominant-like representations and their dominant counterparts enhances overall performance.

3) Moreover, we give the speculation that 30% of model layers with the lowest distance are likely focused on information aggregation and show that directly applying MCL to original representations may compromise the language-specific information within representations, which impedes the model's ability to generate in that language.

2 The Framework

Our *ShifCon* (shown in Fig. 2) includes two modules: 1) Shift Projection (§ 2.1), which maps the representations of non-dominant language into the dominant language subspace to obtain its dominant-like representations during internal forward process, and then shifts backwards to its native space before generation; 2) Multilingual Contrastive Learning (§ 2.2), which further aligns dominant-like representations of non-dominant languages with their dominant language counterparts.

2.1 Shift Projection

2.1.1 Shift-toward and Shift-backward

To obtain the dominant-like representations for non-dominant languages, thereby enabling them to access more information encoded in the model parameters during the internal forward process, our shift-toward module maps non-dominant language representations into dominant language subspace.

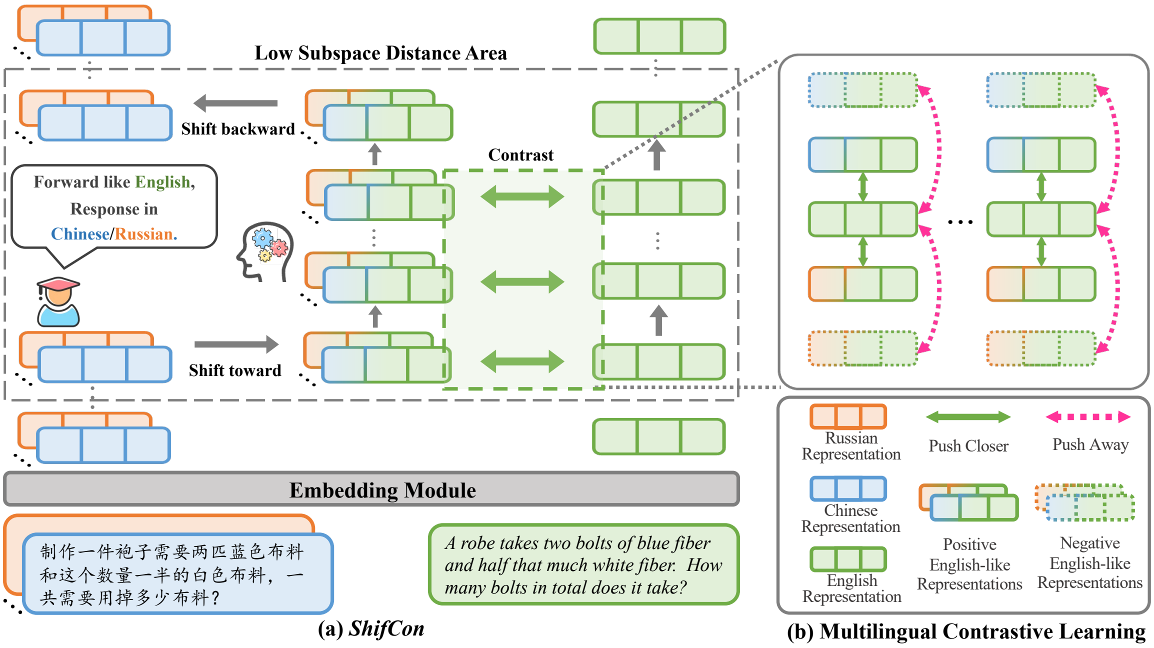

Figure 2: An illustration of our *ShifCon* framework: (I) We shift non-dominant language representations (e.g., Chinese and Russian) into the dominant language subspace (e.g., English) to obtain their dominant-like representations. (II) Using parallel translation inputs between the non-dominant and dominant languages as positive samples, multilingual contrastive learning pushes non-dominant language's dominant-like representations closer to the dominant language and pushes away them from other representations.

Specifically, given an input query in a non-dominant language l, the shift-toward process can be formulated as follows:

$$\tilde{\boldsymbol{h}}_l^{L_{\text{to}}} = \boldsymbol{h}_l^{L_{\text{to}}} - \boldsymbol{v}_l^{L_{\text{to}}} + \boldsymbol{v}_d^{L_{\text{to}}} \; (1 \leq L_{\text{to}} < L) \quad (1)$$

where L_{to} is the layer we shift the representation toward, $\boldsymbol{h}_l^{L_{\text{to}}} \in \mathbb{R}^{n\times d}$ denotes L_{to}-th layer hidden states of the input query in language l, where n is the number of tokens in the input query, d is the hidden dimension of the LLM. $\boldsymbol{v}_l^{L_{\text{to}}} \in \mathbb{R}^d$ and $\boldsymbol{v}_d^{L_{\text{to}}} \in \mathbb{R}^d$ are the L_{to}-th layer language vectors for the non-dominant language l and the dominant language, respectively.[1] To compute the language vectors across all layers for each language l, a set of sentences in that language is fed into the LLM. From the i-th layer of the LLM, sentence vectors are obtained by pooling the token representations[2] within the sentence. These sentence vectors are then averaged to produce $\boldsymbol{v}_l^i \in \mathbb{R}^d$. In this way, we gather a set of vectors $\mathcal{V}_l = [\boldsymbol{v}_l^1, \boldsymbol{v}_l^2, ..., \boldsymbol{v}_l^L]$, where L denotes the number of layers in the LLM. The obtained dominant-like representations of non-dominant language are then fed to the succeeding layers to access relatively rich information encoded in the model parameters.

Since language-specific information is crucial for models to generate answers in that language, we shift dominant-like representations of the non-dominant language back to its native subspace at the L_{bk}-th layer before generation:

$$\boldsymbol{h'}_l^{L_{\text{bk}}} = \tilde{\boldsymbol{h}}_l^{L_{\text{bk}}} - \boldsymbol{v}_{\text{d}}^{L_{\text{bk}}} + \boldsymbol{v}_l^{L_{\text{bk}}} \; (L_{\text{to}} < L_{\text{bk}} \leq L) \quad (2)$$

where L_{bk} is the layer we shift the representation backward, $\tilde{\boldsymbol{h}}_l^{L_{\text{bk}}}$ represent the L_{bk}-th layer hidden states of non-diminant language l. $\tilde{\boldsymbol{h}}_l^{L_{\text{bk}}}$ are dominant-like representations because of the shift-toward projection. They are shifted back into their original subspace, resulting in $\boldsymbol{h'}_l^{L_{\text{bk}}}$. The representations, now containing language-specific information of l, are then fed into the subsequent layers to produce responses in language l.

2.1.2 Language Subspace Distance

It is crucial to establish an effective criterion for determining the optimal layer area for conducting shift projection procedure. A practical solution is to select layers where the subspace of non-dominant language's dominant-like representations[3] aligns

[1] We utilize language vectors in the shift projection process, as it has been demonstrated to be an effective approach for language space mapping (Libovický et al., 2020; Xu et al., 2023; Tang et al., 2024).

[2] We explore different pooling methods in Appendix A.3.

[3] We term the subspace of non-dominant language's dominant-like representations as "dominant-like subspace".

well with the subspace of the dominant language counterparts, as greater alignment indicates they can be more similar in the internal forward process.

Therefore, we propose a subspace distance metric to measure the alignment between their subspaces, where smaller distances indicating stronger alignment. Specifically, for the language A, we define an affine subspace $\mathcal{S}^{\mathcal{A}}$ using the language's mean representation $\boldsymbol{\mu}_A \in \mathbb{R}^d$ along with k_A principal directions of maximal variance in the language, defined by an orthonormal basis $\boldsymbol{V}_A \in \mathbb{R}^{d \times k_A}$. We consider this basis with k_A directions can best describe the language-specific information of language A. To identify this subspace, we use $\boldsymbol{X}_A \in \mathbb{R}^{n \times d}$ to obtain $\boldsymbol{\mu}_A$ and employ singular value decomposition (SVD) on the $\boldsymbol{X}_A$ to obtain $\boldsymbol{V}_A$, which is selected from the top-k_A singular value by $\boldsymbol{\Sigma}_A \in \mathbb{R}^{k_A \times k_A}$. Here, $\boldsymbol{X}_A$ donates n contextualized token representations with d dimensionality in language A from the desired layer. We select the subspace dimensionality k such that the subspace accounted for 90% of the total variance in the language.[4]

Due to the varying dimensionality k of $\boldsymbol{V}$ across different languages, we adopt a Riemannian distance metric that measures distances between positive definite matrices (Bonnabel and Sepulchre, 2009; Chang et al., 2022) to quantify the distance between dominant-like subspace $\mathcal{S}^{\mathcal{D}'}$ and corresponding dominant language subspace $\mathcal{S}^{\mathcal{D}}$:[5]

$$\text{Dist}(\mathcal{S}^{\mathcal{D}'}, \mathcal{S}^{\mathcal{D}}) = \sqrt{\sum_{i=1}^{d} \log^2(\lambda_i) + ||\boldsymbol{\mu}_{D'} - \boldsymbol{\mu}_D||_2} \quad (3)$$

where λ_i is the i-th positive real eigenvalue of $\boldsymbol{K}_{D'}^{-1}\boldsymbol{K}_D$. Here $\boldsymbol{K}_D \in \mathbb{R}^{d \times d}$ can be calculated from the SVD of the right singular matrices $\boldsymbol{V}_D$:

$$\boldsymbol{K}_D = \frac{1}{n-1}\boldsymbol{V}_D \boldsymbol{\Sigma}_D^2 \boldsymbol{V}_D^T \quad (4)$$

We present the distance results of the $\text{XGLM}_{7.5\text{B}}$ in Fig. 3. We observe that the subspace distances in the middle layers are minimal, while the distances on the sides are larger with steep slopes. This observation suggests that the middle layers in the model achieves superior alignment between dominant-like representations and their dominant language counterparts, enabling them access richer information analogous to dominant language representations, rendering it suitable for shift projection.

[4]See more details of computing process in Appendix A.2.

[5]After applying shift projection, the centroids of two subspaces will coincide, causing $||\boldsymbol{\mu}_{D'} - \boldsymbol{\mu}_D||_2 = 0$.

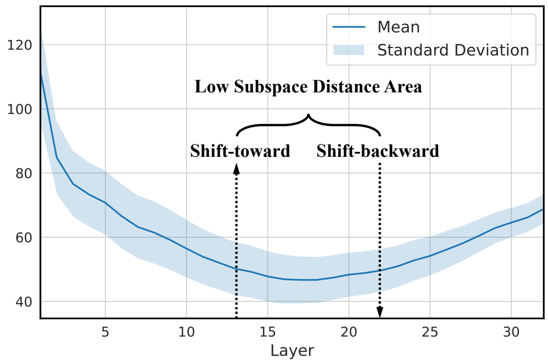

Figure 3: The distance of dominant-like subspace $\mathcal{S}^{\mathcal{D}'}$ and corresponding dominant language subspace $\mathcal{S}^{\mathcal{D}}$ in the $\text{XGLM}_{7.5\text{B}}$ using 1k FLORES samples per language. Its low subspace distance area, [13, 22], identified by β=30% (Finding 1), indicating shifting towards in the 13th layer and backward in the 22nd layer.

To precisely identify these layers, we propose a simple method of sorting the distances in ascending order and selecting the top-β[6] layers with the smallest distances to establish the *low subspace distance area*. We find that the layers within the low subspace distance area are contiguous across models of different families and scales, making them ideally suited for shift projection.

2.2 Multilingual Contrastive Learning (MCL)

However, as shown in Fig. 3, some subspace distance still remains, even in the low subspace distance area (e.g., $\text{XGLM}_{7.5\text{B}}$'s 16th layer still exhibits a subspace distance of about 47), which requires further alignment to reduce. To address this, we employ multilingual contrastive learning to achieve a more refined alignment. We use translation pairs from dominant and non-dominant languages as positive pairs, pulling the dominant-like representations of non-dominant language closer to their dominant language counterparts. While the dominant-like representations of other sentences in the same batch serve as negative samples.

Formally, given a mini-batch of translation pairs from non-dominant and dominant languages $\{(s_l^i, s_d^i)\}_{i=1}^N$, the Multilingual Contrastive Learning (MCL) loss at the t-th layer is:

$$\begin{gathered} \tilde{\boldsymbol{e}}_l^i = g(\left[\tilde{\boldsymbol{h}}_l^t\right]^i); \qquad \boldsymbol{e}_d^i = g(\left[\boldsymbol{h}_d^t\right]^i) \\ \mathcal{L}_{MCL}^t(\theta) = \sum_{i=1}^{N} -\log \frac{\exp(\text{sim}(\tilde{\boldsymbol{e}}_l^i, \boldsymbol{e}_d^i)/\tau)}{\sum_j \exp(\text{sim}(\tilde{\boldsymbol{e}}_l^i, \boldsymbol{e}_d^j)/\tau)} \end{gathered} \quad (5)$$

[6]We test β from 0% to 100%, choosing $\lceil N \times \beta \rceil$ layers to define the low subspace distance area. $\lceil \cdot \rceil$ is ceiling function.

	Generation						Classification					
	MGSM		FLORES (en-xx)		FLORES (xx-en)		XCOPA		XNLI		XStoryCloze	
	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
Llama-2_{7B}	35.2	5.1	33.5	15.9	39.8	21.4	63.2	49.7	45.2	35.2	74.7	56.6
+MSFT	44.9	29.5	34.7	18.4	40.4	24.7	64.2	52.0	46.4	37.6	75.3	58.7
+AFP	46.3	31.7	35.2	19.1	41.0	25.3	65.0	52.8	46.8	38.7	76.0	59.8
+*ShifCon*	**48.2**	**35.1**	**35.6**	**19.7**	**41.8**	**26.4**	**65.5**	**53.5**	**47.2**	**40.1**	**76.6**	**60.8**
$\text{XGLM}_{7.5B}$	4.0	1.9	32.2	31.5	41.2	35.8	63.8	57.3	44.5	41.4	65.2	58.4
+MSFT	10.6	7.0	33.5	32.8	42.3	37.3	64.9	58.3	45.9	42.3	66.7	60.1
+AFP	12.1	9.6	34.0	33.3	43.2	37.7	65.7	58.9	47.0	43.3	67.4	60.9
+*ShifCon*	**13.7**	**11.7**	**34.5**	**34.1**	**43.7**	**38.5**	**66.8**	**60.1**	**48.6**	**44.3**	**68.1**	**62.2**
$\text{BLOOM}_{7.1B}$	13.2	3.7	41.4	24.3	45.7	30.7	57.7	52.1	42.4	36.6	67.3	58.1
+MSFT	21.9	12.5	42.3	25.9	46.5	33.1	59.2	53.9	44.0	38.9	68.6	59.8
+AFP	22.9	15.7	43.0	26.6	47.0	33.6	59.9	54.8	44.9	39.9	68.9	60.2
+*ShifCon*	**24.5**	**18.8**	**43.4**	**27.2**	**47.2**	**34.5**	**60.3**	**56.3**	**45.5**	**40.8**	**69.5**	**60.9**

Table 1: The average results of high- and low-resource languages across five tasks within three distinct model families. Detailed results for each language can be found in Appendix A.7. "en-xx" denotes translation from English to another language, while "xx-en" indicates translation from another language to English. Base model, e.g., Llama-2_{7B}, indicates fine-tuning solely with English data.

where $g(\cdot)$ is the pooling method used to obtain sentence representations, $\left[\tilde{\boldsymbol{h}}_l^t\right]^i$ denotes the t-th layer dominant-like representations of s_l^i, $\left[\boldsymbol{h}_d^t\right]^i$ is the t-th layer representations of s_d^i, and $\text{sim}(,)$ is cosine similarity function. τ is a temperature hyperparameter. MCL is performed on the layers between $[L_{\text{to}}, L_{\text{bk}})$ to achieve better alignment, resulting in the total MCL loss: $\mathcal{L}_{MCL} = \sum_{t=L_{\text{to}}}^{L_{\text{bk}}-1} \mathcal{L}_{MCL}^t$.

We illustrate the process of MCL in Fig. 2 (b) and train our *ShifCon* using the following loss:

$$\mathcal{L}_{ShifCon}(\theta) = \mathcal{L}_{MSFT}(\theta) + \alpha\mathcal{L}_{MCL}(\theta) \quad (6)$$

where $\mathcal{L}_{MSFT}$ denotes the loss of MSFT, computed through autoregressive language modeling on the multilingual dataset, and $\alpha \in \mathbb{R}_+$ is a hyperparameter to balance these two losses. It is important to note that when computing $\mathcal{L}_{MSFT}$ for non-dominant language samples, their dominant-like representations are used during the internal forward process instead of their original ones.[7]

3 Experiment

3.1 Experiment Settings

Evaluation Tasks We conduct evaluations on a variety of multilingual benchmarks, covering both generation and classification tasks. **1)** For generation tasks, we consider FLORES (Team", 2022), a benchmark for machine translation, and MGSM (Shi et al.), a multilingual math reasoning task. **2)** For classification tasks, we utilize XNLI (Conneau et al., 2018), XCOPA (Ponti et al., 2020), and XStoryCloze (Lin et al., 2022), which are widely used generic reasoning datasets.

For the evaluation of MGSM, we utilize MGSM8KInstruct (Chen et al., 2023a) as the training set, which translates the GSM8K into nine non-English languages. For the evaluation of the other tasks, we follow Li et al. (2024) and utilize Bactrian-X (Li et al., 2023b), which has been translated into 52 languages from Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023), as the training set. See Appendix A.4 for more details about the datasets we used in the experiment.

Metrics For MGSM, we implement a rule-based extraction strategy (Chen et al., 2023a) to derive accuracy results in a zero-shot manner. We utilize the evaluation framework introduced by Zhang et al. (2024c) for assessing the other benchmarks in a 4-shot manner. Specifically, we assess the performance on the FLORES dataset using ChrF++ (Popović, 2017) score, while the performance on the other datasets is evaluated based on rank classification accuracy.[8]

[7] In this work, we introduce a new strategy to obtain better language vectors for shift projection in the training phase. The details are illustrated in Appendix A.6.

[8] The scoring function averages per-token logarithmic probabilities, excluding shared prefixes. The candidate with the highest score is chosen as the prediction.

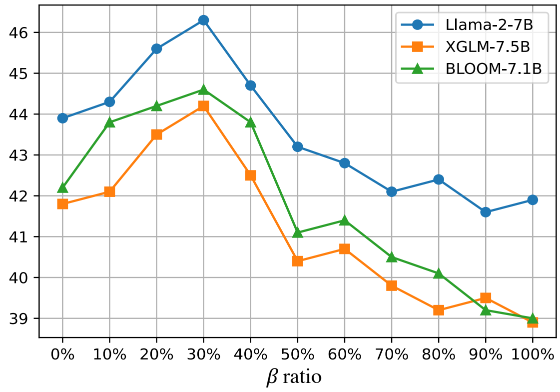

Figure 4: The average results of all benchmarks across different β ratios in three distinct family models.

Training Setup We incorporate LLMs from different families, such as Llama (Touvron et al., 2023), BLOOM (Scao et al., 2022), and XGLM (Lin et al., 2022), in our experiments. We utilize ***English*** as the dominant language in these three model families, as its data predominates in their corresponding pre-training corpus. The models trained using **MSFT** and the state-of-the-art alignment framework **AFP** (Li et al., 2024), serve as the baseline for comparison. Since both MGSM8KInstruct and Bactrian-X are constructed through translation, we directly extract the instruction content from their respective datasets to acquire the translation pairs for MCL. The details of model information and training settings can be found in Appendix A.5.

3.2 Performance of *ShifCon*

We categorize the experimental languages into high- and low-resource languages based on their data ratios in the LLM pre-training corpus, and report their average results across different tasks in Table 1. As shown in Table 1, despite the initial capabilities provided by MSFT for non-dominant languages, our *ShifCon* consistently further boosts their performance. Specifically, for XGLM$_{7.5B}$, our *ShifCon* improves performance by 2.1% for the high-resource languages on XCOPA and a more substantial improvement of 3.5% for the low-resource languages. Moreover, we observe that the enhancement of multilingual understanding also facilitates generation. For example, *ShifCon* exhibits an improvement of 7.3% on high-resource languages on MGSM and a more significant improvement of 18.9% on low-resource languages. Based on these observations, we conclude that: *ShifCon improves the performance of non-dominant languages, especially for low-resource languages.*

	XCOPA		XNLI		XStoryCloze	
	High	Low	High	Low	High	Low
XGLM$_{564M}$	54.3	51.1	37.6	35.2	56.1	53.0
+MSFT	56.5	52.7	40.4	37.8	57.5	55.4
+AFP	57.3	53.9	41.5	39.1	58.0	56.6
+*ShifCon*	**58.4**	**55.8**	**42.6**	**40.5**	**59.8**	**58.1**
XGLM$_{2.9B}$	61.5	54.9	41.8	37.6	61.7	54.9
+MSFT	63.4	57.2	44.6	40.5	64.1	57.6
+AFP	64.0	58.4	45.2	41.4	65.3	58.8
+*ShifCon*	**65.5**	**59.8**	**46.8**	**43.3**	**66.5**	**60.4**
BLOOM$_{560M}$	53.8	51.2	39.8	34.2	60.3	54.2
+MSFT	55.1	52.3	41.7	35.4	62.2	54.1
+AFP	55.8	53.2	42.6	36.6	62.8	55.3
+*ShifCon*	**56.7**	**54.8**	**43.5**	**38.2**	**63.6**	**56.8**
BLOOM$_{1.7B}$	55.4	51.7	41.5	35.3	62.4	54.8
+MSFT	56.9	53.4	43.2	36.3	64.6	56.3
+AFP	57.8	54.5	44.0	37.3	65.2	57.6
+*ShifCon*	**58.7**	**55.8**	**44.8**	**38.9**	**66.8**	**59.2**
Llama-3$_{8B}$	68.6	54.3	50.6	41.5	78.5	63.9
+MSFT	69.0	55.1	51.1	42.4	78.8	64.7
+AFP	69.3	56.0	51.3	43.1	79.1	65.6
+*ShifCon*	**69.7**	**56.9**	**51.6**	**44.2**	**79.5**	**66.4**

Table 2: The average performance of high- and low-resource languages across three classification tasks under model of different scales and families. Base model indicates fine-tuning solely with English data.

3.3 Further Analysis

Suitable β for Shift Projection We conduct extra experiments to determine the number of layers for non-dominant languages to perform in their dominant-like representation during the internal forward process. In Fig. 4, the average performance of all benchmarks across three model families is shown for various selection ratios β (as defined in § 2.1), ranging from 0% to 100%. The results indicate a trend of initially increasing, peaking at a value of 30%, and subsequently declining. Similar trends can be observed in three models of different families. Therefore, we set β to 30% by default to obtain the *low subspace distance area* in our *ShifCon* framework and give the following speculation:

> **Finding 1.** *$N \times 30\%$ of layers with lowest subspace distance are likely focused on information aggregation, making them suitable for non-dominant languages to forward in dominant-like representations.*

Where N denotes the number of layers in the model, and this speculation also aligns with the findings observed by Zhang et al. (2024a).

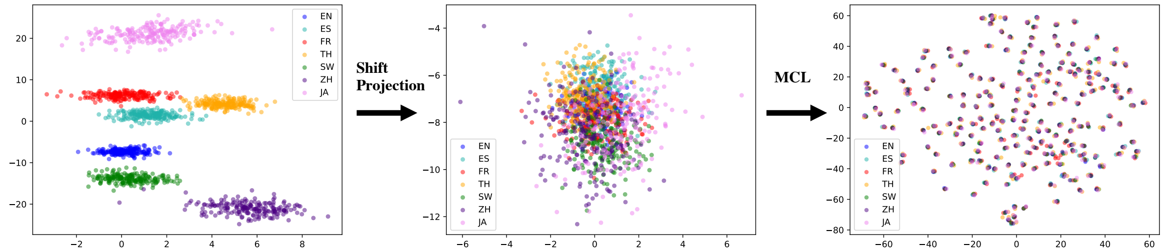

Figure 5: Pooled sentence representations obtained with 300 FLORES samples per language from 15th layer of Llama-2_{7B} after utilizing shift projection and MCL modules. Visualization is based on LDA components 1 and 3.

	Llama-2_{7B}	XGLM$_{7.5B}$	BLOOM$_{7.1B}$
ShifCon	**46.0**	**43.8**	**44.1**
w/o Shift Projection	42.2	39.9	40.7
w/o MCL	44.5	43.1	42.8

Table 3: The impact of Shift Projection and MCL in *ShifCon* on the average results of all benchmarks. "w/o" means excluding this module from *ShifCon*.

	Llama-2_{7B}	XGLM$_{7.5B}$	BLOOM$_{7.1B}$
ShifCon	**96.9**	**97.6**	94.9
w/o Shift Projection	87.6	91.6	88.8
w/o MCL	96.6	97.3	**95.5**

Table 4: The average results of the language consistency on the MGSM task. "w/o" means excluding this module from *ShifCon*.

Performance of *ShifCon* across Different Scales Having verified the effectiveness of our *ShifCon* across different model families, we further assess its generalization on different model scales across three classification datasets. In the BLOOM family models, experiments are conducted at scales of 560M and 1.7B. For the XGLM family models, we utilize 564M and 2.9B scales, and for the Llama family model, we employ the Llama-3_{8B} (Grattafiori et al., 2024). The average results for high- and low-resource languages are presented in Table 2. The results reveal that our *ShifCon* framework continues to exhibit superior performance compared to MSFT. Specifically, in XGLM family models, *ShifCon* demonstrates average improvements of 4.9% and 4.5% for the 564M and 2.9B scales, respectively. For BLOOM family models, *ShifCon* shows average improvements of 4.1% and 4.3% for the 560M and 1.7B scales, respectively. For Llama-3_{8B}, *ShifCon* achieves an average improvement of 2.2%, a relatively modest gain compared to other models. This can be attributed to the inherently stronger multilingual capabilities of Llama-3_{8B}. Nonetheless, the application of *ShifCon* still brings benefits, particularly for low-resource languages. We believe this improvement is due to the notable performance gaps that remain for these languages, which our framework helps to mitigate. Based on these observations, we derive the conclusion below: *ShifCon can generalize to models across different families and scales, which could be attributed to the selection of appropriate layers determined by the subspace distance metric.*

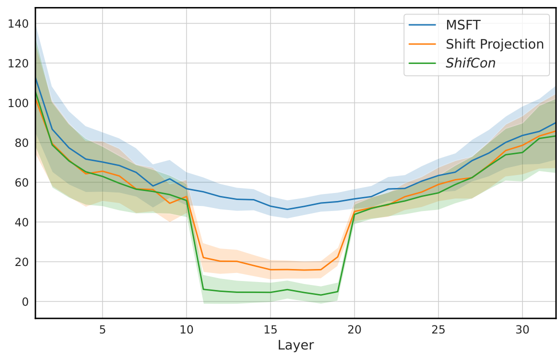

Figure 6: The subspace distances of Llama-2_{7B} after implementing shift projection and MCL.

Impact of Shift Projection and MCL Moreover, we investigate the impact of Shift Projection and MCL within *ShifCon*. Table 3 shows a performance decrease on "*ShifCon* w/o Shift Projection", indicating that directly implementing MCL using original representations of non-dominant languages, instead of their dominant-like counterparts, leads to this decline. We posit that applied MCL directly on original representations may compromise language-specific information within the representations, as it aims to bring representations of different languages with the same meaning closer together, making them become language-agnostic.

To explore this further, we follow Zhang et al. (2024b) to employ a language detector[9] tool to assess the language consistency of input and output

[9] https://pypi.org/project/langdetect

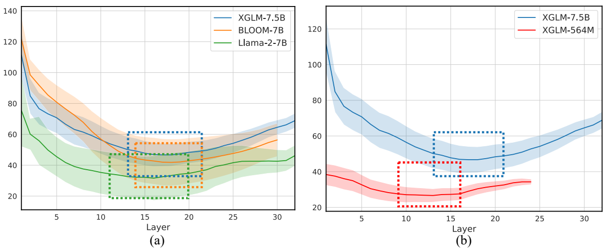

Figure 7: The low subspace distance areas of different models are delineated with dashed boxes. (a) shows the results for different model families; (b) shows the results for different scales of XGLM.

between *ShifCon* and "*ShifCon* w/o Shift Projection". As shown in Table 4, a decrease in language consistency occurs when MCL is directly applied to the original representations. Based on this observation, we give the following conclusion:

Finding 2. *Directly applying MCL to original representations may compromise the language-specific information within representations, which impedes the model's ability to generate in that language, thereby adversely affecting performance.*

Moreover, comparing *ShifCon* and "*ShifCon* w/o MCL", the performance increases. To delve deeper, we visualize the distribution of sentence representations and subspace distance between *ShifCon* and "*ShifCon* w/o MCL" in Fig. 5 and Fig. 6, respectively. The visualization reveals that:

Finding 3. *MCL can further align the dominant-like representations of non-dominant language with its dominant language counterparts, thereby improving overall performance.*

Low Subspace Distance Area In Fig. 7, we show the subspace distance areas of different models utilizing the β value discovered in Finding 1. As depicted in Fig. 7 (a), we observe that the low subspace distance areas of Llama-2$_{7B}$, XGLM$_{7.5B}$, and BLOOM$_{7.1B}$ are [11, 20], [13, 22], and [14, 22] respectively. This indicates that:

Finding 4. *The low subspace distance areas of models from different families vary but generally locate in the middle and late-middle layers.*

Moreover, the subspace distances of XGLM$_{7.5B}$ and BLOOM$_{7.1B}$ are higher than Llama-2$_{7B}$, possibly due to they are being pre-trained on large-scale multilingual data, allowing them to learn more isolated representations for each language.

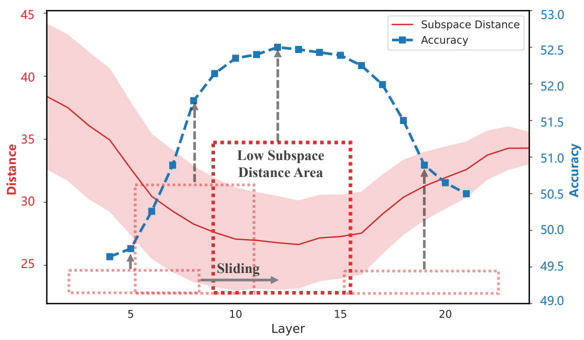

Figure 8: The subspace distance of the XGLM$_{564M}$ and its average performance across three classification tasks using various layer areas. Each point's result denotes a model trained with the specific layer index as the medium of the layer area, such as the 5th layer index indicating a model trained with the [2, 8] layer area.

Another observation we find is that:

Finding 5. *Models from the same family, despite having different layers, exhibit similar locations in the model for their low subspace distance areas.*

Specifically, in Fig. 7 (b), the low subspace distance areas of XGLM$_{7.5B}$ and XGLM$_{564M}$ are [13, 22] and [9, 16], respectively, both situated in the middle of the model. Additionally, the subspace distance of XGLM$_{7.5B}$ is higher than XGLM$_{564M}$, possibly due to larger models showcasing enhanced language discrimination abilities.

Effectiveness of Subspace Distance Area and Metric We conduct extra experiments to verify if the layers within low subspace distance area are suitable for our *ShifCon* framework. Specifically, for the XGLM$_{564M}$ with 24 layers, we select $\lceil 24 \times 30\% \rceil = 8$ layers to apply our *ShifCon*. We explore the performance of shift projection in regions beyond its low subspace distance area [9, 16] in a 8 layers sliding window manner.

As shown in Fig. 8, as we slide the experimental layer area window from left to right, conducting *ShifCon* in layer areas that exhibit great overlap with low subspace distance areas results in improved performance. Moreover, as depicted in Fig. 7, we find that the subspace distances of layers within the low subspace distance area are close. This suggests that the language-specific information within the representations remains relatively unchanged, *resulting in a stable distance between the subspaces of languages*. We speculate the model in these layers may focus on processing semantic information. Based on these two observations, we give the following speculation:

> **Finding 6.** *Layers in the low subspace distance area are likely focused on information aggregation, thus aiding in gathering more information for non-dominant languages and enhancing performance.*

This observation also highlights the effectiveness of our proposed distance metric (§ 2.1.2) in identifying the optimal layer area for our *ShifCon*.

4 Related Work

Multilingual Bias in LLMs Large Language Models (LLMs) have demonstrated remarkable multilingual capabilities as a result of their training on extensive and diverse multilingual datasets. These models have shown proficiency in various aspects of language processing across multiple languages, including multilingual reasoning, understanding, and generation (Xue et al., 2021; Lin et al., 2022; Anil et al., 2023). However, empirical analysis indicates limited proficiency in low-resource languages, stemming from training data imbalances (Huang et al., 2023; Zhu et al., 2024b; Gurgurov et al., 2024) and distinct representation spaces (Wen-Yi and Mimno, 2023; Liu et al., 2024; Yao et al., 2024). Several studies have focused on scaling multilingual corpora through translation, which can provide preliminary capabilities for non-dominant languages. However, this approach is limited in both scale and quality due to the high cost of translated annotations and the presence of translation errors (Muennighoff et al., 2023; Zhang et al., 2023b; Chen et al., 2023b; Tan et al., 2024). In this study, we propose an internal alignment framework to further enhance the performance of non-dominant languages with limited MSFT data.

Representation Alignment Previous studies have shown that projecting representations from the source to the target domain can mitigate domain discrepancies, facilitating effective cross-domain alignment and enhancing performance without disturbing the original domain subspace (Kozhevnikov and Titov, 2014; Chang et al., 2022; Xu et al., 2023; Zhu et al., 2024a). However, this method often results in coarse alignment due to its unsupervised nature. On the other hand, contrastive learning offers a more detailed representation learning approach by utilizing positive and negative pairs to encourage proximity within positive pairs and distance between negative pairs in a supervised manner. This method is better at capturing the complex relationships between representations and achieving precise alignment (Radford et al., 2021; Zhang et al., 2022; Li et al., 2023a; Zhang et al., 2023a, 2025; Li et al., 2024). Drawing from these insights, our framework first employs mean-shifted projection to map non-dominant language representations into the dominant language subspace, preserving language-specific information, and then applies contrastive learning for further alignment.

5 Conclusion

This work aims to improve the performance of non-dominant languages with limited MSFT data. To achieve this, we propose *ShifCon* framework, which aims to align the internal forward process of non-dominant languages with that of the dominant language. It maps the representations of non-dominant languages into the dominant language's subspace to acquire their dominant-like representations, allowing them to access more information encoded in the model parameters. The dominant-like representations are then shifted back to their native subspace to yield answers in their languages. Furthermore, we propose a subspace distance metric to determine the optimal layer area for shift projection, and we apply multilingual contrastive learning to further enhance the internal alignment. The experimental results demonstrate that our proposed *ShifCon* effectively improves the performance of non-dominant languages across models of various families and scales. Our comprehensive analysis offers valuable insights for future research.

6 Limitations

The *ShifCon* framework leverages translation pairs to conduct multilingual contrastive learning, which may pose challenges for low-resource languages or those lacking substantial parallel corpora. Furthermore, due to computational resource limitations, the framework is restricted to multilingual generative language models with parameters not exceeding 8B.

Additionally, our forthcoming research endeavors will delve into exploring alternative model architectures, such as encoder-decoder models, to showcase the full potential and versatility of our proposed framework.

References

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*.

Ashish Agrawal, Barah Fazili, and Preethi Jyothi. 2024. Translation errors significantly impact low-resource languages in cross-lingual learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–329, St. Julian's, Malta. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Silvére Bonnabel and Rodolphe Sepulchre. 2009. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31:1055–1070.

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, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023a. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023b. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

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.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *databricks*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. 2024. Multilingual large language models and curse of multilinguality. *arXiv preprint arXiv:2406.10602*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 397–412. Springer.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. *arXiv preprint arXiv:2203.12926*.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–585.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. 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 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Dawei Li, Hengyuan Zhang, Yanran Li, and Shiping Yang. 2023a. Multi-level contrastive learning for script-based character understanding. *arXiv preprint arXiv:2310.13231*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x:

Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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

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

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 *International Conference on Learning Representations (ICLR)*.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

"NLLB Team". 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. 2024. Data contamination can cross language barriers. *arXiv preprint arXiv:2406.13236*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023a. Assisting language learners: Automated trans-lingual definition generation via contrastive prompt learning. *arXiv preprint arXiv:2306.06058*.

Hengyuan Zhang, Dawei Li, Shiping Yang, and Yanran Li. 2022. Fine-grained contrastive learning for definition generation.

Hengyuan Zhang, Zitao Liu, Chenming Shang, Dawei Li, and Yong Jiang. 2025. A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models. *ACM Transactions on Knowledge Discovery from Data*, 19(2):1–25.

Hengyuan Zhang, Yanru Wu, Dawei Li, Sak Yang, Rui Zhao, Yong Jiang, and Fei Tan. 2024a. Balancing speciality and versatility: a coarse to fine framework for supervised fine-tuning large language model. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7467–7509.

Liang Zhang, Qin Jin, Haoyang Huang, Dongdong Zhang, and Furu Wei. 2024b. Respond in my language: Mitigating language inconsistency in response generation based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4177–4192, Bangkok, Thailand. Association for Computational Linguistics.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024c. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7342–7371, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

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

Mu Zhu, Qingzhou Wu, Zhongli Bai, Yu Song, and Qiang Gao. 2024a. Eeg-eye movement based subject dependence, cross-subject, and cross-session emotion recognition with multidimensional homogeneous encoding space alignment. *Expert Systems with Applications*, 251:124001.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

A Appendix

A.1 Visualization of Sentence Representations across Layers

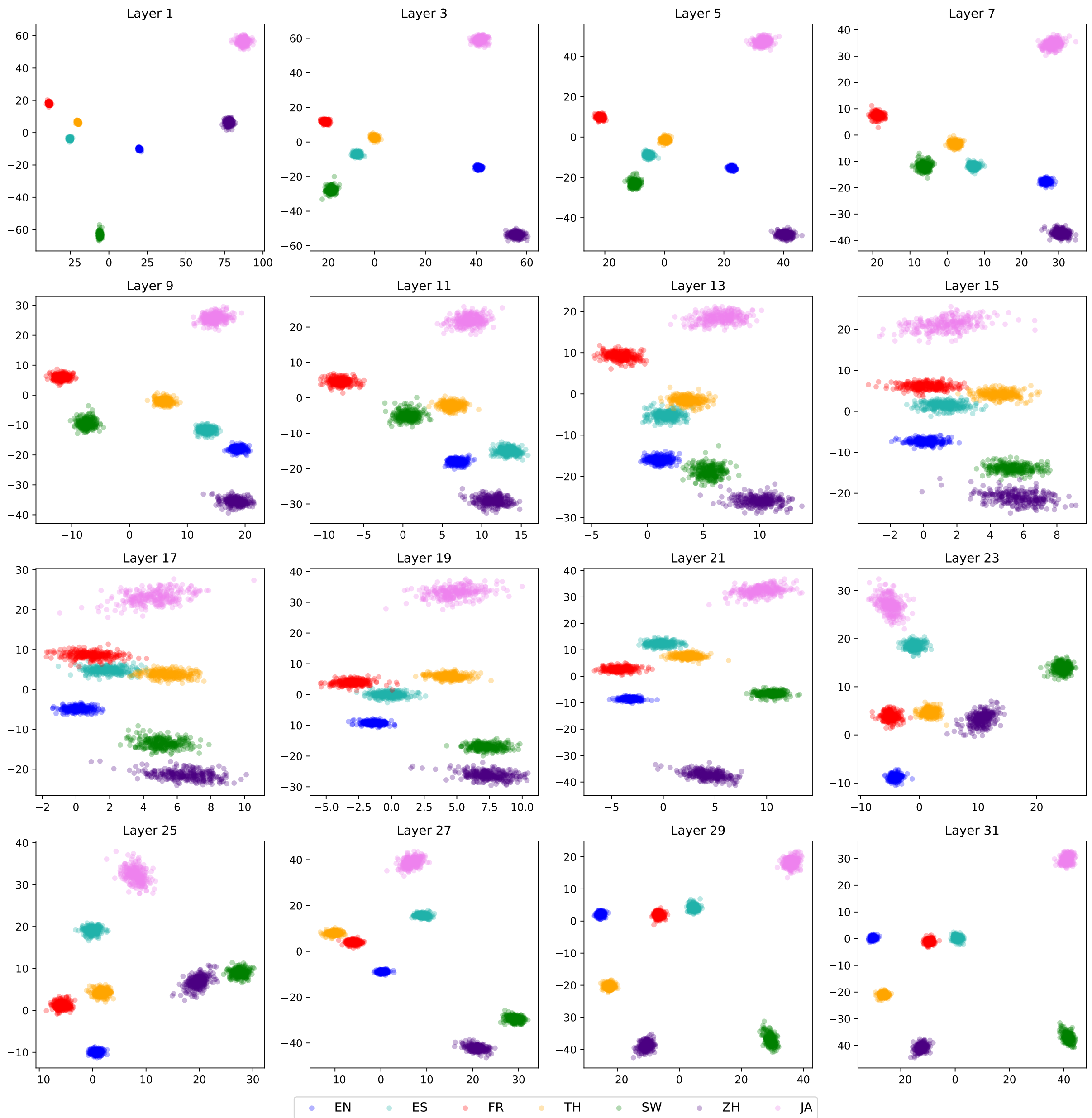

Figure 9: We follow Chang et al. (2022) to conduct LDA and present the visualization of sentence representations obtained by mean-pooling from Llama-2_{7B} across layers along LDA components 1 and 3. We utilize 300 samples for each language from the FLORES dataset.

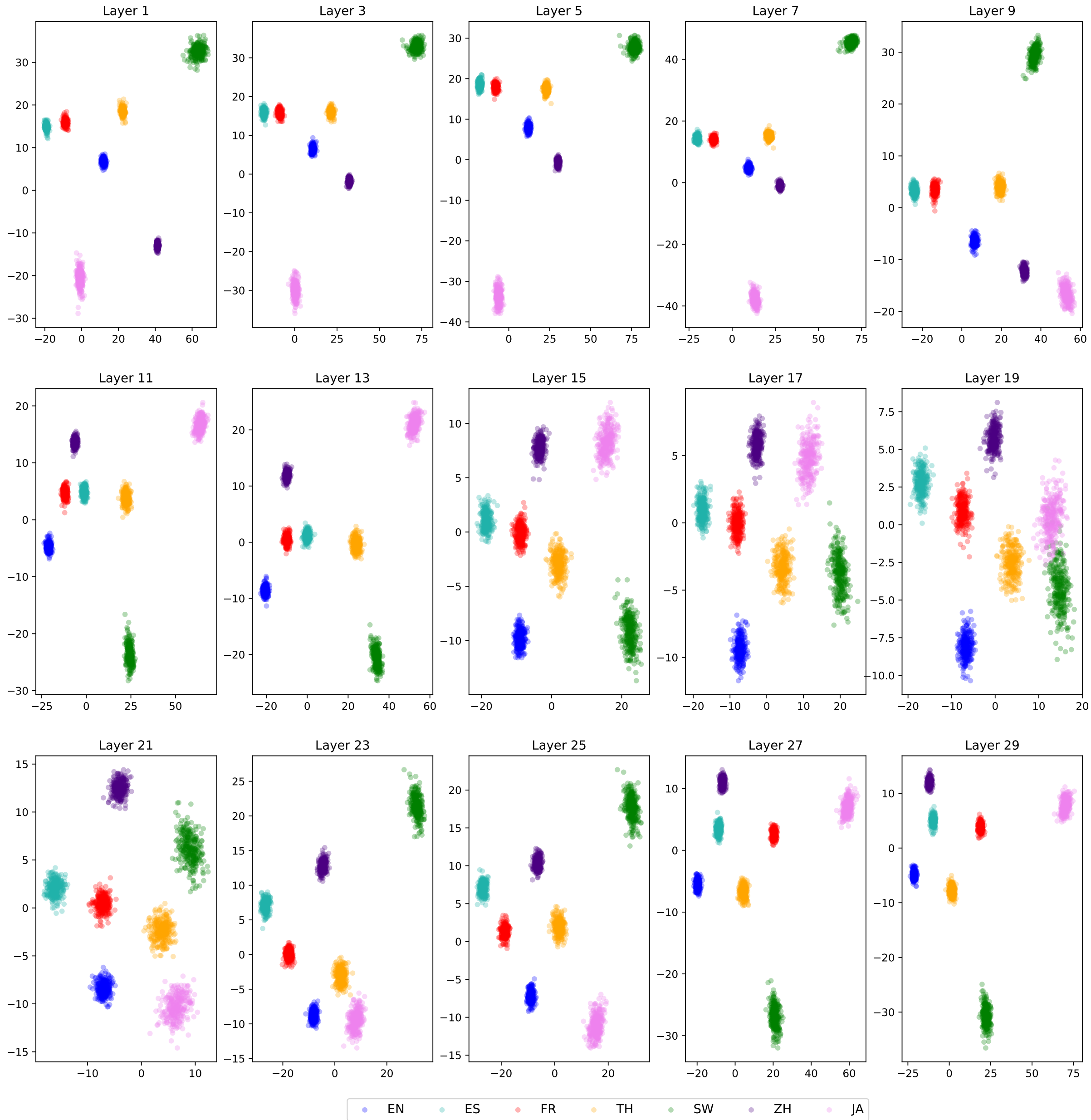

Figure 10: We follow Chang et al. (2022) to conduct LDA and present the visualization of sentence representations obtained by mean-pooling from $\text{BLOOM}_{7.1\text{B}}$ across layers along LDA components 1 and 3. We utilize 300 samples for each language from the FLORES dataset.

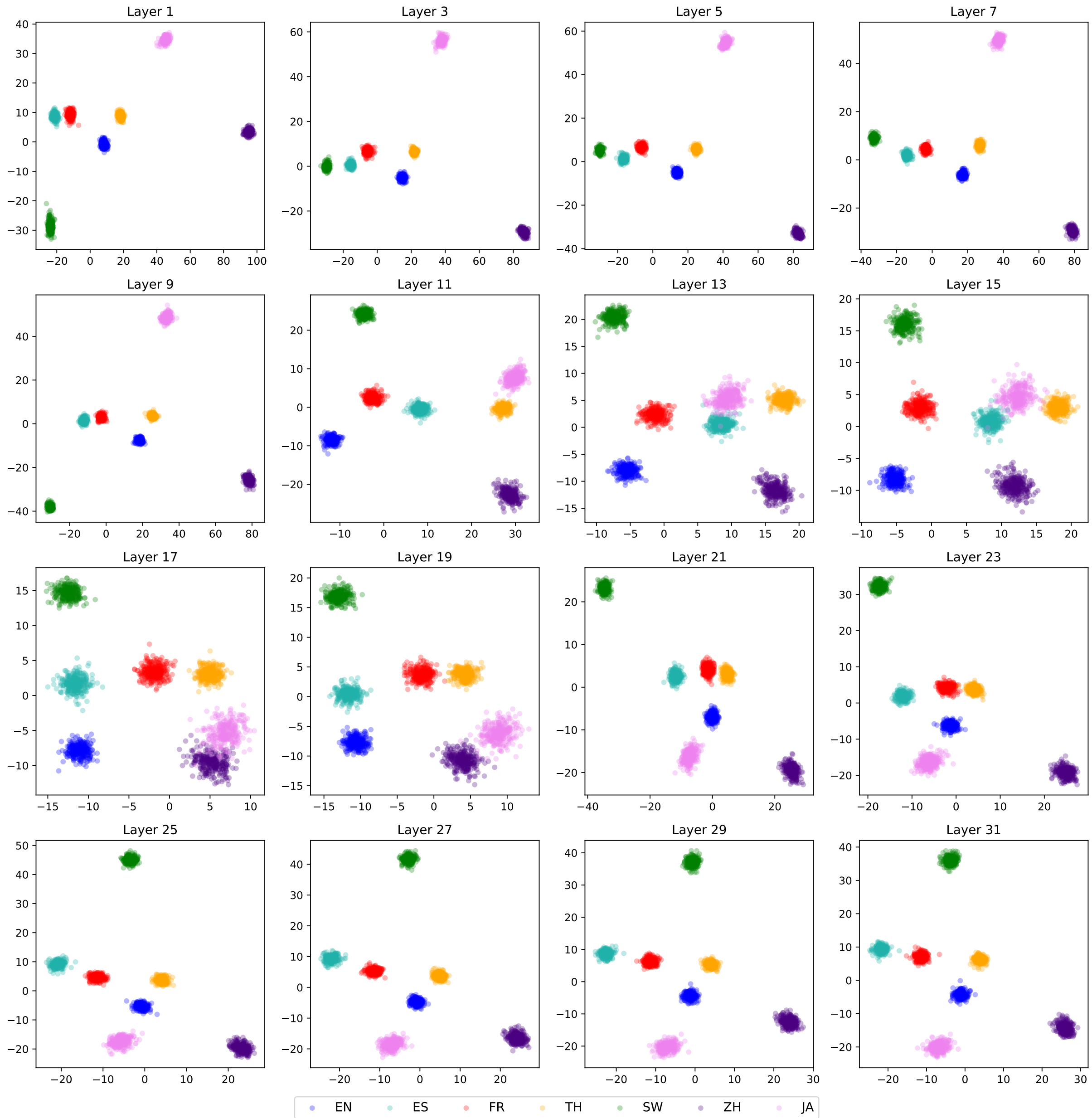

Figure 11: We follow Chang et al. (2022) to conduct LDA and present the visualization of sentence representations obtained by mean-pooling from $XGLM_{7.5B}$ across layers along LDA components 1 and 3. We utilize 300 samples for each language from the FLORES dataset.

A.2 Details of Language Subspace Distance

For each language A, we obtain a data matrix $\boldsymbol{X}_A \in \mathbb{R}^{n \times d}$ of n contextualized token representations with d dimensionality in language A using 1k FLORES samples per language from the desired layer.

The language subspace $\mathcal{S}_A$[10] is described by the language's mean representation $\boldsymbol{\mu}_A \in \mathbb{R}^d$ along with k principal directions of maximal variance in the language, defined by an orthonormal basis $\boldsymbol{V}_A \in \mathbb{R}^{d \times k_A}$.

In particular, $\boldsymbol{\mu}_A$ can be calculated as the mean value of X_A along the token dimension n. As for $\boldsymbol{V}_A$, we first perform a singular value decomposition (SVD) of $\boldsymbol{X}_A$: $\boldsymbol{X}_A = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ are orthogonal. $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times d}$ consists of a diagonal matrix $\boldsymbol{\Sigma}' \in \mathbb{R}^{d \times d}$ and a zero matrix, where $\boldsymbol{\Sigma}' = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$. $\boldsymbol{\Sigma}'$ denotes the direction of greatest change in $\boldsymbol{X}_A$, which can be used for feature selecting. We select the first k_A values to get $\boldsymbol{\Sigma}_A = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{k_A}) \in \mathbb{R}^{k_A \times k_A}$, while at the same time ensuring that the subspace accounted for 90% of the total variance in the language.[11] Therefore, based on $\boldsymbol{\Sigma}_A$, we can obtain the corresponding $\boldsymbol{V}_A$ and leverage $\boldsymbol{U}\boldsymbol{\Sigma}_A\boldsymbol{V}_A^T$ to estimate $\boldsymbol{X}_A$. Since $\boldsymbol{K}_A = \frac{1}{n-1}\boldsymbol{X}_A^{-1}\boldsymbol{X}_A$ (Chang et al., 2022), the $\boldsymbol{K}_A \in \mathbb{R}^{d \times d}$ can be calculated with $\frac{1}{n-1}\boldsymbol{V}_A\boldsymbol{\Sigma}_A^2\boldsymbol{V}_A^T$.

A.3 Impact of Different Pooling Methods

We also investigate the impact of three different pooling methods, namely mean-pooling, max-pooling, and last token representation, to derive sentence embeddings for our *ShifCon* framework.

	Llama-2$_{7B}$	XGLM$_{7.5B}$	BLOOM$_{7.1B}$
Mean-pooling	46.0	43.8	44.1
Max-pooling	45.2	43.3	43.6
Last token	45.8	44.1	43.7

Table 5: The average performance results of our $ShifCon$ framework across all benchmarks for the three different pooling methods.

As demonstrated in Table 5, the last token and mean pooling methods exhibit superior performance, and our approach shows less sensitivity to the choice of pooling method.

A.4 Details of Evaluation

Due to the extensive training time required to train all languages included in Bactrian-X, we opt to sample a subset of representative languages, covering both high and low-resource languages for training. During evaluation, we focus on assessing the performance of the selected languages with corresponding benchmarks. Detailed information regarding the languages used, evaluation metrics for each dataset are presented in Table 6. The evaluation prompt template are presented in Table 7.

Dataset	**|Lang.|**	**Languages**	**Metric**	**Data Type**
Bactrian-X	8	English, Chinese, Indonesian, Spanish, Swahili, Thai, Turkish, Hindi	-	Train
MGSM8KInstruct	10	English, Chinese, Spanish, French, German, Russian, Japanese, Swahili, Thai, Bengali	-	Train
MGSM	10	English, Chinese, Spanish, French, German, Russian, Japanese, Swahili, Thai, Bengali	Accuracy	Test
XNLI	7	English, Spanish, Chinese, Turkish, Thai, Hindi, Swahili	Accuracy	Test
XCOPA	5	Chinese, Indonesian, Turkish, Thai, Swahili	Accuracy	Test
XStoryCloze	6	English, Spanish, Chinese, Indonesian, Hindi, Swahili	Accuracy	Test
FLORES	6	Spanish, Chinese, Indonesian, Turkish, Thai, Swahili	ChrF++	Test

Table 6: Multilingual datasets used in our experiments. We utilize ChrF++ (Popović, 2017) metric to evaluate the translation performance.

[10]We follow Chang et al. (2022) to define the language subspace.

[11]Results were qualitatively similar for subspaces accounting for variance proportions in [75%, 90%, 95%, 99%].

Task	**Pattern**	**Verbalizer**
XNLI	`{premise}` Based on the previous passage, is it true that `{hypothesis}`? Yes, No, or Maybe? `{label}`	Yes ‖ Maybe ‖ No
XCOPA	`{premise}` {% if question == “cause" %}This happened because... {% else %} As a consequence...{% endif %} Help me pick the more plausible option: - `{choice1}` - `{choice2}` `{label}`	`{choice1}` ‖ `{choice2}`

XStoryCloze	`{input_sentence_1}` `{input_sentence_2}`
	`{input_sentence_3}` `{input_sentence_4}`
	What is a possible continuation for the story given the following options?
	- `{sentence_quiz_1}`
	- `{sentence_quiz_2}`
	`{label}`
		`{sentence_quiz_1}` ‖
		`{sentence_quiz_2}`

FLORES	Translate the following `{src_language}` text to `{tgt_language}`: `{src_sentence}` `{tgt_sentence}`	`{tgt_sentence}`

Table 7: The prompt templates used for evaluation following Muennighoff et al. (2023) and Zhang et al. (2024c).

A.5 Implementation Details

	Dimension	Heads	Layers
Llama-2_{7B}	4096	32	32
Llama-3_{8B}	4096	32	32
$BLOOM_{7.1B}$	4096	32	30
$BLOOM_{1.7B}$	2048	16	24
$BLOOM_{560M}$	1024	16	24
$XGLM_{7.5B}$	4096	32	32
$XGLM_{2.9B}$	2048	16	48
$XGLM_{564M}$	1024	16	24

Table 8: The detailed information of the models utilized in our experiment. “Dimension”, “Heads”, and “Layers” denote the dimension of representation, attention heads, and number of layers, respectively.

Model Information In Table 8, we provide comprehensive details about the models utilized in our experiment. Here, “Dimension”, “Heads”, and “Layers” represent the representation dimension, attention heads, and number of layers, respectively.

Training Settings Our experiments are conducted with 4xA100 GPUs. Each experiment is run with three different random seeds, and the results are averaged to obtain the final outcome. The temperature τ is set to 0.05 in the multilingual contrastive learning procedure. We follow previous multitasking works (Kong et al., 2022; Zhang et al., 2023a) to explore α values in Eq. 6 within [0.5, 1.0, 1.5, 2.0] to determine the best performance. We set the learning rate for training models with parameters exceeding 7 billion to 1e-5, while for others to 3e-5. We set the maximum sequence length to 512 and the global batch size to 128. In generation tasks, we utilize a greedy decoding strategy to help replicate our results accurately. A cosine scheduler with a 3% warm-up period is implemented. Mixed precision training and ZeRO are employed within the DeepSpeed training framework to accelerate the training process and conserve memory usage. The AdamW (Loshchilov and Hutter, 2019) optimizer is utilized to update the model parameters during the training process.

For the AFP baseline method, we adhere to the training configuration outlined by Li et al. (2024) to train the models. Specifically, we define p_{src} for cross-lingual guidance during training and perform multilingual contrastive learning on the first layer.

Additionally, we explore our *ShifCon* framework with a two-stage training strategy, which involves initial training solely with MSFT loss to establish a preliminary model, followed by further fine-tuning using our *shifCon* framework. As depicted in Table 9, the results indicate that implementing a two-stage training strategy leads to better performance. We posit that the preliminary model obtained by MSFT in the first stage could offer better representations for each language, facilitating shift projection and multilingual contrastive learning. Consequently, all results are reported based on the two-stage training strategy in our paper.

	Llama-2_{7B}	$XGLM_{7.5B}$	$BLOOM_{7.1B}$
MSFT	43.8	41.6	42.2
ShifCon w/ Two-Stage	46.0	43.8	44.1
ShifCon w/ One-Stage	44.8	41.7	42.5

Table 9: The average performance results of our $ShifCon$ framework across all benchmarks for the three model families, comparing the two-stage and one-stage training strategies.

A.6 New Strategy for Obtaining Better Language Vectors

Given that model parameters are updated at each training step, it is essential for the language vectors to be updated correspondingly. Inspired by the batch normalization paradigm, we introduce a novel strategy aimed at improving the quality of language vectors. As calculating the mean representation of all samples in language a after updating parameters for each batch is computationally expensive, we utilize the mean representation of language a samples in the t-th batch to estimate. Specifically, for the representations of language a in t-th batch at l-th layer, let $\boldsymbol{v}_t$ denote the mean representation of language a samples from first batch to t-th batch and $\boldsymbol{u}_t$ denote the mean representation of the samples in language a from the t-th batch (Noted that, $\boldsymbol{v}_t$ is computed by t-th step's model). The estimation of $\boldsymbol{v}_t$, i.e., $\hat{\boldsymbol{v}}_t$, can be obtained by using the representations of t-th batch computed by corresponding t-th step's model:

$$\hat{\boldsymbol{v}}_t = \frac{\sum_{i=1}^{t} \eta^{i-1} \boldsymbol{u}_i}{\sum_{i=1}^{t} \eta^{i-1}} \tag{7}$$

where $\eta \geq 1$ denotes the enhancement factor. η^{i-1} denotes the $i-1$-th power of η. As t increases, the model becomes more accurate, leading to more precise representation $\boldsymbol{u}_t$. Consequently, the corresponding weight factors are larger.

Subsequently, we can estimate the mean representation of next batch's $\boldsymbol{v}_t$ through the following approach:

$$\begin{aligned} \hat{\boldsymbol{v}}_{t+1} &= \frac{\sum_{i=1}^{t+1} \eta^{i-1} \boldsymbol{u}_i}{\sum_{i=1}^{t+1} \eta^{i-1}} \\ &= \frac{1}{\sum_{i=1}^{t+1} \eta^{i-1}} \eta^t \boldsymbol{u}_{t+1} + \frac{\sum_{i=1}^{t} \eta^{i-1}}{\sum_{i=1}^{t+1} \eta^{i-1}} \Big(\frac{1}{\sum_{i=1}^{t} \eta^{i-1}} \sum_{i=1}^{t} \eta^{i-1} \Big) \\ &= \frac{\eta^t}{\sum_{i=0}^{t} \eta^i} \boldsymbol{u}_{t+1} + \frac{\sum_{i=0}^{t-1} \eta^i}{\sum_{i=0}^{t} \eta^i} \hat{\boldsymbol{v}}_t \end{aligned} \tag{8}$$

Here, we only need the estimated mean representation $\hat{\boldsymbol{v}}_t$ and the true mean representation of the samples from the $t+1$ batch $\boldsymbol{u}_{t+1}$, to generate an estimation of the mean representation of $\hat{\boldsymbol{v}}_{t+1}$. For simplicity, we directly set $\frac{\eta^t}{\sum_{i=0}^{t} \eta^i} = \frac{1}{4}$ and $\frac{\sum_{i=0}^{t-1} \eta^i}{\sum_{i=0}^{t} \eta^i} = \frac{3}{4}$ in this work.

We conduct an extra ablation experiment on $\text{XGLM}_{564\text{M}}$ to verify the effectiveness of our proposed strategy. As shown in Table 10, when compared with the straightforward method (i.e., simply mean pooling the representations), our strategy can yield better performance.

	XCOPA		**XNLI**		**XStoryCloze**	
	High	**Low**	**High**	**Low**	**High**	**Low**
w/ New Strategy	58.4	55.8	42.6	40.5	59.8	58.1
w/ Mean Pooling	58.1	55.3	42.3	40.1	59.6	57.6

Table 10: The average performance of high- and low-resource languages across three classification tasks with two different language vector strategies.

A.7 Detailed Results of Each Language across All the Benchmarks

	High							Low		
	EN	ZH	DE	ES	FR	JA	RU	SW	BN	TH
Llama-2_{7B}	51.4	29.6	37.2	34.8	36.4	26.2	30.8	2.8	7.2	5.2
+MSFT	59.8	43.2	45.2	46.0	42.4	34.4	43.6	31.6	22.8	34.2
+AFP	60.0	42.8	46.4	47.2	45.6	37.2	45.2	34.4	25.2	35.6
+*ShifCon*	58.2	48.4	48.8	45.6	47.2	40.4	48.8	38.0	28.4	38.8
$XGLM_{7.5B}$	7.6	4.8	3.6	3.2	2.8	2.8	2.8	1.2	2.0	2.4
+MSFT	14.4	9.6	10.0	10.4	10.8	8.0	10.8	6.8	7.2	6.8
+AFP	16.4	9.2	12.4	13.2	12.4	11.2	10.4	9.6	9.2	10.0
+*ShifCon*	15.6	12.8	14.0	12.8	15.2	12.0	13.6	11.2	11.6	14.4

Table 11: The detailed results of each language on the MGSM task in Llama-2_{7B} and $XGLM_{7.5B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

	High				Low					
	EN	ZH	ES	FR	SW	BN	TH	DE	JA	RU
$BLOOM_{7.1B}$	20.0	9.2	11.6	12.0	2.4	5.2	1.6	4.0	2.4	6.8
+MSFT	26.8	18.8	21.6	20.4	11.6	13.2	10.4	13.6	12.4	14.0
+AFP	28.4	18.0	23.2	22.0	14.8	15.6	14.4	15.2	16.4	18.0
+*ShifCon*	28.0	21.2	24.8	24.0	19.2	18.8	17.6	19.6	18.4	19.4

Table 12: The detailed results of each language on the MGSM task in $BLOOM_{7.1B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

	High			Low		
	ES	ZH	ID	SW	TH	TR
Llama-2_{7B}	42.6	17.1	40.9	14.7	12.9	20.0
+MSFT	43.4	18.9	41.8	18.1	15.4	21.8
+AFP	43.9	19.5	42.4	18.9	16.2	22.4
+*ShifCon*	44.5	19.8	42.6	20.2	16.6	22.3
$XGLM_{7.5B}$	36.1	17.8	42.9	33.2	31.5	30.0
+MSFT	36.8	19.1	44.5	35.7	32.1	30.5
+AFP	37.4	19.8	45.0	35.9	32.9	31.2
+*ShifCon*	37.8	20.8	44.9	36.8	33.8	31.8
$BLOOM_{7.1B}$	40.2	35.2	48.8	37.1	16.2	19.6
+MSFT	40.5	36.0	50.5	37.8	17.6	22.3
+AFP	41.2	36.9	51.1	38.5	18.2	23.1
+*ShifCon*	41.6	36.8	51.7	39.2	18.4	23.9

Table 13: The detailed results of each language on the FLORES (en-xx) task in Llama-2_{7B}, $XGLM_{7.5B}$, and $BLOOM_{7.1B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

	High			Low		
	ES	ZH	ID	SW	TH	TR
Llama-2_{7B}	49.2	18.8	51.5	23.5	11.6	29.1
+MSFT	48.6	19.4	53.2	26.9	16.4	30.8
+AFP	49.0	19.7	54.4	27.5	17.1	31.6
+*ShifCon*	49.5	21.2	54.8	29.4	17.8	32.5
$XGLM_{7.5B}$	41.8	33.4	48.3	42.9	26.7	37.9
+MSFT	43.0	33.9	50.1	43.9	28.3	39.6
+AFP	44.0	34.8	51.2	44.5	28.9	39.9
+*ShifCon*	43.8	35.6	51.7	45.2	29.8	40.4
$BLOOM_{7.1B}$	45.8	39.6	51.6	43.8	20.3	28.1
+MSFT	46.4	39.9	53.3	45.4	23.0	30.8
+AFP	46.9	40.5	53.8	46.0	23.7	31.6
+*ShifCon*	47.6	41.3	52.8	46.8	24.5	32.4

Table 14: The detailed results of each language on the FLORES (xx-en) task in Llama-2_{7B}, $XGLM_{7.5B}$, and $BLOOM_{7.1B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

	High		Low		
	ZH	ID	TR	TH	SW
Llama-2_{7B}	63.8	62.6	49.0	51.4	48.8
+MSFT	65.0	63.4	51.8	52.6	51.5
+AFP	65.8	64.2	52.9	53.4	52.3
+*ShifCon*	66.8	64.2	54.1	53.2	53.2
$XGLM_{7.5B}$	63.6	64.0	56.8	57.1	58.2
+MSFT	64.4	65.4	58.4	58.8	57.6
+AFP	65.3	66.2	59.3	59.2	58.3
+*ShifCon*	66.8	66.8	60.2	59.4	60.6
$BLOOM_{7.1B}$	57.1	58.4	53.2	50.8	52.1
+MSFT	58.6	59.8	55.5	51.6	54.6
+AFP	59.4	60.5	56.3	52.8	55.5
+*ShifCon*	60.2	60.4	57.6	54.4	56.8

Table 15: The detailed results of each language on the XCOPA task in Llama-2_{7B}, $XGLM_{7.5B}$, and $BLOOM_{7.1B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

	High			Low			
	EN	ES	ZH	TR	TH	HI	SW
Llama-2_{7B}	49.1	42.6	44.0	35.8	37.2	37.1	30.8
+MSFT	50.8	43.8	44.5	37.5	39.5	38.8	34.6
+AFP	50.8	44.4	45.3	38.6	40.7	39.6	35.8
+*ShifCon*	50.4	45.0	46.1	40.8	41.8	40.2	38.1
$XGLM_{7.5B}$	46.9	41.6	45.0	39.8	43.2	42.6	40.1
+MSFT	48.7	42.4	46.7	41.3	44.4	42.2	41.2
+AFP	49.9	43.3	47.8	43.1	45.2	43.1	42.0
+*ShifCon*	51.2	45.8	48.9	44.7	44.8	43.8	43.8
$BLOOM_{7.1B}$	46.0	40.2	41.1	34.9	35.4	38.6	37.5
+MSFT	47.1	42.1	42.9	36.8	38.2	41.1	39.7
+AFP	47.9	43.4	43.6	37.9	39.3	41.8	40.6
+*ShifCon*	48.3	43.2	45.0	39.3	40.7	41.8	41.5

Table 16: The detailed results of each language on the XNLI task in Llama-2_{7B}, $XGLM_{7.5B}$, and $BLOOM_{7.1B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

	High				Low	
	EN	ES	ZH	ID	HI	SW
Llama-2_{7B}	84.4	75.5	69.4	69.4	57.9	55.3
+MSFT	85.5	76.9	70.5	68.3	59.6	57.8
+AFP	86.4	77.3	71.6	69.2	60.5	59.2
+*ShifCon*	86.2	77.5	72.8	70.1	60.2	61.5
$XGLM_{7.5B}$	73.5	63.7	60.4	63.2	59.5	57.2
+MSFT	74.4	65.8	62.8	64.0	61.2	59.1
+AFP	75.5	66.7	63.0	65.1	61.9	60.0
+*ShifCon*	75.2	67.4	62.4	67.2	62.8	61.5
$BLOOM_{7.1B}$	72.2	66.3	66.2	64.7	60.4	55.8
+MSFT	72.8	66.8	67.1	67.5	61.6	58.0
+AFP	72.2	67.5	67.9	67.2	61.9	58.5
+*ShifCon*	72.6	68.2	68.5	68.8	62.8	59.1

Table 17: The detailed results of each language on the XStoryCloze task in Llama-2_{7B}, $XGLM_{7.5B}$, and $BLOOM_{7.1B}$. High- and low-resource languages are categorized based on their data ratios in the pre-training corpus.

A.8 Low Subspace Distance Areas of Models across Different Families and Scales

	Low Subspace Distance Area	Layers
Llama-2_{7B}	[11, 20]	32
Llama-3_{8B}	[11, 20]	32
$BLOOM_{7.1B}$	[14, 22]	30
$BLOOM_{1.7B}$	[10, 17]	24
$BLOOM_{560M}$	[10, 17]	24
$XGLM_{7.5B}$	[13, 22]	32
$XGLM_{2.9B}$	[9, 23]	48
$XGLM_{564M}$	[9, 16]	24

Table 18: The low subspace distance areas of models in our experiments.

A.9 Language Code

ISO 639-1	Language	Family
BN	Bengali	Indo-European
DE	German	Indo-European
EN	English	Indo-European
ES	Spanish	Indo-European
FR	French	Indo-European
HI	Hindi	Indo-European
ID	Indonesian	Austronesian
JA	Japanese	Japonic
RU	Russian	Indo-European
ZH	Chinese	Sino-Tibetan
TH	Thai	Kra-Dai
SW	Swahili	Niger-Congo
TR	Turkish	Turkic

Table 19: Details of Language codes in this work.