LENS: Rethinking Multilingual Enhancement for Large Language Models

Anonymous ACL submission

Abstract

As global demand for multilingual large language models (LLMs) grows, most LLMs still remain overly focused on English, leading to the limited access to advanced AI for non-English speakers. Current methods to enhance multilingual capabilities largely rely on data-driven post-training techniques, such as multilingual instruction tuning or continual pre-training. However, these approaches exhibit significant limitations, including high resource cost, exacerbation of off-target issue and catastrophic forgetting of central language abilities. To this end, we propose LENS, a novel approach that enhances multilingual capabilities by leveraging LLMs' internal language representation spaces. LENS operates on two subspaces: the language-agnostic subspace, where it aligns target languages with the central language to inherit strong semantic representations, and the language-specific subspace, where it separates target and central languages to preserve linguistic specificity. Experiments on three English-centric LLMs show that LENS significantly improves multilingual performance while maintaining the model's English proficiency, achieving better results with less computational cost compared to existing post-training approaches.[1]

1 Introduction

In an increasingly interconnected world, large language models (LLMs) are expected to cater to a diverse range of users across various linguistic backgrounds (Ouyang et al., 2023; Zhao et al., 2024a; Zheng et al., 2024; Qin et al., 2024). However, most state-of-the-art LLMs remain heavily English-centric (Brown et al., 2020; Touvron et al., 2023a,b; Jiang et al., 2023; AI@Meta, 2024), performing far better in English than other languages, thereby potentially marginalizing large portions of the global population from accessing advanced AI services (Wang et al., 2024a; Zhu et al., 2024b).

This disparity has directly spurred research efforts to enhance multilingual capabilities of LLMs. Current approaches are predominantly based on data-driven post-training paradigm, such as multilingual instruction tuning (Zhang et al., 2023; Zhu et al., 2023; Üstün et al., 2024) or continual pre-training (Cui et al., 2023; Kuulmets et al., 2024; Jaavid et al., 2024), which primarily seeks to elicit cross-lingual alignment (Schuster et al., 2019) or inject multilingual knowledge with the supervision signals from *external* datasets.

While this paradigm is widely embraced and demonstrates certain successes, it faces several significant limitations. (1) The efficacy often depends on training with large-scale multilingual datasets (Cui et al., 2023; Zhu et al., 2023), which incur large computational overhead. (2) It overly emphasizes alignment across languages, neglecting the modeling of language-specific features, which exacerbates the off-target issue (Zhang et al., 2020). As a result, the model often struggles to generate accurate responses in the intended language when prompted (Lai et al., 2024; Sennrich et al., 2024). (3) The model's performance in languages it previously handled well is risking at catastrophic forgetting (McCloskey and Cohen, 1989).

In this work, we seek to provide a new perspective on addressing the aforementioned limitations by exploring and manipulating the *internal* representation within the language-related latent spaces of LLMs (Zou et al., 2023; Park et al., 2024). Taking the enhancement of multilingual capabilities for English-centric LLMs as an example. This is based on the intuitive idea that the well-established English representations in existing English-centric LLMs can act as a pivot to improve the performance of other languages. More specifically, for the target language to be enhanced, its *language-agnostic* semantic representations should be *pulled*

[1] Our code and data can be found in supplementary files.

close to those of English, facilitating cross-lingual alignment. Meanwhile, the *language-specific* linguistic representations should be *pushed away* from English to preserve the unique representation of each language. Also, during this process, it is crucial to ensure that the English pivot representation remains unchanged to effectively prevent catastrophic forgetting.

To achieve this, we propose LENS, a novel multi**L**ingual **E**nhancement method based on the hidden represe**N**tations within language **S**pace of LLMs. To be more specific, LENS comprises two stages: Language Subspace Probing (LSP) and Language Subspace Manipulation (LSM). During LSP, the multilingual hidden space within a single layer of the backbone are decoupled into two orthogonal components, a *language-agnostic* subspace and a *language-specific* subspace, via singular value decomposition. Then in LSM, we align the parallel multilingual input representations of the target language and the central language in the language-agnostic subspace. Simultaneously, the projection components of the target language within the language-specific space are pushed away from those of the central language, guiding the target language toward its distinct linguistic expression and ensuring the target language is properly expressed thereby mitigating the off-target issue. Finally, we align the central language's current representations with its original ones to preserve its proficiency during multilingual enhancement. It is crucial to note that, building on recent findings that language-related parameters are primarily concentrated in the top layers of LLMs (Wendler et al., 2024), LENS only updates the higher layers of the backbone with just a few hundred data points, exhibiting high resource efficiency.

We conduct extensive experiments on bilingual and multilingual enhancement setups, targeting languages from diverse linguistic families and varying resource levels. Results on three English-centric LLMs (LLaMA-3-8B-Instruct, LLaMA-3.1-8B-Instruct and Phi-3.5-mini-Instruct) show that LENS outperform baselines and open-source multilingual-enhanced LLMs on both multilingual comprehension and generation tasks. Notably, LENS achieves these improvements without compromising the central language's strong capabilities, while requiring significantly less computational overhead, highlighting its effectiveness, efficiency, and scalability.

The main contributions of this work are summarized as follows: (1) We provide a novel perspective for the multilingual enhancement of LLMs with their internal language representation space leveraged. (2) We propose LENS, an efficient and effective multilingual enhancement method that operates within the language representation space of large language models. (3) Extensive experiments demonstrate the effectiveness, efficiency, scalability of our method to obtain multilingual enhanced chat-style backbones without sacrificing original central language performance.

2 Methodology

2.1 Overview of LENS

We propose LENS, a novel method for effective and efficient multilingual enhancement of LLMs based on their internal language representation spaces. The overall diagram of LENS is displayed in Figure 1, consisting of two key stages: (1) Language Subspace Probing (LSP) and (2) Language Subspace Manipulation (LSM). The subsequent section offers a detailed introduction to them.

2.2 Language Subspace Probing

In this section, we first introduce our method to decouple and probe the language-agnostic and language-specific subspace within a single model layer in an unsupervised manner.

Assuming we aim to enhance the multilingual capabilities of a backbone model for L languages, which include one central language and $L-1$ target languages to be enhanced. In each layer of the backbone, we can obtain a mean representation for each language l:

$$\boldsymbol{m}_l = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{e}_l^i \tag{1}$$

where $\boldsymbol{e}_l^i \in \mathbb{R}^d$ is the embedding of the last token for the i-th sample in language l, and n is the total number of samples for each language. Concatenating $\boldsymbol{m}_l$ of L languages column-by-column results in the mean embedding matrix $\boldsymbol{M} \in \mathbb{R}^{d\times L}$ specifying the multilingual latent space.

Follow previous works (Pires et al., 2019; Libovický et al., 2020; Yang et al., 2021), we hypothesize that such multilingual latent space M could be decomposed into two orthogonal components (1) a language-agnostic subspace $\boldsymbol{M}_a$ representing what is commonly shared across languages and (2) a language-specific one $\boldsymbol{M}_s$ specifying on which different languages express different linguistic signals. Following Piratla et al. (2020); Xie et al.

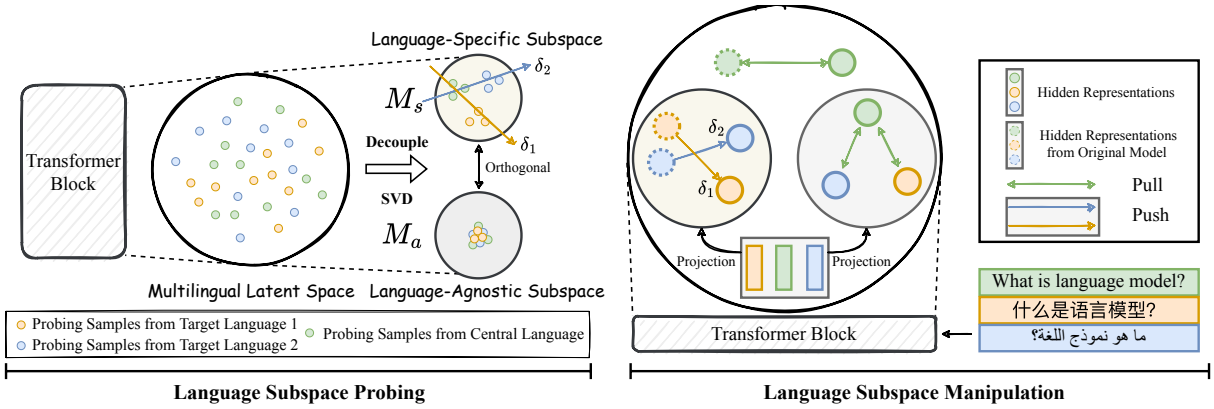


Figure 1: The overall architecture of our proposed LENS for multilingual enhancement. (1) In the LSP, we begin by decomposing the multilingual latent space, which is formed by the representations of probing samples from both the target and central languages. Using singular value decomposition (SVD), we separate this space into two orthogonal components: a language-agnostic subspace, $\boldsymbol{M}_a$, and a language-specific subspace, $\boldsymbol{M}_s$. (2) Then in LSM, the parallel multilingual representations of the target languages are pushed toward their respective linguistic expression directions within $\boldsymbol{M}_s$, while being pulled closer to the central language in $\boldsymbol{M}_a$. Additionally, the representations of the central language are carefully constrained to remain largely intact.

(2022), the objective can be formulated as:

$$\min_{\boldsymbol{M}_a,\boldsymbol{M}_s,\boldsymbol{\Gamma}} \quad \left\|\boldsymbol{M}-\boldsymbol{M}_a\mathbb{1}^\top-\boldsymbol{M}_s\boldsymbol{\Gamma}^\top\right\|_F^2 \qquad (2)$$
$$\text{s.t.} \quad \text{Span}\left(\boldsymbol{M}_a\right)\perp\text{Span}\left(\boldsymbol{M}_s\right),$$

where $\boldsymbol{M}_a \in \mathbb{R}^{d\times 1}$, $\boldsymbol{M}_s \in \mathbb{R}^{d\times r}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{L\times r}$ is the coordinates of language-specific signals along the subspace's r components. And a lower dimensionality for $\boldsymbol{M}_a$ is reasonable because the semantic consistency across different languages can be captured in a simpler form. Meanwhile, $\boldsymbol{M}_s$ requires a higher dimensionality to account for the distinct features of each language.

The optimal solution of Equation 2 can be computed efficiently via Singular Value Decomposition (SVD), where Algorithm 1 in Appendix A presents the detailed procedure.

After obtaining the language-specific subspace $\boldsymbol{M}_s$, we aim to identify a direction of language expression within this subspace, which points from the projection of mean representation from target language $\boldsymbol{m}_l$ to that from central language $\boldsymbol{m}_c$. Formally, the linguistic language expression direction $\boldsymbol{\delta}_l \in \mathbb{R}^d$ for each target language l is calculated as:

$$\boldsymbol{\delta}_l = \boldsymbol{M}_s^T\boldsymbol{M}_s(\boldsymbol{m}_l-\boldsymbol{m}_c) \qquad (3)$$

2.3 Language Subspace Manipulation

To eliminate the heavy reliance on hard-to-access high-quality multilingual datasets, we leverage the well-trained hidden representations of the central language in LLMs as a pivot to derive supervision signals for multilingual enhancement within the model's internal language space.

First, we propose to pull parallel multilingual representations closer within the shared language-agnostic subspace $\boldsymbol{M}_a$. This allows us to directly inherit the well-established general capabilities of the central language. Formally, this goal is accomplished by projecting multilingual representations (at the position of the last token) onto the subspace $\boldsymbol{M}_a$, with the optimization objective defined as:

$$\mathcal{L}_1 = \left\|\boldsymbol{M}_a^T\boldsymbol{M}_a(\boldsymbol{x}_l-\boldsymbol{x}_c)\right\|^2 \qquad (4)$$

where $\boldsymbol{x}_l$ and $\boldsymbol{x}_c$ are parallel multilingual representations from target language l and central one.

Second, to ensure that each target language can be accurately expressed and to alleviate the off-target issue, we need to push the multilingual representations in the language-specific subspace $\boldsymbol{M}_s$ towards their respective language-specific expression directions. This can be achieved through the projection onto the subspace $\boldsymbol{M}_s$ and optimizing the following objective:

$$\mathcal{L}_2 = \left\|\boldsymbol{M}_s^T\boldsymbol{M}_s(\boldsymbol{x}_l-\boldsymbol{x}_l^{\text{ref}})-\lambda_l\boldsymbol{\delta}_l\right\|^2 \qquad (5)$$

where $\boldsymbol{x}_l^{\text{ref}}$ is the representation of target language l obtained from original reference model and λ_l is a scalar of push strength for the corresponding language. The above process can be interpreted as directing the language-specific representations of each target language to shift a specific distance from their original positions toward a direction that enables accurate expression.

Finally, to ensure that the capabilities of the central language are not compromised and maintain a stable alignment objective for the target language, we constrain the representations of central language to remain predominantly intact:

$$\mathcal{L}_3 = \left\| \boldsymbol{x}_c - \boldsymbol{x}_c^{\text{ref}} \right\|^2 \tag{6}$$

where $\boldsymbol{x}_c^{\text{ref}}$ is the representation of central language c obtained from original reference model.

The final optimization objective of LENS is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 \tag{7}$$

where λ_1 and λ_3 are hyper-parameters to balance the impact of these two losses.

3 Experiments

3.1 Experimental Setup

Models We select three English-centric LLMs: LLaMA-3-8B-Instruct (AI@Meta, 2024), LLaMA-3.1-8B-Instruct (Dubey et al., 2024) and Phi-3.5-mini-instruct (Abdin et al., 2024), to fully validate the efficacy and scaleability of our LENS. Notably, although the latter two models undergo more extensive multilingual alignment training, English continues to exhibit the strongest performance.

Languages to be Enhanced We conduct experiments in both bilingual and multilingual settings to address various multilingual enhancement needs.

In the bilingual setting, English (En) serves as the central language, while Chinese (Zh) is the target language for expansion. Chinese is selected due to its growing prominence in the academic focus on multilingual enhancement for LLMs.

In the multilingual setting, we select six target languages, reflecting diverse linguistic families and resource levels. The high-resource languages are Chinese (Zh) and Japanese (Jp); the medium-resource languages are Korean (Ko) and Arabic (Ar); and the low-resource languages are Bengali (Bn) and Swahili (Sw), with English (En) continuing to serve as the central language.

It is important to note that these target languages are classified as *out-of-scope* in the official model card of the above LLMs, which further underscores their relevance for enhancement.

Training Data We sample 300 data points from Aya Dataset (Üstün et al., 2024) for each language to probe the language space and 200 data points from the Bactrian-X dataset (Li et al., 2023a) per language to manipulate the language space. Please see Appendix B for more details of these datasets.

Benchmarks To comprehensively measure the efficacy of our LENS on various multilingual tasks, we employ 5 mainstream benchmarks for evaluation, which can be categorized into multilingual understanding and multilingual generation:

Multilingual Understanding: (1) **XCOPA** (Ponti et al., 2020), (2) **XWinograd** (Muennighoff et al., 2023), (3) **XStoryCloze** (Lin et al., 2022) and (4) **M-MMLU** (Hendrycks et al., 2021; Lai et al., 2023). **Accuracy** is the evaluation metric and we randomly sample up to 1,000 data points from each benchmark for evaluation.

Multilingual Generation: (5) **MT-Bench** (Zheng et al., 2023): A benchmark for open-ended generation to evaluate a model's ability to follow multi-turn instructions. The evaluation follows the **LLM-as-a-judge** approach, where GPT-4o is prompted to assign a score directly to a single response on a scale of 1 to 10. It is essential to highlight that the languages targeted for enhancement, as mentioned above, are all within the capability range of GPT-4o, especially given that its official model card (OpenAI, 2024) emphasizes support for low-resource languages such as Swahili (Sw) and Bengali (Bn). This underscores the validity and reliability of the evaluation approach. We employ **Language Fidelity** (Holtermann et al., 2024) as a metric to assess the consistency between input and output languages, offering a clear measure of how effectively off-target issues are mitigated.

Please refer to Appendix C for the detailed description of the benchmarks.

3.2 Baseline Methods

For comparison, we consider the following baseline methods that enhance LLMs' multilingual capabilities using multilingual instruction fine-tuning technique: (1) **xSFT & xSFT-Full** (Ouyang et al., 2022), (2) **QAlign** (Zhu et al., 2024a), (3) **SDRRL** (Zhang et al., 2024c) and (4) **CLA** (Li et al., 2024).

Please refer to Appendix D.1 for the detailed description of the baseline methods.

3.3 Implementation Details

Our experiments are implemented with PyTorch (Paszke et al., 2019) and Transformer library (Wolf et al., 2020) on a single NVIDIA A800-SXM4-80GB GPU. The training duration is set to one epoch with the learning rate of 1e-5 and batch size of 8 across all backbones. For more detailed settings, please refer to the Appendix E.

	Multilingual Understanding										Multilingual Generation	
	XCOPA		XWinograd		XStoryCloze		M-MMLU		AVG.		MT-Bench	
	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh
LLaMA-3	-	83.40	63.50	54.37	95.40	88.90	64.90	49.40	74.60	69.02	6.99 (100%)	2.72 (43.75%)
xSFT	-	87.20	**64.30**	63.49	**95.10**	90.60	62.80	46.10	74.07	71.85	4.79 (100%)	2.94 (88.75%)
xSFT-Full	-	84.60	58.80	60.11	93.50	90.30	60.60	43.20	70.97	69.55	5.80 (100%)	4.44 (92.50%)
QAlign	-	52.20	55.10	47.02	89.20	71.90	56.40	34.00	66.90	51.28	3.59 (100%)	1.23 (37.50%)
SDRRL	-	85.20	64.80	55.95	92.60	84.30	63.80	47.80	73.73	68.31	6.60 (100%)	3.84 (73.75%)
CLA	-	85.60	61.70	56.70	95.00	89.80	**64.70**	**48.90**	73.80	70.26	6.47 (100%)	4.41 (81.25%)
LENS (Ours)	-	**87.60**	63.80	**66.67**	94.70	**91.80**	64.40	48.60	**74.30**	**73.67**	**7.21** (100%)	**5.77** (**97.50%**)
LLaMA-3.1	-	90.40	64.10	68.65	95.80	91.40	69.30	52.50	76.40	75.74	7.31 (100%)	5.38 (93.75%)
xSFT	-	88.00	63.70	67.46	96.20	**92.70**	68.10	**53.10**	76.00	75.32	5.33 (100%)	3.32 (90.00%)
xSFT-Full	-	86.80	60.40	62.50	90.60	83.80	66.10	49.90	72.37	70.75	6.02 (100%)	4.18 (92.50%)
QAlign	-	55.00	56.00	48.02	94.10	52.30	64.10	33.50	71.40	47.20	4.13 (100%)	2.65 (83.75%)
SDRRL	-	87.20	63.20	58.83	95.30	89.80	63.50	45.30	74.00	70.31	6.49 (100%)	3.14 (58.75%)
CLA	-	89.00	63.30	65.90	**94.20**	90.50	**67.40**	52.50	**77.20**	75.36	6.49 (100%)	4.49 (88.75 %)
LENS (Ours)	-	**90.20**	64.60	**69.44**	95.90	91.80	69.10	52.60	76.53	**76.01**	**7.41** (100%)	**5.96** (**93.75%**)

Table 1: Detailed results on the multilingual understanding and multilingual generation benchmarks under the bilingual setting (English and Chinese). The values in parentheses represent language fidelity. Results highlighted in green indicate an improvement or performance of central language comparable (within a gap of 0.5) to the original backbone, while those highlighted in red signal a decline in performance relative to the original backbone. The best and second-best results in our method and baselines are in bold and underlined, respectively.

3.4 Overall Results

Table 1 present the performance comparison between LENS and recent multilingual enhancement baseline methods under bilingual settings. For additional results under multilingual configuration across all three backbones, please see Appendix F.1. From the results across all backbones, we have drawn the following key insights:

LENS achieves comprehensive improvements across diverse languages with different resource levels. It outperforms baselines on both multilingual understanding and generation benchmarks, effectively mitigating the off-target issue. In contrast, baselines primarily favor multilingual understanding, offering minimal or no improvement in generation tasks. In some cases, they compromise language fidelity, exacerbating the off-target problem. Finally, LENS safeguards the central language from catastrophic forgetting. We also compare LENS with LoRA-based SFT, as recent work suggests that it can effectively prevent catastrophic forgetting (Huang et al., 2024). Detailed results and discussions are provided in Appendix F.2.

Using the central language representations within the backbone as a supervision signal proves more effective, efficient and scalable. The key distinction between LENS and baseline methods lies in how multilingual performance is enhanced: LENS relies on the model's internal representation of the central language, while baseline methods depend on external data. This difference make baselines not only fail to improve the target languages but also lead to performance degradation. However, the Aya Dataset and Bactrain-X datasets we used are already considered high-quality multilingual resources, widely employed and proven effective in boosting multilingual capabilities in previous models such as mT5 and LLaMA-2 (Li et al., 2023a; Üstün et al., 2024). This highlights that for current extensively trained LLMs such as LLaMA-3 (which has been trained on over 15T data), an over-reliance on external supervision signals may fall short of scalability needs (Cao et al., 2024). In addition, the results in Table 6 of Appendix F demonstrate that LENS achieves the best performance with minimal computational overhead.

4 Analysis

4.1 Ablation Study

We conduct ablation studies to validate the effectiveness of the three optimization objectives in LSM. The results under the bilingual enhancement setting with LLaMA-3-8B-Instruct are shown in Figure 2, leading to the following key findings:

Aligning multiple languages in language-agnostic subspaces primarily enhances multilingual comprehension rather than generation. As the coefficient λ_1 of the alignment loss $\mathcal{L}_1$ increases, Chinese comprehension improves, while its generation ability remains slightly unaffected.

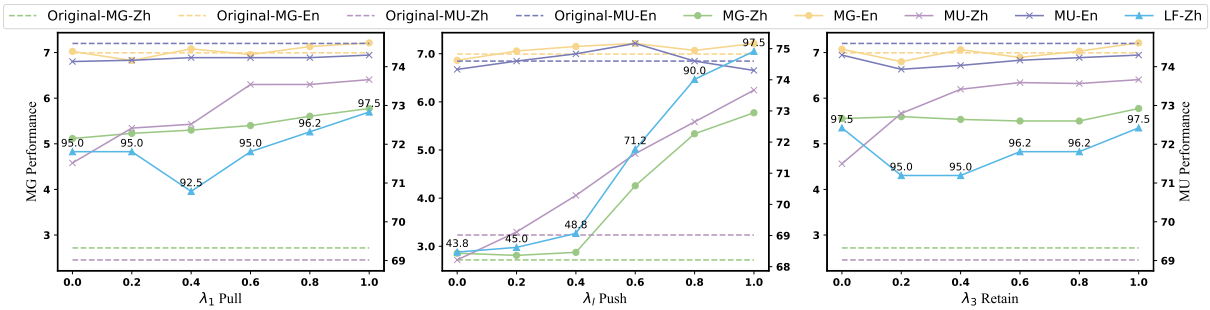


Figure 2: The ablation results to verify the effectiveness and impact of different optimization objectives in LSM. MU Performance stands for the average performance on all multilingual understanding benchmarks, while MG Performance is the results on MT-Bench. LF represents language fidelity.

	Zh	Jp	Ar	Ko	Bn	Sw
COMET scores on X-to-English Tasks						
LLaMA-3	85.40	86.15	84.77	86.07	85.51	78.15
xSFT	70.41	72.40	67.09	72.43	59.52	73.56
QAlign	85.52	85.26	83.11	84.96	83.13	73.66
SDRRL	44.78	45.73	40.87	45.29	45.05	41.51
CLA	85.28	85.35	85.11	85.15	84.36	77.84
LENS	**85.64**	**86.23**	**85.15**	**86.07**	**85.67**	**80.05**
COMET scores on English-to-X Tasks						
LLaMA-3	85.28	88.32	76.51	84.53	80.14	71.44
xSFT	83.78	82.22	74.30	81.08	73.40	58.48
QAlign	61.65	58.66	49.41	57.16	41.10	50.96
SDRRL	62.52	57.65	43.83	64.11	68.74	60.00
CLA	80.19	85.75	72.50	82.45	56.73	55.36
LENS	**85.59**	**88.47**	**79.52**	**85.77**	**80.20**	**71.88**

Table 2: Results of baseline methods and our LENS on FLORES-101 benchmark.

Separating language representations in language-specific subspaces is crucial for multilingual performance. As shown in the middle of Figure 2, increasing λ_l significantly boosts both comprehension and generation in Chinese, yielding greater benefits than mere language alignment. This suggests that the conventional approach of aligning languages (Cao et al., 2020; Zhu et al., 2023; Li et al., 2024; Hua et al., 2024) may be insufficient for fully optimizing multilingual LLMs. Further discussion in Appendix F.3 elaborates on this insight.

To further illustrate the benefit of the distinction between representations of different languages within the language-specific subspace, we conduct additional experiments on the multilingual machine translation task and results are shown in Table 2. The consistent performance improvements achieved by LENS confirm that it captures the subtle linguistic nuances required for precise and fluent translations. Detailed experimental setups and analysis are provided in Appendix F.4.

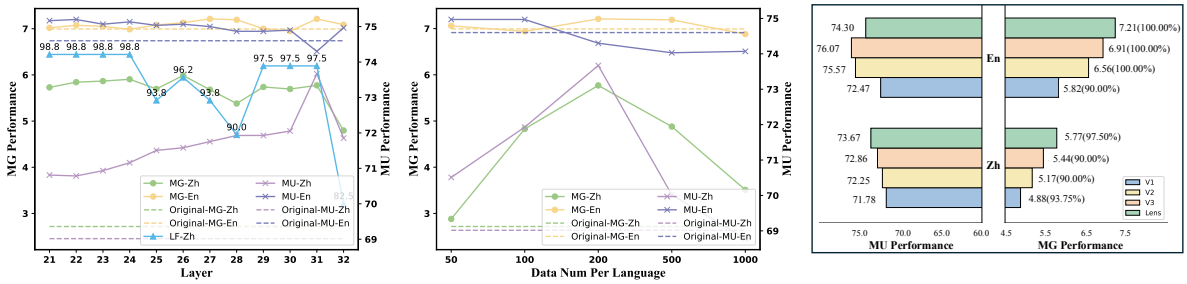
Maintaining stable representations of the central language provides reliable alignment supervision. As shown in the rightmost part of Figure 2, removing the retention objective for English leads to a significant drop in Chinese performance, likely due to misalignment caused by changes in English representations. However, since our modifications only affect the upper layers (layers 31 and 32) of LLaMA-3-8B-Instruct, most parameters remain frozen, preventing catastrophic forgetting of English capabilities. Further analysis in §4.2 confirms that increasing the number of updated layers does not negatively impact English performance, highlighting the efficacy of the retention objective.

4.2 Impact of Varying the Number of Manipulated Layers and Training Data

Building on recent studies on LLM interpretability (Zhao et al., 2024c; Zhong et al., 2024), we focus on updating only the upper layers of the backbone. Figure 3(a) examines the impact of modifying different layers under the bilingual enhancement setting with LLaMA-3-8B-Instruct. The horizontal axis denotes the starting layer for updates, with the final layer as the default endpoint.

Intermediate-layer processing in English benefits multilingual understanding. If the target language is partitioned into the language-specific subspace too early, it weakens multilingual comprehension. However, modifying only the final layer is insufficient for enhancing either understanding or generation, as language-specific information remains underdeveloped, leading to off-target errors.

LENS further validates the conclusions of existing works on LLM interpretability and applies these findings to multilingual enhancement. Please



(a) The impact of manipulating different backbone layers on multilingual performance enhancement.

(b) The impact of varying the number of training data per language.

(c) Comparison between the bilingual enhanced model via our LENS with Chinese-LLaMA series.

Figure 3: (a) The impact of varying the number of manipulated layers. (b) The impact of training data volume. (c) Comparison with open-source multilingual-enhanced LLMs.

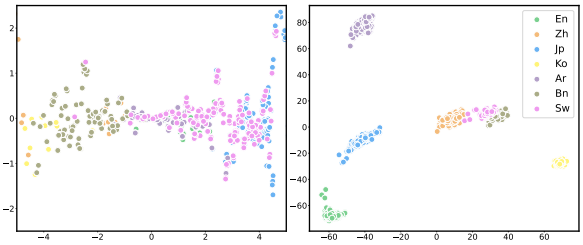


Figure 4: The PCA visualization of multilingual representations projected in the obtained language-agnostic subspace (right) and the language-specific (left) subspace. The backbone model is LLaMA-3-8B-Instruct after multilingual enhanced with LENS.

refer to Appendix F.5 for more analysis.

Results in Figure 3(b) indicate that increasing the amount of training data leads to diminishing returns for LENS. This finding reinforces our claim that for extensively pre-trained LLMs such as LLaMA-3 (trained on over 15T tokens), over-reliance on more training data falls short of meeting scalability needs. Instead of focusing on larger training datasets, it is more critical to identify supervision signals that are both reliable and scalable. This directly motivates us to seek internal supervision from the central language with the backbone itself.

4.3 Comparison with Open-Source Multilingual-Enhanced LLMs

In §3.4, we compare with *reproducible* baselines. Additionally, we extend our comparisons to open-source LLMs that leverage private datasets and large-scale post-training to improve multilingual performance. In particular, we focus on the Chinese-LLaMA-3 series, which builds on LLaMA-3 series to enhance Chinese capabilities and includes three versions. For a detailed description of these models, please refer to Appendix D.2.

Results and the resource consumption of different methods are presented in Figure 3(c) and Table 6 in Appendix F, respectively. The resulting model applied with LENS is identical to the one utilized for bilingual enhancement in Table 1. LENS demonstrates more comprehensive enhancement of the Chinese capabilities with extremely low resource overhead compared to these three models. Also, all the data leveraged by LENS is publicly accessible, which eliminates the need for laboriously gathering extensive high-quality multilingual datasets and makes it easily shareable with the community.

4.4 Visualization Analysis

To further confirm whether LENS manipulates language representations within different language subspaces as anticipated, as shown in Figure 4, we perform Principal Component Analysis (PCA) to visualize the projection of multilingual representations in the language-agnostic and -specific subspace. Parallel inputs in seven languages are sourced from the MultiQ datasets (Holtermann et al., 2024). Results indicate that representations of different languages converge within a narrow range in the language-agnostic subspace, while forming distinct clusters in the language-specific subspace, supporting our claim. This also highlights the advantages of LENS in delivering transparent, controllable, and interpretable solutions for the multilingual enhancements of LLMs.

4.5 Impact of Language Proximity

In §3.4, we focus on languages that are out-of-scope and under-represented in the backbone, selected based on their typological and script differences from the central language, English. To

further highlight the generalizability of LENS, we expand our analysis to include three additional languages: Spanish, French, and German. These languages, being typologically closer to English, are already better represented in the backbone model. As illustrated in Table 7 of Appendix F.6, LENS still exhibit notable improvements. For a more comprehensive discussion, please refer to Appendix F.6.

5 Related Works

Multilingual Large Language Model Pretraining on multilingual data is a common approach to gain multilingual capabilities (Conneau and Lample, 2019; Xue et al., 2020; Lin et al., 2022; Shliazhko et al., 2022; Wei et al., 2023; Xue et al., 2022; Le Scao et al., 2023; Blevins et al., 2024). However, due to the uneven distribution of data in pretraining corpora, current LLMs or MLLMs exhibit uneven language capabilities, with most models heavily biased towards English (Jiang et al., 2023; AI@Meta, 2024; Abdin et al., 2024). Moreover, pretraining from scratch is computationally intensive. These limitations have directly sparked research into expanding or enhancing the language capabilities of current LLMs or MLLMs.

Multilingual Enhancement for LLMs Current methods for multilingual enhancement of LLMs can be categorized into two types: 1) prompt-based methods and 2) post-training-based methods.

The former focuses on leveraging the LLMs' own translation capabilities to translate low-resource language inputs into the central language, and then generating a response (Shi et al., 2023; Huang et al., 2023; Qin et al., 2023; Etxaniz et al., 2024; Zhang et al., 2024b). However, Liu et al. (2024a) reveal the limitations of these methods, showing they are not optimal for real-world scenarios and highlighting the necessity of more comprehensive multilingual enhancement.

The latter aims to conduct further multilingual post-training to inject language knowledge or achieve cross-lingual alignment, including ways of continual pre-training (Zhang et al., 2021b; Cui et al., 2023; Chen et al., 2023b; Lin et al., 2024b; Kuulmets et al., 2024; Jaavid et al., 2024) and instruction tuning (Muennighoff et al., 2023; Chen et al., 2023c; Indurthi et al., 2024; Ahuja et al., 2024; Lai and Nissim, 2024; Zhang et al., 2024c; Zhu et al., 2024a; Li et al., 2024; Zhao et al., 2024d). For example, Cui et al. (2023) attempt to inject Chinese knowledge into LLaMA by conducting continual pre-training on a large-scale Chinese corpus, while Zhu et al. (2023) focus more on building language alignment through cross-lingual instruction tuning and translation training.

Our proposed LENS stands out from existing methods in that we seek multilingual supervision signals from the *internal* language representation space of the LLMs, rather than relying heavily on *external* multilingual datasets as in the above methods, which offers fresh insights and new opportunities for enhancing the multilingual capabilities of LLMs both efficiently and effectively. For more discussions on the theoretical foundation of language-agnostic and language-specific subspaces within LLMs, please refer to Appendix G.

Representation Engineering Editing or manipulating representation within LLMs has garnered increasing attention due to its transparency and lightweight properties (Zou et al., 2023). This is theoritically rooted from Linear Representation Hypothesis (Mikolov et al., 2013; Nanda et al., 2023; Park et al., 2024). Building upon this, exist works attempt to edit representations at inference time to develop models that are more truthful (Li et al., 2023b; Campbell et al., 2023; Zhang et al., 2024a), and harmless (Lee et al., 2024; Uppaal et al., 2024). We expand and implement this paradigm for the multilingual enhancement of LLMs by focusing on representations during the training phase, ensuring that the efficiency of LLMs remains unaffected during the inference phase.

6 Conclusion

In this paper, we introduce LENS, a novel method designed for the effective, efficient and comprehensive multilingual enhancement of large language models (LLMs). LENS first decouple the multilingual hidden spaces of the backbone into two orthogonal components: a language-agnostic subspace and a language-specific subspace. Then taking well-established representations of the central language as a pivot, representations of target languages are pulled closer and pushed away from them in language-agnostic subspace and language-specific subspace, respectively. Experimental results on 3 representative cutting-edge LLMs demonstrate that LENS outperforms baseline methods with much lower training costs, underscoring its efficacy, efficiency and scalability.

7 Limitation and Future Work

Despite our LENS achieving comprehensive and efficient multilingual enhancement, there are still limitations and future directions worth exploring.

First, due to limited computational resources, our experiments are not conducted on larger-scale models (larger than 8B). This remains a valuable direction to apply LENS on larger LLMs.

Second, our current operations on language representation are still relatively coarse-grained. Future work could delve into more specific parameter areas for finer operations.

Finally, as we find that relying too much on external datasets to enhance multilingual capabilities may be limited, we instead seek higher quality supervision signals from within the model itself. Future work could consider combining these two paradigms by incorporating data selection strategies (Albalak et al., 2024; Liu et al., 2024b), thereby providing higher quality multilingual supervision signals to the model from both internal and external sources.

8 Ethical Considerations

This work is conducted solely for academic research purposes, aiming to enhance the multilingual capabilities of large language models in a responsible and interpretable manner. Our methods focus on improving cross-lingual alignment and representation learning without introducing biases or harmful content. We do not intend for this research to be used in applications that could lead to misinformation, discrimination, or unethical manipulation of language models. Additionally, we adhere to ethical AI principles, ensuring that our approach respects linguistic diversity and maintains fairness across different languages.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Awadallah, Monojit Choudhary, et al. 2024. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. *arXiv preprint arXiv:2407.09879*.

AI@Meta. 2024. Llama 3 model card.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Gábor Berend. 2020. Massively multilingual sparse word representations. In *International Conference on Learning Representations*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv preprint arXiv:2401.10440*.

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

James Campbell, Phillip Guo, and Richard Ren. 2023. Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching. In *Socially Responsible Language Modelling Research*.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, et al. 2024. Towards scalable automated alignment of llms: A survey. *arXiv preprint arXiv:2406.01252*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2023a. Is bigger and deeper always better? probing llama across scales and layers. *CoRR*.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023b. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, pages 5383–5395. PMLR.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023c. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660.

William Croft. 2002. *Typology and universals*. Cambridge university press.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

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*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in english? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Joseph H Greenberg. 1963. Universals of language. *The Massachusetts Institute of Technology*.

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

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4476–4494.

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

Jinyi Hu, Yuan Yao, Chongyi Wang, SHAN WANG, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2024. Large multilingual models pivot zero-shot multimodal learning across languages. In *The Twelfth International Conference on Learning Representations*.

Tianze Hua, Tian Yun, and Ellie Pavlick. 2024. mothello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1585–1598.

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*, pages 12365–12394.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficient boosting llm reasoning in non-english languages. *arXiv preprint arXiv:2405.17386*.

Sathish Reddy Indurthi, Wenxuan Zhou, Shamil Chollampatt, Ravi Agrawal, Kaiqiang Song, Lingxiao Zhao, and Chenguang Zhu. 2024. Improving multilingual instruction finetuning via linguistically natural and diverse datasets. *arXiv preprint arXiv:2407.01853*.

J Jaavid, Raj Dabre, M Aswanth, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models via romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615.

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

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

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 6912–6964.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026. Association for Computational Linguistics.

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

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8186–8213.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. In *Forty-first International Conference on Machine Learning*.

Seanie Lee, Hae Beom Lee, Juho Lee, and Sung Ju Hwang. 2022. Sequential reptile: Inter-task gradient alignment for multilingual learning. In *International Conference on Learning Representations*.

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 8051–8069.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

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.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024a. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024b. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 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.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024a. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 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.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30.

OpenAI. 2024. Gpt-4o system card. *OpenAI*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. 2020. Efficient domain generalization via common-specific low-rank decomposition. In *International conference on machine learning*, pages 7728–7738. PMLR.

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. 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.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, and Ekaterina Shutova. 2023. Probing llms for joint encoding of linguistic categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7158–7179.

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. 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.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rheeya Uppaal, Apratim De, Yiting He, Yiquao Zhong, and Junjie Hu. 2024. Detox: Toxic subspace projection for model editing. *arXiv preprint arXiv:2405.13967*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Aiti Aw, and Nancy Chen. 2024a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. 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 370–390.

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

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. 2021. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

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. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Roberts Adam, Kale Mihir, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yuzi Yan, Jialian Li, Yipin Zhang, and Dong Yan. 2024. Exploring the llm journey from cognition to expression with linear representations. In *Forty-first International Conference on Machine Learning*.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021a. Share or not? learning to schedule language-specific capacity for multilingual translation. In *Ninth International Conference on Learning Representations 2021*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

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

Shaolei Zhang, Tian Yu, and Yang Feng. 2024a. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.

Yongheng Zhang, Qiguang Chen, Min Li, Wanxiang Che, and Libo Qin. 2024b. AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9191–9200. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024c. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204. Association for Computational Linguistics.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, et al. 2021b. Cpm-2: Large-scale cost-effective pre-trained language models. *AI Open*, 2:216–224.

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

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024a. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Xinyu Zhao, Xuxi Chen, Yu Cheng, and Tianlong Chen. 2024b. Sparse moe with language guided routing for multilingual machine translation. In *The Twelfth International Conference on Learning Representations*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024c. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024d. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *arXiv preprint arXiv:2402.18913*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

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

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*, pages 2765–2781.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Algorithm 1: Language Subspace Probing

In: languages' mean embeddings $\boldsymbol{M}$, rank of subspace r

Out: language-agnostic subspace $\boldsymbol{M}_a$, language-specific subspace $\boldsymbol{M}_s$, coordinates $\boldsymbol{\Gamma}$

```
/* 1) Approximate M in low rank */
```

$\boldsymbol{M}'_a \leftarrow \frac{1}{d}\boldsymbol{M}\mathbb{1}$;

$\boldsymbol{M}'_s, _, \boldsymbol{\Gamma}' \leftarrow$ Top-r SVD $\left(\boldsymbol{M} - \boldsymbol{M}'_a\mathbb{1}^\top\right)$;

$\boldsymbol{M}' \leftarrow \boldsymbol{M}'_a\mathbb{1}^\top + \boldsymbol{M}'_s\boldsymbol{\Gamma}'^\top$;

```
/* 2) Force orthogonality        */
```

$\boldsymbol{M}_a \leftarrow \frac{1}{\|\boldsymbol{M}'^+\mathbb{1}\|^2}\boldsymbol{M}'^+\mathbb{1}$;

$\boldsymbol{M}_s, _, \boldsymbol{\Gamma} \leftarrow$ Top-r SVD $\left(\boldsymbol{M}' - \boldsymbol{M}_a\mathbb{1}^\top\right)$

A Probing for Language Subspace

The optimal solution of Equation 2 can be computed efficiently via Singular Value Decomposition (SVD). Algorithm 1 presents the detailed procedure. Readers interested in more details can consult the proof provided in Xie et al. (2022). The only hyperparameter $r < L$ controls the amount of language-specific information captured by the identified subspace. The larger r is, the more language-specific signals we can identify.

B Training Data

The multilingual data used for the language subspace probing stage is sourced from the Aya Dataset (Üstün et al., 2024), a human-annotated, non-parallel multilingual instruction fine-tuning dataset with 204,000 instances in 65 languages. For the language subspace manipulation stage, we rely on parallel multilingual data from the Bactrian-X dataset (Li et al., 2023a), which contains 3.4 million instruction-response pairs in 52 languages. These pairs are generated by translating 67,000 English instructions (derived from alpaca-52k (Taori et al., 2023) and dolly-15k) into 51 languages using the Google Translate API, and then obtaining natural responses from ChatGPT.

C Multilingual Benchmarks

We comprehensively measure the efficacy of our LENS on various multilingual tasks, including 5 mainstream benchmarks for evaluation. They can be categorized into the evaluation of multilingual understanding and multilingual generation.

For multilingual understanding:

- **XCOPA** (Ponti et al., 2020):[2] A benchmark to evaluate the ability of machine learning models to transfer commonsense reasoning across languages. The dataset is the translation and re-annotation of the English COPA (Roemmele et al., 2011) and covers 11 languages from 11 families and several areas around the globe. The dataset is challenging as it requires both the command of world knowledge and the ability to generalise to new languages. In our experimental setup, this benchmark covers both Chinese (Zh) and Swahili (Sw).

- **XWinograd** (Muennighoff et al., 2023):[3] A well-established tool for evaluating coreference resolution (CoR) and commonsense reasoning (CSR) capabilities of computational models. The dataset is the translation of the English Winograd Schema datasets and it adds 488 Chinese schemas from CLUEWSC2020 (Xu et al., 2020), totaling 6 languages. Formulated as a fill-in-a-blank task with binary options, the goal is to choose the right option for a given sentence which requires commonsense reasoning. In our experimental setup, this benchmark covers English (En), Chinese (Zh) and Japanese (Jp).

- **XStoryCloze** (Lin et al., 2022):[4] A commonsense reasoning framework for evaluating story understanding, story generation, and script learning. The dataset consists of the professionally translated version of the English StoryCloze dataset (Spring 2016 version) to 10 non-English languages. The dataset is challenging and is designed to evaluate story understanding, story generation, and script learning. In our experimental setup, this benchmark covers English (En), Chinese (Zh), Arabic (Ar) and Swahili (Sw).

- **M-MMLU** (Hendrycks et al., 2021; Lai et al., 2023):[5] A benchmark designed to measure knowledge acquired during pretraining by evaluating models exclusively in zero-shot and few-shot settings. The datasets is a ma-

[2] https://huggingface.co/datasets/cambridgeltl/xcopa

[3] https://huggingface.co/datasets/Muennighoff/xwinograd

[4] https://huggingface.co/datasets/juletxara/xstory_cloze

[5] https://huggingface.co/datasets/alexandrainst/m_mmlu

chine translated version of the MMLU dataset by GPT-3.5-turbo and covers 34 languages. This is a massive multitask test consisting of multiple-choice questions from various branches of knowledge. To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability. In our experimental setup, this benchmark covers English (En), Chinese (Zh), Arabic (Ar), Korean (Ko), and Swahili (Sw).

For multilingual generation:

- **MT-Bench** (Zheng et al., 2023): The dataset is designed for open-ended generation to evaluate a model's ability to follow multi-turn instructions. In our experimental setup, this benchmark covers English (En), Chinese (Zh), Arabic (Ar), Japanese (Jp), Korean (Ko), Swahili (Sw) and Bengali (Bn). We collect data in English[6], Japanese[7], Korean[8], and Arabic[9] from huggingface, and Chinese[10] from github. In addition, we use GPT-4o to translate the English data into Swahili and Bengali, and performed manual proofreading to ensure correctness.

D Baseline Methods

D.1 Multilingual Enhancement Baselines

For comparison, we consider the following baseline methods that enhance LLMs' multilingual capabilities using multilingual instruction fine-tuning technique: (1) **xSFT & xSFT-Full** (Ouyang et al., 2022): xSFT performs multilingual instruction fine-tuning using the same data volume as our LENS. In contrast, xSFT-Full utilizes the full dataset for each target language from the Aya Collection and Bactrian-X. (2) **QAlign** (Zhu et al., 2024a): It explores the benefits of question alignment, where the model is trained to translate inputs into English by finetuning on X-English parallel question data. (3) **SDRRL** (Zhang et al., 2024c): It is based on self-distillation from resource-rich languages that effectively improve multilingual performance by leveraging self-distillated data. (4) **CLA** (Li et al., 2024): It aligns the internal sentence representations across different languages via multilingual contrastive learning and aligns outputs by following cross-lingual instructions in the target language.

[6] https://huggingface.co/datasets/HuggingFaceH4/mt_bench_prompts

[7] https://huggingface.co/datasets/shi3z/MTbenchJapanese

[8] https://huggingface.co/datasets/StudentLLM/Korean_MT-Bench_questions

[9] https://huggingface.co/spaces/QCRI/mt-bench-ar/tree/main/data/mt_bench_ar

[10] https://github.com/HIT-SCIR/huozi

	Manipulated Layer	λ_{Zh}
LLaMA-3-8B-Instruct	31	1
LLaMA-3.1-8B-Instruct	30	0.05
Phi-3.5-mini-Instruct	27	0.25

Table 3: Detailed hyper-parameter settings for bilingual enhancement. The number under the column of Manipulated Layer represents the starting point of the layers where manipulation is applied, with the default endpoint being the final layer.

D.2 Open-Source Multilingual-Enhanced LLMs

The Chinese-LLaMA-3 series, which builds on LLaMA-3 series to enhance Chinese capabilities and includes three different versions:

- **Chinese-LLaMA-3-Instruct-V1**:[11] This model is continually pre-trained on 120GB of Chinese text and fine-tuned with 500 million carefully curated instruction data points, based on the LLaMA-3-8B. These training datasets is not available to the public.
- **Chinese-LLaMA-3-Instruct-V2**:[12] This version is directly fine-tuned on the same 500 million instruction data points using the LLaMA-3-8B-Instruct model.
- **Chinese-LLaMA-3-Instruct-V3**:[13] This model is created by merging V1, V2, and the original LLaMA-3-8B-Instruct, followed by fine-tuning on 5,000 instruction data points.

E Implementation Details

Our experiments are implemented with PyTorch (Paszke et al., 2019) and Transformer library (Wolf et al., 2020) on a single NVIDIA A800-SXM4-80GB GPU. The training duration is set to one epoch with the learning rate of 1e-5, cosine learning rate scheduler with warm up ratio of 0.05 and

[11] https://huggingface.co/hfl/llama-3-chinese-8b-instruct

[12] https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v2

[13] https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v3

	Manipulated Layer	λ_{Zh}	λ_{Jp}	λ_{Ko}	λ_{Ar}	λ_{Bn}	λ_{Sw}
LLaMA-3-8B-Instruct	29	1	0.6	1	0.5	0.2	0.2
LLaMA-3.1-8B-Instruct	30	0.01	0.01	0.03	0.01	0.01	0.01
Phi-3.5-mini-Instruct	29	0.2	0.2	0.2	0.2	0.15	0.15

Table 4: Detailed hyper-parameter settings for multilingual enhancement. The number under the column of Manipulated Layer represents the starting point of the layers where manipulation is applied, with the default endpoint being the final layer.

batch size of 8 across all backbones. And all backbones are trained with their official chat template with $\lambda_1 = 1$ and $\lambda_3 = 1$. The hyper-parameter r specifying the dimension of language-specific subspace in language subspace probing stage is set to $L - 1$, where L is the total number of languages participated in this process. We use GlotLID (Kargaran et al., 2023) to identify the response language to obtain the language fidelity. GlotLID is an open-source language identification model that supports more than 1,600 languages. GlotLID returns iso_636_9 language codes, which we manually map to the language codes in this work.

More detailed hyper-parameter settings for bilingual and multilingual enhancement across different backbones are listed in Table 3 and Table 4.

Further, we carefully evaluate the official implementations of all baselines, in order to make the comparison as fair as possible. All baselines are trained using the same language pairs as those in LENS. For SDRRL and CLA, we adhere to their respective configurations to reconstruct the training data from these language pairs. We strictly follow the hyper-parameter settings in their original code. If this could not reach the expected performance, we carry out the hyper-parameter search of the learning rate and batch size.

F Additional Experimental Results

F.1 Results under Multilingual Settings

Results of the comparison between LENS and baseline methods on Phi-3.5-mini-Instruct under bilingual and multilingual setups are shown in Table 5 and Figure 7, respectively. We report the results of multilingual enhancement settings on LLaMA-3-8B-Instruct and LLaMA-3.1-8B-Instruct are displayed in Figure 5 and 6, respectively.

The results demonstrate that our LENS is still capable of achieving the comprehensive multilingual enhancement. Similarly, LENS continues to improve the model's multilingual generation capability, enhancing the quality of the model's responses in specific languages. However, the improvement in language fidelity is more pronounced in the LLaMA-3-8B-Instruct backbone than in the multilingual backbones (LLaMA-3.1-8B-Instruct and Phi-3.5-mini-Instruct), which the latter ones undergo more extensive multilingual alignment training. Notably, while the baseline method considerably decreases the language fidelity of the multilingual backbones, LENS has minimal impact on it. These extensive experimental results demonstrate that LENS can serve as an effective, efficient, and scalable multilingual enhancement solution. We hope that our method can provide inspiration for future work to seek multilingual supervision more from the LLM itself rather than heavily relying on external dataset.

We also note that xSFT-Full exhibits uneven performance gains, especially in Swahili (Sw). We believe this may be attributed to the uneven quality of the training data. Specifically, the Bactrian-X dataset used for training derives its input from Google Translate and its output from GPT-3.5-turbo, meaning the dataset quality depends heavily on these two sources. As a result, inconsistencies in translation and generation quality can introduce noise, leading to uneven performance gains from data-driven post-training approaches like xSFT-Full. This highlights one of the key limitations of the current data-driven paradigms.

In contrast, LENS seeks supervision signals internally from the backbone itself, bypassing the need for extensive reliance on potentially noisy external datasets. This intrinsic approach allows LENS to achieve consistent improvements over the backbone model across a wide range of languages, demonstrating better scalability and robustness. We have also demonstrated this phenomenon in our experiments, showcasing LENS's broader applicability. In our future work, we propose combining the LENS training paradigm with advancements in data selection and filtering methods. We believe this hybrid approach holds great potential for further enhancing multilingual performance.

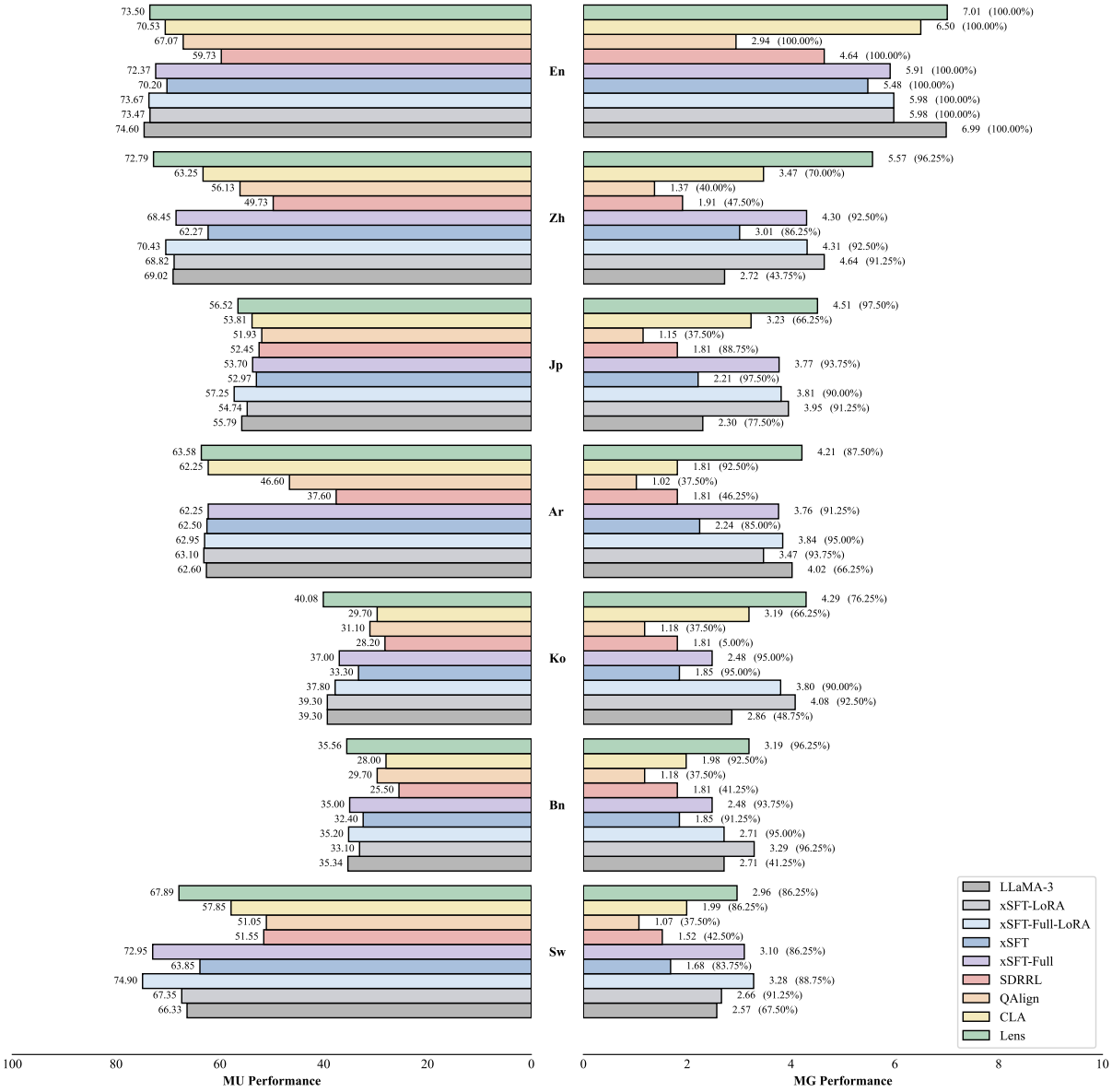


Figure 5: Results on the multilingual understanding and generation benchmarks with LLaMA-3-8B-Instruct backbone under the multilingual setting.

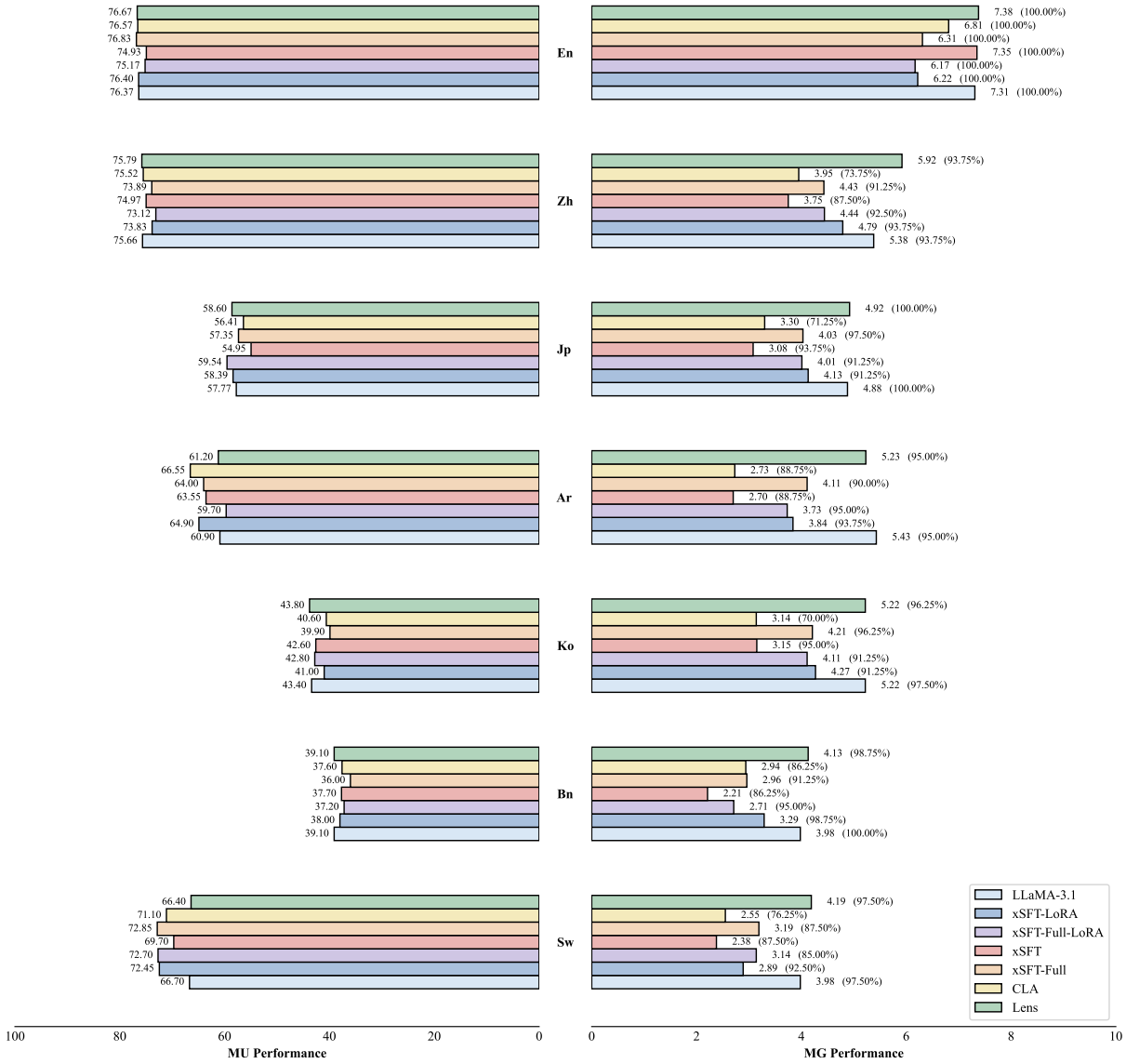


Figure 6: Results on the multilingual understanding and generation benchmarks with LLaMA-3.1-8B-Instruct backbone under the multilingual setting.

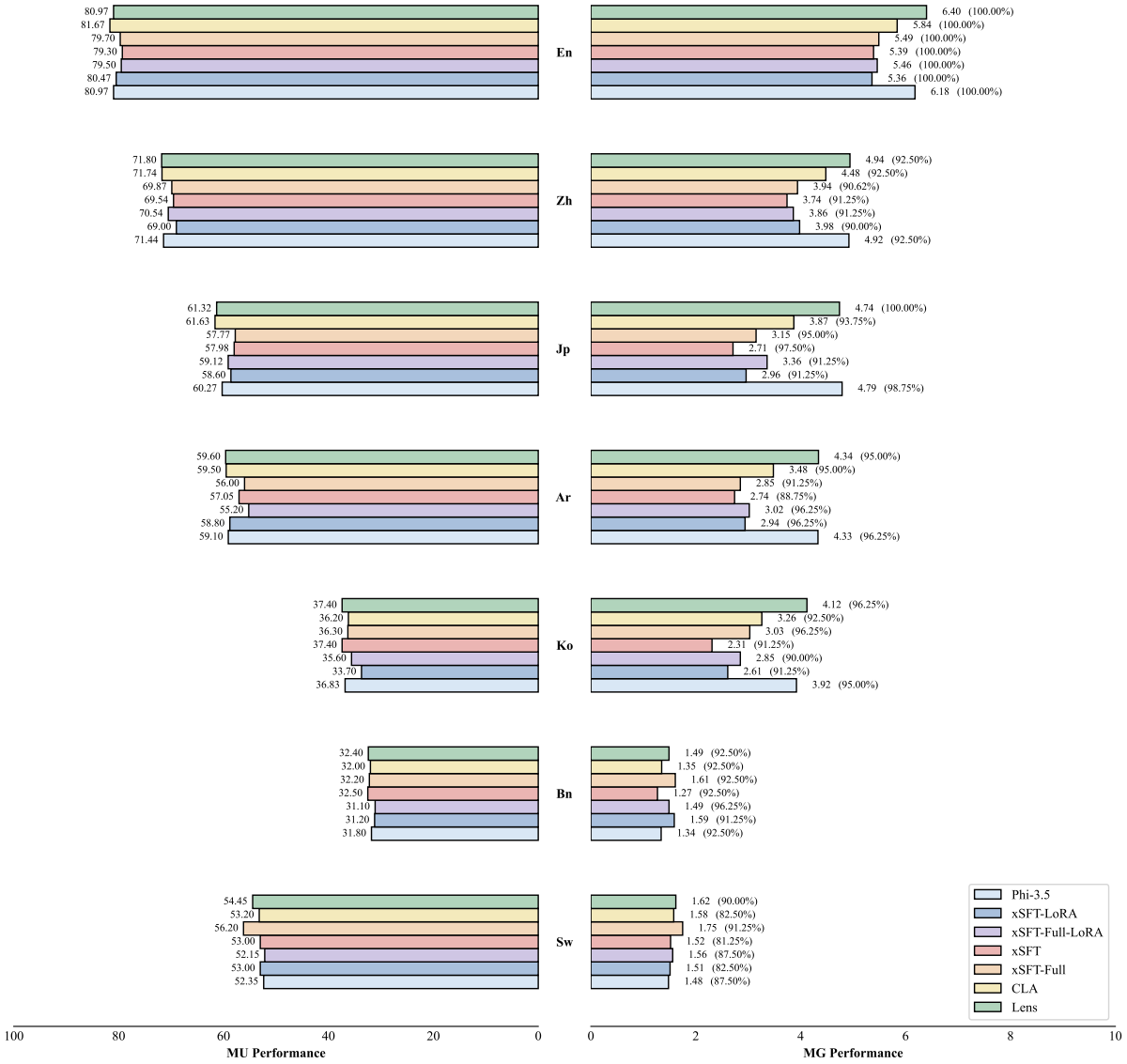


Figure 7: Results on the multilingual understanding and generation benchmarks with Phi-3.5-mini-Instruct backbone under the multilingual setting.

	Multilingual Understanding										Multilingual Generation	
	XCOPA		XWinograd		XStoryCloze		M-MMLU		AVG.		MT-Bench	
	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh
Phi-3.5	-	81.40	75.80	67.70	95.40	89.40	71.70	47.30	81.00	71.40	6.18 (100%)	4.92 (90.50%)
xSFT	-	80.80	**77.20**	**69.64**	95.40	89.40	71.70	46.80	**81.43**	71.66	5.29 (100%)	3.31 (88.75%)
xSFT-Full	-	80.40	73.10	65.67	95.20	88.20	**71.90**	44.70	80.07	69.74	5.25 (100%)	3.84 (87.50%)
QAlign	-	78.00	69.60	58.73	95.10	84.70	70.80	46.60	78.50	67.01	5.28 (100%)	3.15 (88.75%)
SDRRL	-	81.80	76.30	66.87	95.60	90.20	71.60	46.90	81.17	71.44	6.15 (100%)	4.03 (90.00%)
CLA	-	80.40	76.50	66.50	95.70	89.40	71.70	47.10	81.30	70.85	6.08 (100%)	4.26 (90.00%)
LENS (Ours)	-	**82.60**	75.80	68.73	**95.60**	**90.60**	71.80	**47.40**	81.07	**72.33**	**6.44** (100%)	**5.16** (**92.50%**)

Table 5: Detailed results on the multilingual understanding and multilingual generation benchmarks with Phi-3.5-mini-Instruct backbone under the bilingual setting (English and Chinese). Accuracy serves as the evaluation metric for multilingual understanding, while GPT-4o ratings (on a scale of 1 to 10) are provided for MT-Bench. The values in parentheses represent language fidelity. Results highlighted in green indicate an improvement or performance comparable to the original backbone, while those highlighted in red signal a decline in performance relative to the original backbone. The best and second-best results in our method and baselines are in bold and underlined.

	Lens	xSFT	xSFT-Full	SDRRL	QAlign	CLA	V1	V2	V3
Training time	2m08s	5m33s	192m35s	11m30s	12m03s	12m33s	-	-	-
Trainable parameters rate	5.43%	100.00%	100.00%	100.00%	100.00%	100.00%	13.08%	13.08%	-
Instruction data	1K	1K	111.5K	4K	1K	1K	5M	5M	5K
Pre-training data	-	-	-	-	-	-	120G	-	-

Table 6: Resource consumption of different multilingual enhancement methods under the bilingual enhancement setup. The backbone model is LLaMA-3-8B-Instruct.

F.2 Comparison with LoRA-Based SFT

In Figures 5, 6 and 7, We also perform comparison with LoRA-based (Hu et al., 2022) SFT as recent work suggests that LoRA-based SFT can effectively prevent catastrophic forgetting (Huang et al., 2024). Based on our experimental results, we derived the following key conclusions:

For preserving the central language's capabilities, incorporating LoRA-based SFT is indeed more effective at preventing catastrophic forgetting than its full-parameter counterpart. However, it primarily protects multilingual understanding (MU) tasks while multilingual generation (MG) capabilities are also significantly affected.

For target language enhancement, LoRA-based SFT methods also show a trend for improving MU tasks more over MG tasks.

By contrast, our proposed LENS achieves a more comprehensive performance, simultaneously enhancing understanding and generation for target languages while maintaining both the understanding and generation capabilities of the central language across different base models.

F.3 Broader Insights and Connection to Superficial Alignment Hypothesis

Our experimental findings suggest that eliciting language-specific representations benefits multilingual capability more than aligning language-agnostic representations, which may also lend support to the superficial alignment hypothesis (Zhou et al., 2023; Lin et al., 2024a; Yan et al., 2024). It posits that LLMs acquire their core knowledge and abilities during pretraining, while post-alignment training primarily guides the model towards a desirable subdistribution of formats to use when prompted. In the multilingual settings, this is specifically manifested in:

(1) Despite the imbalance distribution of training data in pretraining resources for different languages, the majority of language-agnostic knowledge is already well-comprehended and aligned during pretraining, especially for current LLMs exposed to super-large-scale pretraining corpora (e.g., over 15T tokens for LLaMA-3).

(2) Current post-alignment training, which disproportionately focuses on English data, limits other languages to a subdistribution aligned with English-specific formats.

Thus, further aligning multilingual representations may have less impact compared to stimulating language-specific expressiveness in the target languages, but both mechanisms contribute to performance improvement in our method, with separation playing a more significant role.

F.4 Results on Multilingual Machine Translation

We carry out evaluations on the FLORES-101 dataset (Goyal et al., 2022). Specifically, we assess the bidirectional translation performance between the target language and English, reporting scores using the COMET metric with the WMT22-comet-da model (Rei et al., 2022). The choice of target language and training settings align with the experimental setup used for multilingual enhancement on LLaMA-3-8B-Instruct.

The experimental results in Table 2 demonstrate that LENS still effectively enhances the multilingual machine translation performance, further validating its robustness across diverse multilingual tasks. Additionally, methods like QAlign, SDRRL and CLA, which heavily rely on translation-based training for multilingual alignment, fall short in acquiring deep linguistic understanding for each language. Thus, their ignorance of language-specific modeling hinders their efficacy in multilingual machine translation that require fine-grained linguistic nuances necessary for accurate translations.

F.5 Impact of Varying the Number of Manipulated Layers

Recent studies on the interpretability of LLMs has sought to reveal the mechanisms underlying their multilingual capabilities (Zhao et al., 2024c; Zhong et al., 2024). A growing consensus suggests that language-specific parameters or neurons are primarily concentrated in the upper layers of these models, while the middle layers tend to process inputs from various languages using a shared and language-agnostic mechanism (Chen et al., 2023a; Wendler et al., 2024; Tang et al., 2024; Kojima et al., 2024; Zhang et al., 2024d). Drawing inspiration from this, our main experiments focus on performing updates solely within the upper layers of the backbone, resulting in a notable improvement in multilingual performance. In Figure 3(a), we explore the effect of increasing the number of layers involved on the model's multilingual enhancement. The horizontal axis represents the starting point of the layers where manipulation is applied, with the default endpoint being the final layer. This experiment is performed under the bilingual enhancement with LLaMA-3-8B-Instruct.

"Thinking" in English at the intermediate layers is more favorable for improving multilingual understanding. If we partition representations of target language into the language-specific subspace too early at the middle layers, it may impair its multilingual understanding capability. On the contrary, inheriting more from the shared representations at the middle layers, while emphasizing language-specific representations only at the higher layers (where most language-specific parameters and neurons are concentrated), is more beneficial for enhancing multilingual performance.

It is important to note that modifying only the final layer does not significantly improve either multilingual understanding or generation. This is because language-specific information is not sufficiently enhanced, causing the model to suffer from off-target issues and struggle to represent specific languages accurately. The lack of improvement in multilingual understanding aligns with the findings in §4.1, which highlight the critical role of supervision provided by the Push loss ($\mathcal{L}_2$).

LENS further validates the conclusions of existing works on LLM interpretability and applies these findings to multilingual enhancement.

	MU			MG		
	Es	Fr	De	Es	Fr	De
LLaMA-3	53.50	52.50	56.50	5.88	5.27	4.56
xSFT	50.80	49.90	56.00	4.36	4.47	4.00
xSFT-Full	50.90	48.90	51.50	4.60	4.42	4.33
SDRRL	48.90	47.80	50.30	2.74	3.06	2.54
QAlign	48.50	46.60	51.30	2.88	2.91	2.31
LENS	**53.70**	52.10	**57.10**	**5.90**	**5.63**	**4.90**

Table 7: Detailed results on the multilingual understanding and multilingual generation benchmarks with LLaMA-3-8B-Instruct backbone for Spanish (Es), French (Fr), and German (De). Accuracy serves as the evaluation metric for multilingual understanding, while GPT-4o ratings (on a scale of 1 to 10) are provided for MT-Bench. The values in parentheses represent language fidelity. The best and second-best results in our method and baselines are in bold and underlined.

F.6 Impact of Language Proximity

Here are our supplemented experimental results based on LLaMA-3-8B-Instruct to further highlight the generalizability of LENS, where these 3 languages, Spanish (Es), French (Fr), and German (De), are typologically and scripturally closer to English. For Multilingual Understanding (MU) evaluation, we adopt M-MMLU dataset which covers all 4 languages En, Es, Fr and De. And Multilingual Generation (MG) evaluation is performed on MT-Bench. Results are shown in Table 7.

The horizontal comparison in Table 7 and Figure 5 reveals that it demonstrates effectiveness across languages from different typological families, further highlighting its generalizability and adaptability to diverse linguistic characteristics.

G Additional Related Works

Here we provide additional discussion on the theoretical foundation of language-agnostic and language-specific subspaces, dividing it into two aspects: linguistic theory and LLM interpretability.

Linguistic Theory From a linguistic standpoint, the idea of separating representations into language-agnostic and language-specific spaces is grounded in established theories of language universals and typology. Language-agnostic features align with universal linguistic structures, such as shared syntactic patterns or semantic primitives (Greenberg, 1963; Comrie, 1989), while language-specific features capture unique aspects like phonology, morphology, or syntax (Croft, 2002; Cotterell et al., 2016). These distinctions have also been studied in computational linguistics, such as in multilingual embeddings (Artetxe et al., 2018) and cross-lingual representation learning (Ruder et al., 2019), supporting the conceptual basis in LENS.

LLM Interpretability Recent interpretability studies have provided compelling evidence that LLMs internally encode language-agnostic and language-specific subspaces. For example, specific neurons or groups of neurons have been identified as responsible for mapping multilingual input representations into either a shared language-agnostic space (Chen et al., 2023a; Starace et al., 2023; Wang et al., 2024b; Chen et al., 2024; Wendler et al., 2024) that different languages share the common knowledge or distinct language-specific spaces (Tang et al., 2024; Kojima et al., 2024; Zhang et al., 2024d) that are crucial for the accurate expression for specific languages. These findings support our assumption that LLMs naturally exhibit such separable structures, and our work leverages this inductive bias to improve multilingual performance.

Building upon such two theoretical foundations, particularly from linguistic theory, most previous works regarding multilingual enhancement have focused on aligning representations in the language-agnostic space (Hu et al., 2024; Berend, 2020; Cao et al., 2020; Karthikeyan et al., 2020; Alaux et al., 2019; Wang et al., 2019) or aligning gradients during optimization (Lee et al., 2022; Wang et al., 2021) to leverage shared features across languages. However, few works in multilingual machine translation have considered language-specific characteristics, primarily to implement routing mechanisms or modular designs to improve performance (Zhao et al., 2024b; Zhang et al., 2021a).

In contrast, our proposed LENS goes a step further that it leverages both language-agnostic and language-specific subspaces to comprehensively enhance multilingual performance both inheriting the theoretical soundness and demonstrating practical utility of our approach.