LENS: RETHINKING MULTILINGUAL ENHANCEMENT FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the growing global demand for large language models (LLMs) that serve users from diverse linguistic backgrounds, most cutting-edge LLMs remain predominantly English-centric. This creates a performance gap across languages, restricting access to advanced AI services 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 encounter significant challenges, including the scarcity of high-quality multilingual datasets and the limited enhancement of multilingual capabilities. They often suffer from off-target issues and catastrophic forgetting of central language abilities. To this end, we propose LENS, a novel approach to enhance multilingual capabilities of LLMs by leveraging their internal language representation spaces. Specially, LENS operates by manipulating the hidden representations within the language-agnostic and language-specific subspaces from top layers of LLMs. Using the central language as a pivot, the target language is drawn closer to it within the language-agnostic subspace, allowing it to inherit well-established semantic representations. Meanwhile, in the language-specific subspace, the representations of the target and central languages are pushed apart, enabling the target language to express itself distinctly. Extensive experiments on one English-centric and two multilingual LLMs demonstrate that LENS effectively improves multilingual performance without sacrificing the model's original central language capabilities, achieving superior results with much fewer computational resources 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). However, despite this global trend, most state-of-the-art LLMs remain predominantly English-centric (Brown et al., 2020; Touvron et al., 2023a;b; Jiang et al., 2023; AI@Meta, 2024). These models exhibit significantly better performance in English than in other languages, leading to an imbalance in user experience and potentially excluding large segments 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 the multilingual capabilities of LLMs, aiming to provide more equitable access and performance across various linguistic communities. 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 inject or elicit multilingual knowledge with the supervision signals from *external* datasets.

While this paradigm is widely embraced and demonstrates certain successes, it faces several significant challenges. (1) The efficacy of multilingual enhancement heavily depend on large-scale and high-quality multilingual datasets (Zhou et al., 2023; Liu et al., 2024b), which are both time-consuming and labor-intensive to obtain for each language. (2) It favors improving multilingual understanding over generation capabilities, which leaves the off-target issue inadequately addressed.

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

As a result, the model often struggles to generate accurate responses in the intended language when prompted (Zhang et al., 2020; 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 & Cohen, 1989), such as English in the LLaMA family (Touvron et al., 2023a;b; AI@Meta, 2024).

In this work, we seek to provide a new perspective on addressing the aforementioned challenges 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 extended, its *language-agnostic* semantic representations should be *pulled close* to those of English, enabling it to quickly inherit general abilities in English without the need for supervision signals from external multilingual training data. Conversely, the *language-specific* linguistic representations of the target language should be *pushed away* from English to avoid off-target issues, ensuring accurate responses in the target 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. These subspaces are efficiently derived using 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. This allows the target language to directly inherit the well-established semantic representations of the central language. 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.

We conduct extensive experiments under bilingual and multilingual enhancement setups. Results on one English-centric (LLaMA-3-8B-Instruct) and 2 multilingual LLMs (LLaMA-3.1-8B-Instruct and Phi-3.5-mini-Instruct), demonstrate that LENS succeed to improve target languages on both multilingual comprehension and generation tasks without sacrificing the strong capability of central language, showing the efficacy and scalability of our method. Deeper analysis highlights the significance of steering the target language towards its unique expressions within its own language-specific subspace to fully enhance both comprehension and generation capabilities. This is overlooked by most existing approaches, which primarily focus on aligning representations across different languages to boost multilingual performance. 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 achieves high resource efficiency compared to baselines, with much fewer computational costs by only updating the model's higher layers using just a few hundred data points.

The main contributions of this work are summarized as follows:

- We provide a novel perspective for the multilingual enhancement of large language models with their internal language representation space leveraged.
- We propose LENS, an efficient and effective multilingual enhancement method that operates within the language representation space of large language models.
- Extensive experiments on one English-centric and two multilingual LLMs demonstrate the effectiveness, efficiency, scalability of our method to obtain truly multilingual enhanced chat-style backbones without sacrificing original central language performance.

2 RELATED WORKS

Multilingual Large Language Model With the acceleration of globalization, multilingual large language models (MLLMs) are gaining significant attention for their ability to handle multiple languages comprehensively (Qin et al., 2024). Pretraining on multilingual data is a common approach to gain the multilingual capabilities (Conneau & 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 state-of-the-art 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). For example, Huang et al. (2023) introduce cross-lingual-thought prompting to minimize language disparities. 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 or elicit extensive language knowledge for specific languages, including ways of continual pre-training (Zhang et al., 2021b; Cui et al., 2023; Chen et al., 2023b; Lin et al., 2024; 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 & 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 primarily 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.

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), which posits that various human-interpretable concepts are encoded in linear subspaces of model representations. 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.

3 METHODOLOGY

3.1 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.

3.2 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

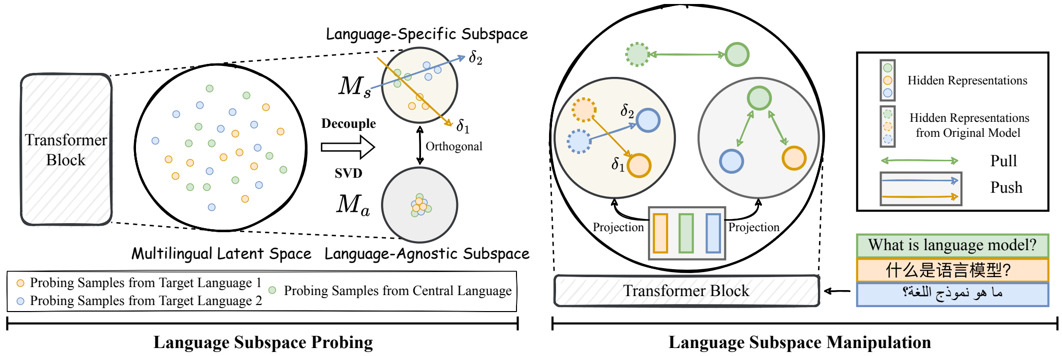

Figure 1: The overall architecture of our proposed LENS for multilingual enhancement. (1) In the LSP, we begin by decomposing the multilingual latent space, which is formed by the representations of probing samples from both the target and central languages. Using singular value decomposition (SVD), 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.

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. (2022), the objective can be formulated as:

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

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 B 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:

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

3.3 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:

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

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:

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

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.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models We select one representative English-centric LLMs: LLaMA-3-8B-Instruct (AI@Meta, 2024) and 2 MLLMs: LLaMA-3.1-8B-Instruct[2] and Phi-3.5-mini-instruct (Abdin et al., 2024), to fully validate the effectiveness and generality of our LENS in enhancing multilingual 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 chosen as 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 for enhancement based on the availability of language resources. 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 and MLLMs, which further underscores their relevance for enhancement.

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

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

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.

We sample 300 data points from the Aya Dataset for each language to probe the language space and 200 data points from the Bactrian-X dataset per language to manipulate the language space.

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 adopted as 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): The dataset is designed 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[3] emphasizes support for low-resource languages such as Swahili (sw) and Bengali (bn). This underscores the validity and reliability of the evaluation approach. In addition, 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 different methods mitigate the model's off-target issues.

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

4.2 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): 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.

4.3 IMPLEMENTATION DETAILS

LENS is a model-agnostic multilingual enhancement method that is compatible with different transformer-based LLM. 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 D.

4.4 OVERALL RESULTS

Table 1 and Figure 2 present the performance comparison between LENS and recent multilingual enhancement baselines on multilingual understanding and generation benchmarks, under bilingual and multilingual configurations, respectively. For additional results, including those on Phi-3.5-mini-instruct and multilingual configuration for the other two backbones, please see Appendix E. From the outcomes across all backbones, we have drawn the following key insights:

LENS succeed to achieve a comprehensive improvement for the multilingual capabilities of (M)LLMs without sacrificing original central language performance. Specifically, it enhances

[3] https://cdn.openai.com/gpt-4o-system-card.pdf.

Table 1: Detailed results on the multilingual understanding and multilingual generation benchmarks with the English-centric LLaMA-3-8B-Instruct backbone and the multilingual LLaMA-3.1-8B-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.

	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	-	<u>87.20</u>	**64.30**	<u>63.49</u>	**95.10**	<u>90.60</u>	<u>62.80</u>	46.10	<u>74.07</u>	<u>71.85</u>	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%)	<u>4.44</u> (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	<u>47.80</u>	73.73	68.31	<u>6.60</u> (100%)	3.84 (73.75%)
LENS (Ours)	-	**87.60**	<u>63.80</u>	**66.67**	<u>94.70</u>	**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	-	<u>88.00</u>	<u>63.70</u>	<u>67.46</u>	**96.20**	**92.70**	<u>68.10</u>	**53.10**	<u>76.00</u>	<u>75.32</u>	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	<u>6.02</u> (100%)	<u>4.18</u> (<u>92.50%</u>)
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%)
LENS (Ours)	-	**90.20**	**64.60**	**69.44**	<u>95.90</u>	<u>91.80</u>	**69.10**	<u>52.60</u>	**76.53**	**76.01**	**7.41** (100%)	**5.96 (93.75%)**

the multilingual capabilities of the backbone on both multilingual understanding and generation benchmarks, showing a marked increase in language fidelity during multilingual generation. This effectively mitigates the off-target issue. Moreover, LENS is the only approach that enhances multilingual performance across all languages. In contrast, baseline methods primarily focus on boosting multilingual understanding with little to no improvement in generation tasks. Additionally, methods like QAlign and SDRRL, which rely on translation-based training for multilingual alignment, fall short in effectively enhancing large models' overall multilingual performance. This suggests that aligning multilingual representations alone is insufficient for fully optimizing multilingual capabilities (Hua et al., 2024). Finally, LENS safeguards the central language from catastrophic forgetting, allowing the resulting model to effectively serve users from diverse linguistic backgrounds.

Using the central language representations within the backbone as a supervision signal proves more effective than relying on external data for supervision. 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. This phenomenon becomes more pronounced in xSFT-Full when trained with more data. 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). We hope LENS could inspire further research to explore more efficient, scalable, and automated supervision signals for multilingual enhancement of the most advanced LLMs.

5 ANALYSIS

5.1 ABLATION STUDY

We further conduct ablation studies to demonstrate the effectiveness of our proposed three optimization objectives in LSM. The overall results under the bilingual enhancement setting with LLaMA-3-8B-Instruct backbone are shown in Figure 3. We can draw the following key findings.

The alignment of multiple languages within language-agnostic subspaces mainly impacts multilingual comprehension rather than generation capabilities. As we incrementally raise the coef-

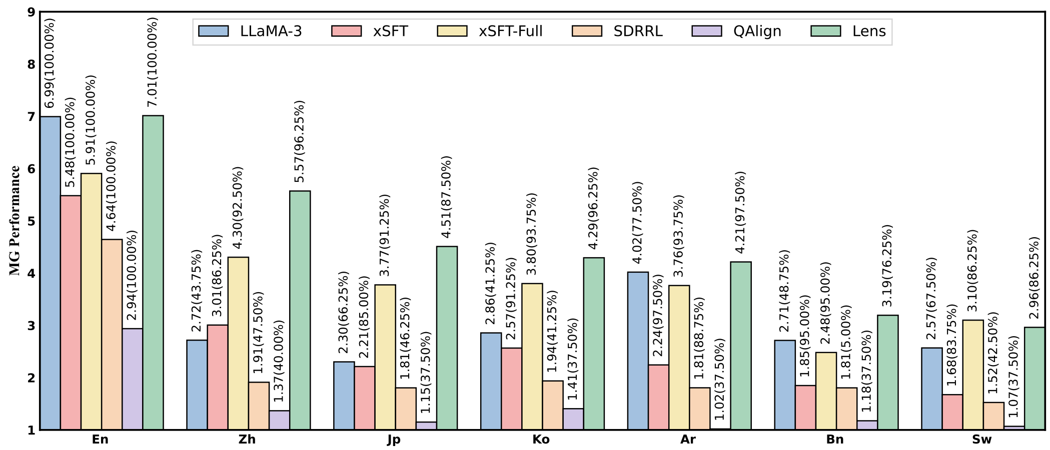

Figure 2: Results on the multilingual generation benchmark with LLaMA-3-8B-Instruct backbone under the multilingual setting. GPT-4o ratings (on a scale of 1 to 10) are provided for MT-Bench.

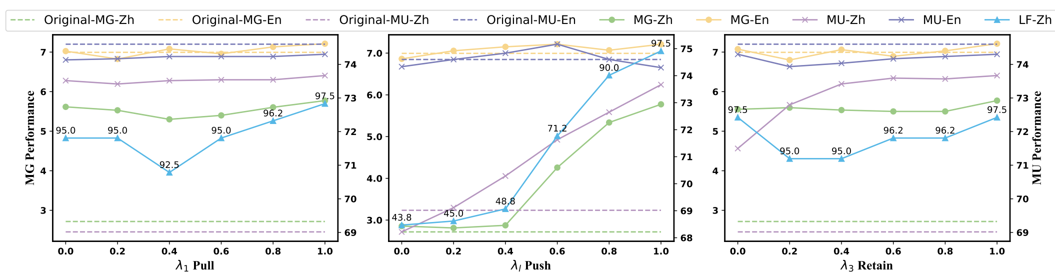

Figure 3: The ablation results to verify the effectiveness and impact of different optimization objectives in LSM. MU Performance stands for the average performance on all multilingual understanding benchmarks, while MG Performance is the results on MT-Bench. LF represents language fidelity.

ficient λ_1 of the multilingual alignment objective $\mathcal{L}_1$ from zero, Chinese comprehension improves, but its generation ability stays largely unaffected.

Enhancing the separation between representations of different languages in the language-specific subspace is vital for boosting multilingual performance. In particular, as illustrated in the middle part of Figure 3, both comprehension and generation abilities in Chinese improve significantly as the coefficient λ_l increases. This finding indicates that the commonly accepted notion of merely aligning languages to enhance multilingual capabilities (Cao et al., 2020; Zhu et al., 2023; Li et al., 2024; Hua et al., 2024) may not be sufficient for fully optimizing the multilingual performance of current LLMs. We hope this result encourages future research to focus more on eliciting and leveraging language-specific information within LLMs.

Maintaining the representations of the central language without significant changes can provide a stable and reliable alignment supervision for the target language to be enhanced. As illustrated on the right side of Figure 3, removing the objective to retain English representations leads to a significant decline in the backbone's Chinese performance. This could be due to alterations in the English representations during optimization, which may cause misalignment in the target for Chinese, thus impacting its performance. However, regarding the English capability, since our modifications are applied to the upper layers of the backbone (layers 31 and 32 in LLaMA-3-8B-Instruct), most of the parameters remain frozen and unaffected. As a result, even without the retaining objective, the backbone's English ability does not suffer from significant catastrophic forgetting. In the analysis shown in Figure 4, we observe that as the number of updated layers increases, the backbone's English understanding and generation capabilities are also not impacted by catastrophic forgetting. This can manifest the effectiveness of the retain optimization objective in safeguarding the performance of the central language.

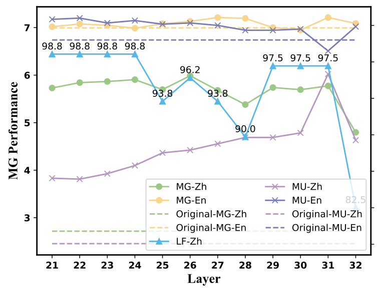

Figure 4: The impact of manipulating different backbone layers on multilingual performance enhancement.

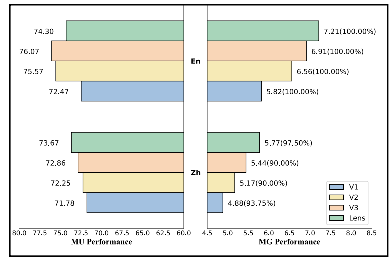

Figure 5: Comparison between our bilingual enhanced model with Chinese-LLaMA series.

5.2 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 4, 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 Section 5.1, which highlight the critical role of supervision provided by the Push loss ($\mathcal{L}_2$).

Our proposed Lens further validates the conclusions of existing works on LLM interpretability and applies these findings to multilingual enhancement of LLMs.

5.3 Comparison with Open-Sourced Multilingual-Enhanced LLMs

In Section 4.4, our main experiment primarily compares with *reproducible* baseline methods for multilingual enhancement. Additionally, we extend our comparisons to several open-source LLMs from the community 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 different versions:

- **Chinese-LLaMA-3-Instruct-V1**:[4] 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.

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

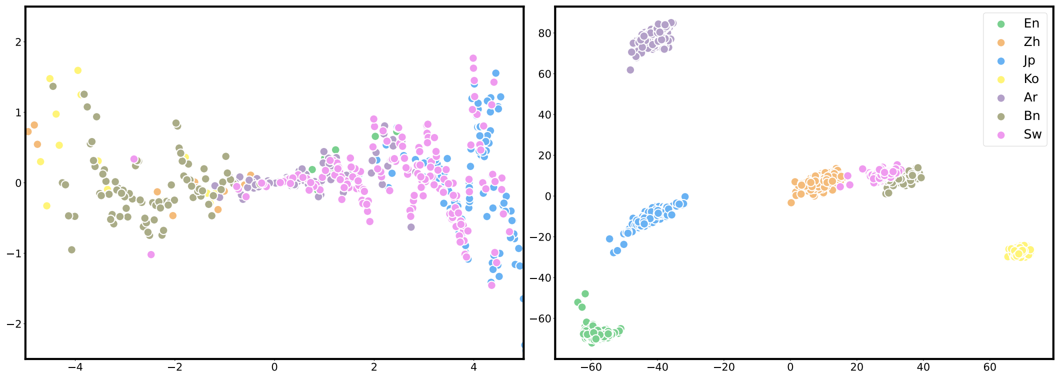

Figure 6: The PCA visualization of multilingual representations projected in the obtained language-agnostic subspace (right) and the language-specific (left) subspace. The backbone model is LLaMA-3-8B-Instruct after multilingual enhanced with LENS.

- **Chinese-LLaMA-3-Instruct-V2**:[5] 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**:[6] 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.

The experimental results and the resource consumption of different methods are presented in Figure 5 and Table 5 in Appendix E, respectively. The resulting model applied with LENS is identical to the one utilized for bilingual enhancement in Table 1. Remarkably, LENS demonstrates more comprehensive enhancement of the Chinese capabilities with extremely low resource overhead compared to these three models. This reinforces our claim that LENS is an efficient and effective approach for boosting the multilingual capabilities of LLMs. Additionally, 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.

5.4 VISUALIZATION ANALYSIS

To further confirm whether LENS manipulates language representations within different language subspaces as anticipated, we perform a visualization analysis. Specifically, as shown in Figure 6, we perform Principal Component Analysis (PCA) to visualize the projection of multilingual representations in our obtained language-agnostic subspace and the language-specific subspace. Parallel inputs in seven languages are sourced from the MultiQ datasets (Holtermann et al., 2024). The visualization 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.

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.

[5] https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v2
[6] https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v3

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. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. *arXiv preprint arXiv:2407.09879*, 2024.

AI@Meta. Llama 3 model card. 2024.

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

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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)*, pp. 789–798, 2018.

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

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

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

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

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

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

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

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

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 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, pp. 17817–17825, 2024.

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

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

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

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

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

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

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 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)*, pp. 550–564, 2024.

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

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

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

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

Tianze Hua, Tian Yun, and Ellie Pavlick. 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*, pp. 1585–1598, 2024.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 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*, pp. 12365–12394, 2023.

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

J Jaavid, Raj Dabre, M Aswanth, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 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)*, pp. 15593–15615, 2024.

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=dl4e3EBz5j.

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

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 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)*, pp. 6912–6964, 2024.

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

Huiyuan Lai and Malvina Nissim. 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)*, pp. 12012–12026. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.649.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuật Nguyễn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 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*, pp. 318–327, 2023.

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

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

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

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

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 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)*, pp. 8051–8069, 2024.

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

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

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

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, 2022.

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

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

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

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 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*, pp. 746–751, 2013.

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. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, 2023.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 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*, pp. 16–30, 2023.

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

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 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*, pp. 2375–2393, 2023.

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

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

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

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

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

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 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*, pp. 2695–2709, 2023.

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

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011. URL `https://people.ict.usc.edu/~gordon/publications/AAAI-SPRING11A.PDF`.

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

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 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)*, pp. 21–33, 2024.

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

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

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

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 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)*, pp. 5701–5715. Association for Computational Linguistics, 2024.

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

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

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

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

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. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Aiti Aw, and Nancy Chen. 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)*, pp. 370–390, 2024a.

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

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

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

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

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 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)*, pp. 15366–15394. Association for Computational Linguistics, 2024.

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

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

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

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

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 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*, pp. 5825–5832, 2021.

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

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

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

Shaolei Zhang, Tian Yu, and Yang Feng. 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)*, pp. 8908–8949, 2024a.

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

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 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)*, pp. 11189–11204. Association for Computational Linguistics, 2024c.

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

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

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

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

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

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

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

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

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

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

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

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

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781, 2024b.

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

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 */
```
$\boldsymbol{M}'_a \leftarrow \frac{1}{d}\boldsymbol{M}\mathbb{1}$;
$\boldsymbol{M}'_s, \_, \boldsymbol{\Gamma}' \leftarrow \text{Top-}r \text{ SVD}\left(\boldsymbol{M} - \boldsymbol{M}'_a\mathbb{1}^\top\right)$;
$\boldsymbol{M}' \leftarrow \boldsymbol{M}'_a\mathbb{1}^\top + \boldsymbol{M}'_s{\boldsymbol{\Gamma}'}^\top$;
```
/* 2) Force orthogonality */
```
$\boldsymbol{M}_a \leftarrow \frac{1}{\|\boldsymbol{M}'^{+}\mathbb{1}\|^2}\boldsymbol{M}'^{+}\mathbb{1}$;
$\boldsymbol{M}_s, \_, \boldsymbol{\Gamma} \leftarrow \text{Top-}r \text{ SVD}\left(\boldsymbol{M}' - \boldsymbol{M}_a\mathbb{1}^\top\right)$

A 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.

B 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.

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):[7] 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):[8] A benchmark to evaluate the ability of machine learning models to transfer commonsense reasoning across languages. The dataset is the translation of the English Winograd Schema datasets and it adds 488 Chinese schemas from CLUEWSC2020 (),

[7] https://huggingface.co/datasets/cambridgeltl/xcopa
[8] https://huggingface.co/datasets/Muennighoff/xwinograd

Table 2: 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.

	Manipulated Layer	λ_{Zh}
LLaMA-3-8B-Instruct	31	1
LLaMA-3.1-8B-Instruct	30	0.05
Phi-3.5-mini-Instruct	27	0.3

Table 3: 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.

	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	0.2	0.2

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):[9] A benchmark to evaluate the ability of machine learning models to transfer commonsense reasoning across languages. 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):[10] A benchmark to evaluate the ability of machine learning models to transfer commonsense reasoning across languages. The datasets is a machine 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[11], Japanese[12], Korean[13], and Arabic[14] from huggingface, and Chinese[15] 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 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

[9] https://huggingface.co/datasets/juletxara/xstory_cloze
[10] https://huggingface.co/datasets/alexandrainst/m_mmlu
[11] https://huggingface.co/datasets/HuggingFaceH4/mt_bench_prompts
[12] https://huggingface.co/datasets/shi3z/MTbenchJapanese
[13] https://huggingface.co/datasets/StudentLLM/Korean_MT-Bench_questions
[14] https://huggingface.co/spaces/QCRI/mt-bench-ar/tree/main/data/mt_bench_ar
[15] https://github.com/HIT-SCIR/huozi

Table 4: 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.

	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%)
LENS (Ours)	-	81.60	75.80	67.66	95.40	89.40	71.70	**47.40**	80.97	71.51	**6.44** (100%)	**5.16** (**92.50%**)

Table 5: Resource consumption of different multilingual enhancement methods under the bilingual enhancement setup. The backbone model is LLaMA-3-8B-Instruct.

	Lens	xSFT	xSFT-Full	SDRRL	QAlign	V1	V2	V3
Training time	2m08s	5m33s	192m35s	11m30s	12m03s	-	-	-
Trainable parameters rate	5.43%	100.00%	100.00%	100.00%	100.00%	13.08%	13.08%	-
Instruction data	0.8K	0.8K	111.5K	4K	0.8K	5M	5M	5K
Pre-training data	-	-	-	-	-	120G	-	-

epoch with the learning rate of 1e-5, cosine learning rate scheduler with warm up ratio of 0.05 and 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 2 and Table 3, respectively.

Further, we carefully evaluate the official implementations of all baselines, in order to make the comparison as fair as possible. 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.

E ADDITIONAL EXPERIMENTAL RESULTS

We report the multilingual understanding performance of LLaMA-3-8B-Instruct in Figure 7. Experimental results of the comparison between LENS and baseline methods on Phi-3.5-mini-Instruct under bilingual and multilingual setups are shown in Table 4 and Figure 9, respectively. And the multilingual enhancement results for LLaMA-3.1-8B-Instruct are displayed in Figure 8.

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 English-centric backbone than in the multilingual backbone, which the latter one undergoes more extensive multilingual alignment training. Notably, while the baseline method considerably decreases the language fidelity of the multilingual backbone, 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

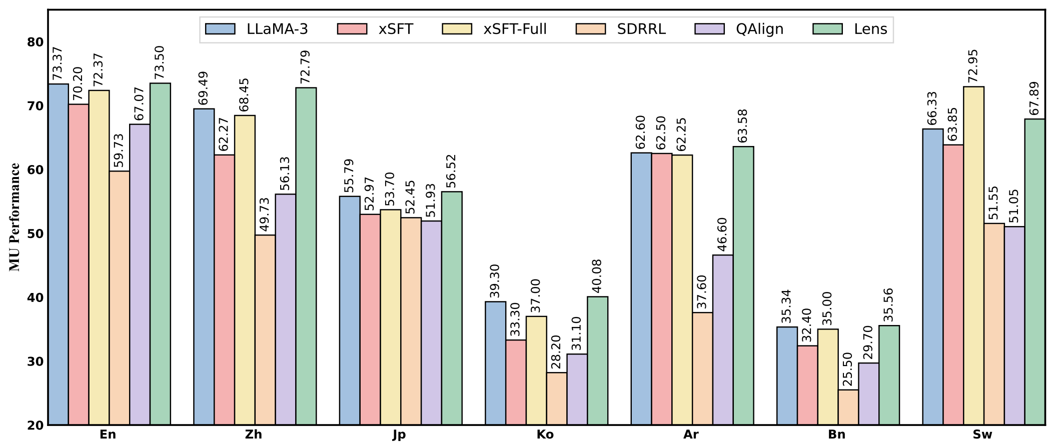

Figure 7: Results on the multilingual understanding benchmark with LLaMA-3-8B-Instruct backbone under the multilingual setting. We report the average performance of each language on the corresponding benchmarks.

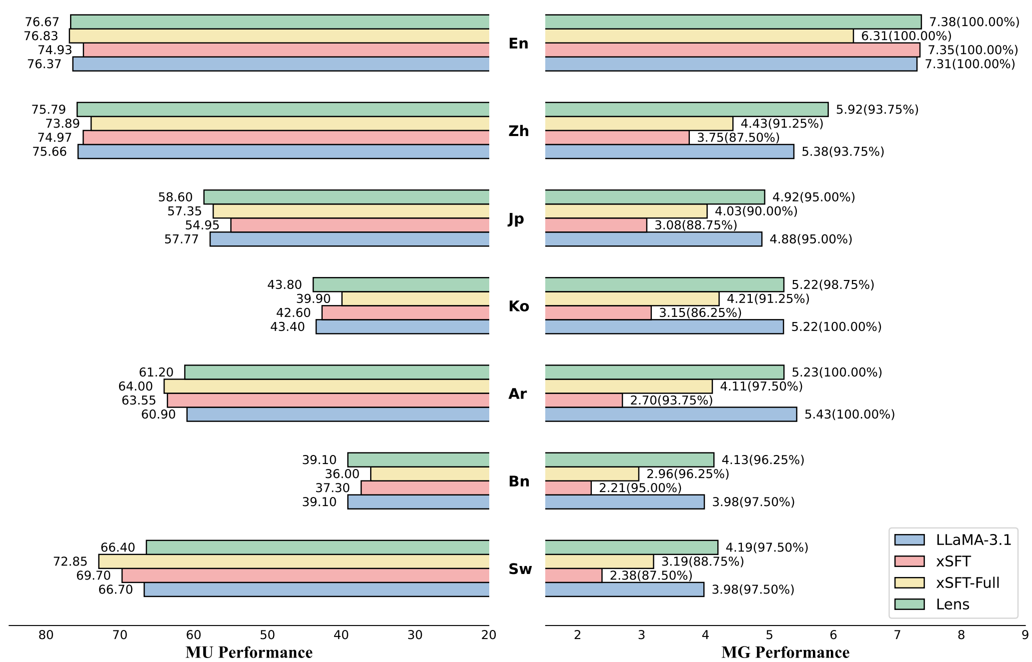

Figure 8: Results on the multilingual understanding and generation benchmarks with LLaMA-3.1-8B-Instruct backbone under the multilingual setting.

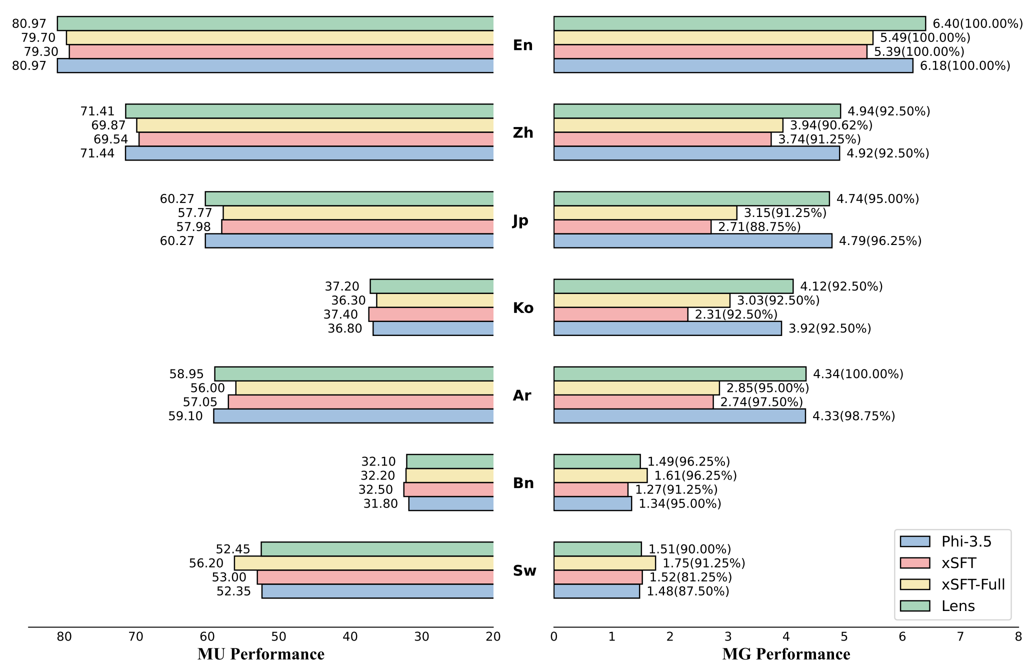

Figure 9: Results on the multilingual understanding and generation benchmarks with Phi-3.5-mini-Instruct backbone under the multilingual setting.

method can provide inspiration for future work to seek multilingual supervision more from the LLM itself rather than heavily relying on external dataset.

F 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.