Getting More from Less: Large Language Models are Good Spontaneous Multilingual Learners

Shimao Zhang♣, Changjiang Gao♣, Wenhao Zhu♣, Jiajun Chen♣, Xin Huang◇,
Xue Han◇, Junlan Feng◇, Chao Deng◇, Shujian Huang♣∗

♣National Key Laboratory for Novel Software Technology, Nanjing University
◇China Mobile Research Beijing, China
{smzhang,gaocj,zhuwh}@smail.nju.edu.cn, {huangsj,chenjj}@nju.edu.cn
{huangxinyjy,hanxueai,fengjunlan,dengchao}@chinamobile.com

Abstract

Recently, Large Language Models (LLMs) have shown impressive language capabilities, while most of them have very unbalanced performance across different languages. Multilingual alignment based on the translation parallel data is an effective method to enhance LLMs' multilingual capabilities. In this work, we first discover and comprehensively investigate the spontaneous multilingual alignment of LLMs. Firstly, we find that LLMs instruction-tuned on the question translation data (i.e. without annotated answers) are able to encourage the alignment between English and a wide range of languages, even including those unseen during instruction-tuning. Additionally, we utilize different settings and mechanistic interpretability methods to analyze the LLM's performance in the multilingual scenario comprehensively. Our work suggests that LLMs have enormous potential for improving multilingual alignment efficiently with great language generalization and task generalization. [1]

1 Introduction

Large Language Models (LLMs) have recently shown impressive language capabilities across numerous downstream language tasks (Zhao et al., 2023). However, most of the existing LLMs are trained on extensive high-resource languages text (Touvron et al., 2023; Brown et al., 2020; Jiang et al., 2023), which lead to a significant performance gap between high-resource languages and low-resource languages (Huang et al., 2023; Zhang et al., 2023b; Gao et al., 2024). For the same task and question contents, using different languages for inputs may have a significant impact on the model's performance.

Some studies have conducted comprehensive exploration about how to enhance the LLMs' ca-

pabilities across different languages. The classical approach typically follows the translate-based paradigm (Liu et al., 2024). Considering LLMs' great performance on the high-resource languages, some cross-lingual alignment and transfer methods are proposed (Eronen et al., 2023; Zhu et al., 2024; Zhao et al., 2024a). Question alignment (Zhu et al., 2024) is an outstanding paradigm among these methods which effectively improves multilingual alignment at lower cost, i.e. only utilizes the X-English parallel question translation data.

Meanwhile, some studies have further explored the LLMs, revealing that English also participate in the intermediate latent reasoning of these models even when LLMs are prompted in non-English (Wendler et al., 2024; Zhao et al., 2024b). These findings suggest that for LLMs, different languages are not isolated, and LLMs are able to leverage the connections between various languages to address problems in the multilingual scenarios. Researchers also reveal the shared semantic space for different languages (Chang et al., 2022), which is consistent with the findings above and indicates the importance of the multilingual alignment. Also, Kew et al. (2023) discover that multilingual instruction-tuning with three languages improves model's cross-lingual transfer abilities on some generative tasks.

Intuitively, LLMs have abilities to acclimatize themselves to the multilingual environment through appropriate training (Shi et al., 2022). Many existing methods rely on instruction-tuning on the multilingual instruction-tuning datasets (Kew et al., 2023; Liu et al., 2024). However, given the question alignment paradigm, utilizing multilingual alignment is also helpful for improving LLMs' multilingual abilities. Additionally, we focus on question alignment in our work to eliminate the interference of task-related data with annotated answers from our analysis of multilingual alignment. Based on the findings above, can LLMs achieve better

---

∗Corresponding author
[1] Our code and data is available at: https://github.com/Shimao-Zhang/LLM-Multilingual-Learner.

multilingual alignment across different languages efficiently through appropriate methods?

In this work, we investigate the multilingual alignment of LLMs, where we only train the LLMs on the parallel data without annotated answers (only queries) in a few languages. Following question alignment, we conduct the experiments on models in different types (English-centric or not) and parameter sizes, and test across a wide range of languages on different benchmarks. We find that question alignment following Zhu et al. (2024) can effectively enhance the multilingual capabilities of LLMs, which indicates that models can effectively utilize the relevant knowledge and capabilities learned during the pretraining process with question alignment, consisting with the "Superficial Alignment Hypothesis" (Zhou et al., 2024). Our results also indicate that conducting question alignment in a small number of languages brings much better multilingual alignment even between English and many languages unseen during instruction-tuning process, which implies good language generalization. Furthermore, we also use logit lens (Nostalgebraist, 2020) and dimensionality reduction techniques (Pearson, 1901) to study the latent states of LLMs, providing more comprehensive perspectives and empirical results for the alignment improvements in our experiments.

2 Background

2.1 Unbalanced Multilingual Performance

With a much larger number of parameters pretrained on a massive corpus, LLMs have shown the impressive capabilities in a variety of language tasks (Zhao et al., 2023). These models are mainly pretrained on English data, which often accounts for 90% or even more of all training data. We present a partial language distribution of LLaMA-2's training data in Table 7 in the Appendix A. Meanwhile, most of the LLMs also show unstable and unbalanced performance in multilingual scenarios, especially for some low-resource languages (Zhang et al., 2023a; Zhu et al., 2024). It's important to enable LLMs to adapt to a wider range of users and scenarios.

2.2 Cross-lingual Enhancement for Large Language Models

While LLMs still exhibit significant shortcomings in multilingual scenarios, many researchers propose multilingual LLMs that are specifically ad-

justed for multilingual tasks (Barbieri et al., 2021; Le Scao et al., 2023; Wei et al., 2023). But for multilingual LLMs, researches indicate a decline in their performance in English because of the limited tokens and parameter size (Lin et al., 2022; Scao et al., 2022).

Based on the existing LLMs, researchers have made great efforts to enhancing the multilingual performance, which include two categories: prompting close-source LLMs and instruction-tuning open-source LLMs. For the former, some studies utilize translation-based strategies which translate the non-English input into English firstly before solving the problem (Huang et al., 2023; Qin et al., 2023). This type of approaches are constrained by the translation quality of the model itself and is cumbersome for users. In addition, cross-lingual prompting mechanism is also widely used to improve LLMs' multilingual performance (Ranaldi et al., 2024).

For the latter, LLMs shows significant improvement of multilingual abilities and good task generalization through multilingual multitask finetuning (Muennighoff et al., 2022). Chen et al. (2023) follow the translation-based approach and instruction-tune the model on a multilingual version of GSM8K, which is translated from English GSM8K (Cobbe et al., 2021). Liang et al. (2024) create a new intermediate language MUL (Machine-created Universal Language) as a translatable unified representation of shared concepts across different languages. "X-English" parallel question translation data have also been used for multilingual question alignment (Zhu et al., 2024). In our work, we mainly analyse based on the question alignment, which is an outstanding alignment methods, and eliminates the interference of the annotated answers from our analysis.

2.3 Mechanistic Interpretability

In addition to improving the performance of LLMs, it is also crucial to understand and explain the principles of neural networks and related methods explicitly. Current works mainly analyze LLMs' actions by observing the internal states during the inference process. Intermediate logits and neuron activation states are both important objects of observation.

Although the English-centric LLMs are mainly trained on English data, they also show good performance across some non-English languages (Shi et al., 2022). Logit lens (Nostalgebraist, 2020) is an

early proposed technique that uses the model head in the final layer to project the intermediate latent logits directly to the vocabulary space. It have been evidenced that LLaMA 2 (Touvron et al., 2023), a open-source English-centric LLMs, have explicit English output in its latent states even when having non-English inputs (Wendler et al., 2024). There is also a hypothesis about how LLMs handle multilingualism that LLMs solve task by English with the help of multilingual knowledge, and output in the target language finally (Zhao et al., 2024b). All these results indicate that there are connections between various languages for LLMs, and LLMs have the capability to spontaneously learn to utilize multiple languages to solve problems. Zhao et al. (2024b) calculate the overlapping ratio of the language-specific neurons of different languages in different layers. The results indicate that neurons belonging to different languages exhibit clear distribution differences. In our experiments, we utilize logit lens and dimensionality reduction techniques to help us better understand the mechanism behind our findings.

3 Analysis Pipeline

We investigate the effect of question translation parallel data on LLMs' performance across a wide range of languages even unseen during the fine-tuning process.

We define the universal set of languages as $\mathbf{U}$:

$$\mathbf{U} = \{l_0,\ l_1,\ l_2,\ ...\ ,\ l_{n-1}\} \quad (1)$$

where l_i is the i-th language in $\mathbf{U}$, n is the total number of languages. We let l_0 refer to English specially here.

We select a few of non-English languages $\mathcal{L}_s = \{l_i, ..., l_k\} \subseteq \mathbf{U}$, and a target language $l_t \in \mathbf{U}$, $l_t \notin \mathcal{L}_s$. Then we will construct translation parallel data from every language $l \in \mathcal{L}_s$ to target language l_t. When constructing the translation data, we only use the questions without annotated answers. Then we get a translation dataset $\mathcal{Q}_{train}$ including source question $\mathcal{Q}_s$ and the corresponding target question $\mathcal{Q}_t$, which means $\mathcal{Q}_{train} = \{(q_s, q_t) \mid q_s \in \mathcal{Q}_s \ and \ q_t \in \mathcal{Q}_t\}$. We instruct-tune the model on the translation task:

$$\arg\min_{\theta} \sum_{(q_s,q_t) \in \mathcal{Q}_{train}} -\log p_\theta(q_t \mid q_s) \quad (2)$$

where θ is the model parameters, $\mathcal{Q}_{train}$ is the whole training translation dataset, q_s is the question in the source language, q_t is the question in the target language. Then we get the trained model:

$$\theta^{'} = \theta + \Delta\theta \quad (3)$$

We use question translation data for training to eliminate the impact of annotated answers themselves. And we use in-context learning for test while the model haven't been trained on the corresponding task.

We test the trained model on all languages $l \in \mathbf{U}$. We construct the testing dataset $\mathcal{Q}_{test} = \{\mathcal{Q}_l \mid l \in \mathbf{U}\}$ for every language, where $\mathcal{Q}_l$ consists of all test questions in the language l.

$$\text{Accuracy}_l = \sum_{q \in \mathcal{Q}_l} \mathbf{I}_{\theta^{'}}(\hat{a} = a \mid q) \quad (4)$$

$$\text{Accuracy} = \frac{\sum_{l \in \mathbf{U}} \text{Accuracy}_l}{|\mathbf{U}|} \quad (5)$$

where $\mathbf{I}$ is a function that takes 1 when the proposition is true and 0 otherwise. $\mathcal{Q}_l$ denotes the test dataset of language l. $\mathbf{U}$ is the universal set of languages we use in our work. $\hat{a}$ is the answer that the model predicts base on q, and a is the golden answer corresponding to q.

4 Experimental Setup

We conduct our experiments on both English-centric and non-English-centric models. And we utilize different representative tasks and different model parameter sizes to further strengthen our conclusions. In this section, we introduce our experimental settings in detailed.

Models We choose representative open-source LLMs for our experiments:

- **Mistral:** Mistral-7B-v0.1 (Jiang et al., 2023) is an advanced open-source English-centric large language model, which is one of the most popular open-source LLMs.

- **Qwen:** To enhance the generalization and reliability of our conclusions, we also choose models of different types and parameter sizes. Qwen1.5 is a newly released and enhanced version of Qwen (Bai et al., 2023). Qwen1.5 is pretrained on a multilingual dataset with a significant portion of the data being in English and Chinese, which means it is not an English-centric model. We choose Qwen1.5-1.8B, Qwen1.5-4B, and Qwen1.5-14B for our experiments.

Mistral-7B	en	zh	de	fr	es	it	nl	ja	ru	sv
base	89.2	92.4	91.8	93.4	94.2	93.8	93.6	93.0	93.2	93.4
zh ⇒ en	95.2	94.8	94.8	**95.2**	94.4	94.4	94.8	94.4	94.0	**95.4**
sw ⇒ en	**95.4**	93.4	94.2	94.4	94.2	94.4	93.0	93.6	93.8	94.8
zh/es ⇒ en	95.2	95.0	**95.0**	95.0	94.8	92.8	94.6	**95.0**	**94.4**	94.8
zh/de ⇒ en	95.2	95.4	94.8	**95.2**	95.2	95.2	94.8	93.6	94.2	94.6
zh/it ⇒ en	**95.4**	**95.8**	94.8	94.0	95.2	92.6	94.4	93.0	94.2	95.2
sw/hi ⇒ en	**95.4**	94.6	94.4	93.4	93.4	93.6	93.6	94.0	93.8	94.4
sw/th ⇒ en	**95.4**	95.0	93.8	93.4	93.4	92.8	93.6	92.6	93.2	94.0
zh/es/ru ⇒ en	**95.4**	95.4	94.4	94.0	94.6	92.6	94.6	94.2	94.0	94.2
zh/de/it ⇒ en	95.2	95.6	94.4	95.0	94.0	93.8	**95.0**	93.6	94.2	94.6

Mistral-7B	sl	pl	bg	no	ms	is	hi	th	sw	bn
base	87.6	93.2	91.6	92.4	91.8	63.2	81.6	83.0	58.0	71.0
zh ⇒ en	**94.0**	94.0	**94.6**	92.2	89.0	84.0	88.8	**88.4**	75.8	**81.0**
sw ⇒ en	89.8	92.6	93.6	93.4	90.0	72.0	64.4	51.4	**81.2**	54.0
zh/es ⇒ en	93.2	93.6	94.0	93.0	92.2	81.2	87.0	84.8	75.6	75.4
zh/de ⇒ en	93.4	94.0	**94.6**	**93.6**	92.2	**86.6**	84.8	**88.4**	71.8	68.6
zh/it ⇒ en	92.6	93.8	94.2	**93.6**	**92.6**	84.2	77.6	77.2	71.6	60.0
sw/hi ⇒ en	89.2	93.0	93.2	92.6	90.0	71.8	**89.8**	87.0	77.6	79.4
sw/th ⇒ en	92.8	92.0	93.2	87.2	84.4	79.4	86.8	84.0	81.0	74.2
zh/es/ru ⇒ en	93.6	**94.2**	93.4	93.4	91.4	83.8	85.0	86.0	77.0	76.0
zh/de/it ⇒ en	91.2	93.6	94.2	93.4	91.8	83.2	77.2	82.4	69.0	71.4

Table 1: Accuracy of Mistral-7B base model and aligned models on the Amazon Reviews Polarity. We report at least two sets of results for each language quantity to strengthen our conclusions. The accuracy of randomly choosing is 50.0%. "X/Y/Z ⇒ T" means using a randomly mixed dataset including 10k X to T, 10k Y to T, 10k Z to T translation data for instruction-tuning. We highlight the best results for every language.

Datasets Following Wendler et al. (2024), we select test tasks based on two fundamental principles:

1. **Obvious Answers:** Obvious answers reduce the entropy during the inference process, minimizing the impact of irrelevant tokens on our analysis.

2. **Fixed Answers:** Fixed answers (as opposed to open-ended responses) provide clearer observation targets, facilitating analysis through observing the latent outputs of the model. Deterministic outputs also make it easier for us to control the model's outputs.

Based on these, we conduct our experiments on three different types of tasks:

- **Sentiment analysis:** Sentiment analysis is an important and classic NLP task (Alswaidan and Menai, 2020), which always has three common outputs: "positive", "negative", and "neutral". We choose Amazon Reviews Polarity[2] (Zhang et al., 2015), a famous dataset includes two classes "positive" and "negative", to construct the parallel data mentioned in §2.2 and the test data. We extract 10K instances from train subset for parallel data and

500 instances from test subset for test data respectively.

- **Natural Language Inference:** Natural language inference (NLI) aims to judge the relationship between a given premise and a hypothesis sentence. There are always three possible outputs: "entailment", "neutral", and "contradiction". We choose SNLI[3] (Stanford Natural Language Inference) (Bowman et al., 2015) for our experiments. Following the settings above, we extract 10K instances from train subset for parallel data and 600 instances from test subset for test data respectively.

- **Paraphrase Identification:** Model needs to judge if two given sentences are semantically equivalent in the paraphrase identification task, which includes two possible labels. We conduct our experiments on PAWS[4] dataset (Zhang et al., 2019), which is a famous dataset proposed by Google. Following the above tasks, we extract 10K instances from train subset for parallel data and 500 instances from test subset for test data respectively.

---

[2] https://huggingface.co/datasets/amazon_polarity

[3] https://huggingface.co/datasets/stanfordnlp/snli

[4] https://huggingface.co/datasets/google-research-datasets/paws

Languages We conduct our following experiments across 20 languages in this work. As shown in Table 7 in Appendix A, we choose English (en), German (de), French (fr), Swedish (sv), Chinese (zh), Spanish (es), Russian (ru), Dutch (nl), Italian (it), and Japanese (ja) as the top 10 highest-resource languages according to Touvron et al. (2023). Additionally, we choose another 10 representative languages to strengthen our work, including Slovenian (sl), Polish (pl), Bulgarian (bg), Norwegian (no), Malay (ms), Icelandic (is), Hindi (hi), Thai (th), Swahili (sw), and Bengali (bn).

Implementations We use LoRA (Hu et al., 2021) to instruction-tune the pretrained models on the mixed parallel translation data first. We train LLMs on the translation data excluding the golden answers to mitigate the impact of the data of the tasks themselves on the model's capabilities. We use in-context learning which not only doesn't interfere with LLMs' parameters but also help LLMs handle the tasks better. We use constrained decoding rather than sampling that is used for diverse generation (Zhang et al., 2024) to eliminate the interference of irrelevant outputs on the results.

For instruction-tuning process we mentioned above, we use LoRA (rank = 8, α = 16) with 3 epochs (1 epoch for PAWS to mitigate overfitting), batch_size = 16, learning_rate = 5e-5, val_size = 0.05, lr_scheduler_type = cosine, cutoff_len = 2048 based on the settings of LLaMA-Factory[5] (Zheng et al., 2024), a widely used and recognized open-source project for LLMs efficient fine-tuning. We use single NVIDIA RTX A6000 48GB or single NVIDIA Tesla V100 SXM2 32GB for training. Training time varies from 4 hours to 10+ hours depending on the language and the total instance quantity.

More details are shown in Appendix B.

5 Results

In this section, we report the main results across different experimental settings and conduct some discussions based on the results.

5.1 Main Results

We report the accuracy of Mistral-7B on sentiment analysis task in Table 1. Clearly, we can see that the models trained on multilingual translation data outperform the original model obviously across a

lot of languages, which indicates that model have much stronger multilingual capabilities after a multilingual training. We summarize our empirical findings as follows:

1. **Large language models can learn to handle multilingualism better spontaneously.** Traditionally, fine-tuning or alignment on the target languages is needed to help the model adapt. However, our results indicate that LLMs are able to perform effective learning and transfer across multiple languages without parallel data for most of them. As seen, models have much higher overall accuracy across 20 languages after training on the parallel data containing 2-4 languages.

2. **High-resource languages are not only good learners but also good leaders.** Is there any difference when we use high-resource languages or low-resource languages in our training data? Our results in Table 1 show that the accuracy on high-resource language is not obviously related to whether the corresponding language data is used. More importantly, training on high-resource language data enables the model to achieve more stable improvements across multiple languages compared to that on low-resource languages (Swahili, Hindi, and Thai).

3. **A few languages can lead to spontaneous multilingual learning.** We select one, two, or three languages with English for instruction-tuning respectively. In Table 1, although using more languages sometimes leads to more stable improvements, model trained only on Chinese and English have achieved similar overall performance improvements. This is also consistent with the findings of Kew et al. (2023). The multilingual alignment improvement shows great language generalization.

4. **Our findings remain consistent across models of different parameter sizes.** We also present the average accuracy results of Qwen1.5-1.8B, Qwen1.5-4B, and Qwen1.5-14B in Table 2 to strengthen our conclusions. We find significant multilingual performance improvements across all of these models.

We have also validated our findings on the other two tasks to strengthen our conclusions, including

---

[5] https://github.com/hiyouga/LLaMA-Factory

Model	Qwen1.5-1.8B	Qwen1.5-4B	Mistral-7B	Qwen1.5-14B
base	68.35	79.52	87.07	86.27
zh/es ⇒ en	**76.13**	81.99	**90.83**	91.53
zh/de ⇒ en	74.23	82.64	90.81	**92.25**
zh/it ⇒ en	75.70	83.32	89.10	92.13
sw/hi ⇒ en	75.37	**85.32**	90.21	90.28

Table 2: Average accuracy of models of different parameter sizes on the Amazon Reviews Polarity. We highlight the best results for every model.

Qwen1.5-14B	Sentiment Analysis	NLI	Paraphrase Identification
base	86.27	66.94	76.36
zh/es ⇒ en	91.53	73.03	**82.59**
zh/de ⇒ en	**92.25**	**73.28**	82.21
zh/it ⇒ en	92.13	71.94	82.15
sw/hi ⇒ en	90.28	71.48	81.80

Table 3: Average accuracy results of Qwen1.5-14B base model and trained models on the Amazon Reviews Polarity, SNLI and PAWS across 20 different languages. The accuracy of randomly choosing is 33.33% for SNLI and 50.00% for the other two tasks. We highlight the best results for every task. Full results are reported in Appendix E.

Model	Qwen1.5-1.8B	Mistral-7B
base	68.35	87.07
ja/it ⇒ zh	73.32	90.34
sw/hi ⇒ zh	**77.00**	**90.58**
en/ja ⇒ zh	71.92	90.04

Table 4: Average accuracy on Amazon Reviews Polarity. We replace English with Chinese as the target language. We highlight the best results for each model.

Model	Amazon Polarity	SNLI
base	86.27	66.94
zh/es ⇒ en	90.38	68.72
zh/de ⇒ en	90.75	67.50
zh/it ⇒ en	90.46	67.76
sw/hi ⇒ en	90.53	65.76

Table 5: The model tested on Amazon Reviews Polarity is trained on SNLI questions. The model tested on SNLI is trained on Amazon Reviews Polarity questions.

Natural Language Inference (NLI) and Paraphrase Identification. The model needs to determine the relationship between two paragraphs of text in both of these two tasks. We conduct our experiment on SNLI for NLI task and PAWS for Paraphrase Identification task. We report the average accuracy of Qwen1.5-14B across all languages in Table 3. And we report the full results on each language in the Appendix E.

5.2 Analysis

Building upon the above results, we conduct more comprehensive observations and analyses of the model's behavior.

English is not necessary as the target language in the training data. As elaborated in Section 4, we use outputs in English uniformly for all languages in our previous experiments. English has been widely used for multilingual transfer as a pivot language (Zhu et al., 2024; Hu et al., 2023). We further investigate the case of replacing English with Chinese in training data and report the results in Table 4. Mistral and Qwen1.5 represent two

different types of LLMs (English-centric or not) respectively. From the results, we can find that using Chinese as the target language leads to the same conclusions as using English. For both of the two types of LLMs, using Chinese rather than English as the target language is also helpful for models' multilingual performance improvement, which indicates that English is not necessary as the target language in the training data.

It is not necessary but more beneficial to use the train subset corresponding to the test data as the source of translation data. Following Zhu et al. (2024), in our previous experiments, we construct the parallel translation data for instruction-tuning based on the train subset corresponding to the test dataset, which have the similar data characteristics and distributions. We further cross-test the Qwen1.5-14B trained on SNLI on Amazon Reviews Polarity and the Qwen1.5-14B trained on Amazon Reviews Polarity on SNLI. We report the results in Table 5. We can find that although the models trained on data with different distributions

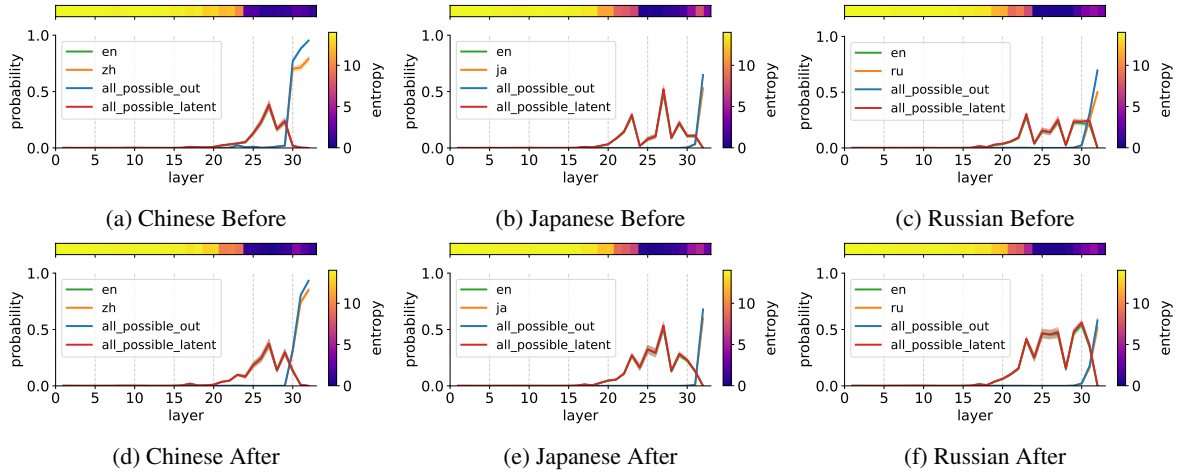| (a) Chinese Before | (b) Japanese Before | (c) Russian Before |
| (d) Chinese After | (e) Japanese After | (f) Russian After |

Figure 1: **Logit lens on Mistral-7B in Chinese, Japanese and Russian scenarios (languages not in training data).** The horizontal axes is the layer num and the vertical axes is the probability. "en" (Green covered by Red) means the latent English output corresponding to the correct answer in the target language. "zh/ja/ru" (Orange) means the correct answer in the target language. "all_possible_out" (Blue) means the probability of all possible outputs in the target language. "all_possible_latent" (Red) means all possible outputs in English.

Model	Same Language	Task-agnostic
base	76.86	50.40
zh/es $\Rightarrow$ en	83.48	**77.61**
zh/de $\Rightarrow$ en	83.69	72.28
zh/it $\Rightarrow$ en	82.33	72.32
sw/hi $\Rightarrow$ en	**84.59**	74.92

Table 6: The results of Mistral-7B on sentiment analysis task for different output types. **Same Language** means the outputs in the same language with the inputs. **Task-agnostic** means using the task-agnostic outputs.

also have better overall performance in most cases, they have a worse performance than that trained on the data corresponding to the target task. That means the multilingual data is crucial for enhancing the model's multilingual capabilities, and similar types of data is more helpful. This is consistent with the "Superficial Alignment Hypothesis" (Zhou et al., 2024), which indicates that model learns knowledge and capabilities almost entirely in pre-training process, while alignment only guides the model to utilize the different "subdistribution of formats". So the data in the same subdistribution of formats is more beneficial.

How about using outputs in different types? Except the outputs in English, we also conduct our experiments by using outputs in different types, including outputs in the same language with the inputs and task-agnostic outputs. When using outputs in the same language with the inputs, as shown in Table 6, the model also perform better after

instruction-tuning, while performing worse compared to using English outputs (shown in Table 2) under the same settings. This confirms our conclusion in Section 4 that generating content in the target language is sometimes another great challenge for LLMs except understanding and solving multilingual problems themselves.

We further replace "positive" with "ox" and replace "negative" with "horse" to investigate the cases of using task-agnostic outputs. We report the results in Table 6. Firstly, we can observe a significant decrease in multilingual performance of the base model when using task-agnostic outputs, which indicates that task-specific outputs are important for effective in-context learning (ICL). Clearly, we find a significant improvement in multilingual performance of the instruction-tuned models. By comparing the results before and after training, we can find that our training greatly improves the model's ICL capability on the specific task, and this capability improvement exhibits excellent multilingual generalization. Based on the Superficial Alignment Hypothesis, we infer that the questions in only a few languages are able to effectively activate the corresponding subdistribution of formats across a wide range of languages.

6 Mechanistic Interpretability Analysis

In this section, we further utilize methods mentioned in §2.3 to analyze the model's changes before and after the training.
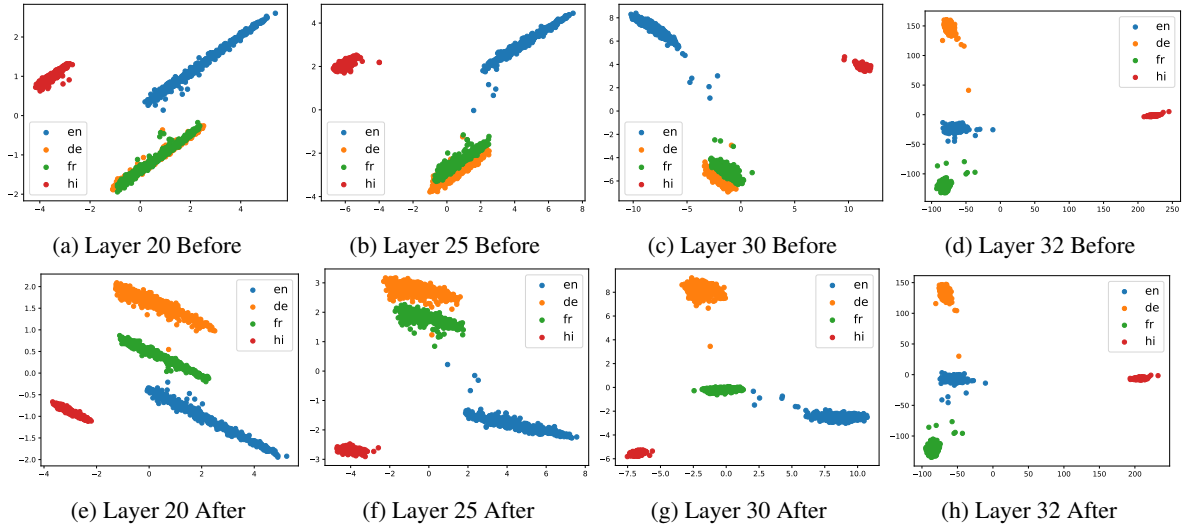
Figure 2: **PCA (Principal Component Analysis) on Mistral-7B in English, German, French and Hindi scenarios.** Before means the base model. After means the trained model. All logits are mapped into the two-dimensional representation. Each point in the plot corresponds to one instance.

6.1 Logit Lens

Following Wendler et al. (2024), we utilize *logit lens* to analyze the changes of the model. Logit lens technique uses the model head in the final layer to project the intermediate latent logits directly to the vocabulary space. We utilize logit lens on Qwen1.5, a series of LLMs that are not English-centric, and find there is not English latent outputs in the intermediate layers. And the prefix token overlapping between target language and English will also bring errors to the results. For example, the English word "positive" is written as "positivo" in Spanish, while they share the same prefix "positiv", which makes it difficult to distinguish the language of the token and accurately calculate probabilities for each language separately through logit lens. So we choose Chinese, Japanese and Russian as three representative languages for our experiment, which shows significant improvement in our results before. Following Wendler et al. (2024), we use the outputs in the same language with the inputs (Table 6). We conduct our experiments on Mistral-7B and its best trained version "sw/hi ⇒ en" in Table 6. We report the results in Figure 1.

Clearly, we can observe the following points: (1) All models generate latent English output before generating outputs in the target language finally; (2) The proportion of the probability of the correct answer increases in the sum of all possible answer probabilities; (3) The probability of all other possible answers (except correct answer) in the latent English outputs is nearly zero; (4) The area of latent

English output obviously increases, which means the trained models perform latent inference in English better and indicates better alignment.

6.2 Principal Component Analysis

We further utilize the dimensionality reduction technique to visualize the intermediate layer latent outputs of the model across different languages. PCA (Principal Component Analysis) (Pearson, 1901) can be used in some scenarios where logit lens doesn't work. Principal components are a few linear combinations of the original variables that maximally explain the variance of all the variables (Greenacre et al., 2022). We utilize PCA to map the latent logits into the two-dimensional representation. Based on the patterns shown in Figure 1, we report layer 20, 25, 30 and the last layer as four representative layers in Figure 2.

We have the following findings: (1) The points of different languages follow the similar patterns in layer 20 and layer 25, where English latent outputs have appeared and outputs in the target language haven't appeared. We further calculate the Pearson correlation coefficient of 1 dimension PCA results (Appendix C). There is a strong linear correlation between representations of different languages, which also indicates an uniform latent representation pattern during latent inference process; (2) Representations belonging to different languages exhibit greater distance from each other after training; (3) The results of the last layer is similar because of the same possible outputs; (4) Based on

the Pearson coefficient results reported in the Appendix C (Table 8 and 9), the correlation between the low-resource languages (hi, th, sw, and ms) and other high-resource languages greatly improves, which suggests better alignment with English.

7 Conclusion

In this paper, we find that LLMs only trained on question translation data without annotated answers are able to get a significant multilingual alignment improvement between English and a wide range of languages, even those unseen during instruction-tuning. We conduct the experiments on different models, different benchmarks and 20 different languages to strengthen our conclusions. Our results indicate that utilizing question alignment greatly enhances the multilingual alignment and the in-context learning capabilities of LLMs. And these improvements demonstrate the excellent model and language generalization. Furthermore, we also conduct comprehensive analysis based on some mechanistic interpretability methods, including logit lens and dimensionality reduction technique. Our work demonstrates the enormous potential of LLMs for efficient multilingual capability improvement. We hope our work can inspire the community to further explore this promising direction for the better multilingual alignment.

8 Limitations

We aim to draw more attention to the multilingual alignment which is a promising research direction. Despite our work has demonstrated LLMs' strong capabilities of multilingual generalization and the great potential of efficient multilingual alignment, there are still some limitations waiting for research. Because we investigate the models trained on the parallel question translation data in our work to eliminate the interference of the task-related data with annotated answers from our analysis of multilingual alignment, we utilize few-shot learning to help models handle the target tasks better. Analyzing LLMs' multilingual alignment in a zero-shot setting properly would further strengthen the conclusions if possible.

Due to the limited resources, we conduct experiments on different LLM scale from 1.8B to 14B in this work. We are willing to verify our conclusions on larger LLMs (70B or larger) if more resources are available in the future. Meanwhile, we mainly utilize automatic translation engine in our work be-cause of the limited resources, while data annotated by native speakers would be more accurate.

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *arXiv preprint arXiv:2104.12250*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. 2022. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*.

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

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

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pre-training and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *arXiv preprint arXiv:2404.04659*.

Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. 2022. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

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. *arXiv preprint arXiv:2305.07004*.

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*.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *arXiv preprint arXiv:2312.12683*.

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.

Yaobo Liang, Quanzhi Zhu, Junhe Zhao, and Nan Duan. 2024. Machine-created universal language for cross-lingual transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18617–18625.

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. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

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. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nostalgebraist. 2020. interpreting gpt: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024. A tree-of-thoughts to broaden multi-step reasoning across languages. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. 2022. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

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

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. *arXiv preprint arXiv:2402.10588*.

Shimao Zhang, Yu Bao, and Shujian Huang. 2024. Edt: Improving large language models' generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023a. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023b. Plug: Leveraging pivot language in cross-lingual instruction tuning. *arXiv preprint arXiv:2311.08711*.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. 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. 2024. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

A LLaMA 2 Language Distribution

Language	Percent
en	89.70%
unknown	8.38%
de	0.17%
fr	0.16%
sv	0.15%
zh	0.13%
es	0.13%
ru	0.13%
nl	0.12%
it	0.11%
ja	0.10%

Table 7: Top-10 (except unknown) lanaguage distribution in LLaMA-2's pretraining data (Touvron et al., 2023). The majority of these data is English data. And the unknown category is partially made up of programming code data.

B Additional Experimental Implementations

We construct 10k parallel data for every language pair used for training. For example, for "zh/de-en" setting of Mistral-7B, we construct a dataset including 10k Chinese-to-English translation instances and a dataset including 10k German-to-English translation instances firstly. Then we instruction-tune the Mistral-7B model only on 20k randomly mixed translation data.

We use test data and few-shot examples translated from English by Google Translate for all languages to minimize the impact of test dataset and few-shot examples themselves and ensure testing fairness across different languages. We choose the few-shot examples which are not in our training data and test data.

Additionally, we find that not only non-English inputs but also non-English outputs have significant impacts on the model's performance. For example, for Mistral-7B and sentiment analysis task, the accuracy on Hindi is 0.5 if we use outputs in Hindi, while the accuracy is 0.816 if the output is "positive" or "negative". This implies that generating content in the target language is another great challenge for LLMs, which is distinct from understanding and solving problems in the corresponding language. Considering we mainly focus on the language understanding and task solving

capabilities, we use English outputs uniformly if it is not specified.

C Pearson Correlation Coefficient Based on PCA

Layer 20	Base	Trained
en-de	0.9727	0.9752
en-fr	0.9804	0.9822
en-hi	0.9268	0.9526
de-fr	0.9825	0.9834
de-hi	0.9564	0.9707
fr-hi	0.9518	0.9674
en-th	0.8727	0.8941
en-sw	0.9501	0.9594
en-ms	0.9552	0.9620

Table 8: Pearson correlation coefficient of 1 dimension PCA results in Mistral-7B layer 20.

Layer 25	Base	Trained
en-de	0.9286	0.9275
en-fr	0.9520	0.9447
en-hi	0.5060	0.7942
de-fr	0.9628	0.9754
de-hi	0.6233	0.9102
fr-hi	0.6272	0.9001
en-th	0.6791	0.7484
en-sw	0.2316	0.8514
en-ms	0.7835	0.8448

Table 9: Pearson correlation coefficient of 1 dimension PCA results in Mistral-7B layer 25.

D Logit Lens and PCA Results for Qwen1.5

We report the logit lens and the PCA results of Qwen1.5-1.8B (total 24 layers) here. In Figure 3, we can find that as we mentioned above, while utilizing logit lens on Qwen1.5, a non-English-centric model, there is no intermediate latent output before generating the output in the target language finally. This indicates that logit lens might not be an effective tool for analyzing the non-English-centric LLMs.

We further report the PCA results in Figure 4, which also indicates a clear similar latent representation pattern for different languages in the non-English-centric LLMs' intermediate layers. This

further reinforces the significance of multilingual alignment, which also provides the basis for the success of question alignment paradigm on Qwen. The Pearson coefficient results reported in Table 10 and Table 11 show the better alignment with English, consisting with the results of Mistral-7B.

Layer 12	Base	Trained
en-de	0.9816	0.9820
en-fr	0.9848	0.9850
en-hi	0.9730	0.9742
de-fr	0.9852	0.9855
de-hi	0.9760	0.9769
fr-hi	0.9732	0.9740
en-th	0.9538	0.9547
en-sw	0.9746	0.9763
en-ms	0.9815	0.9823

Table 10: Pearson correlation coefficient of 1 dimension PCA results in Qwen1.5-1.8B layer 12.

Layer 18	Base	Trained
en-de	0.7331	0.9176
en-fr	0.6243	0.9303
en-hi	0.3623	0.8589
de-fr	0.7493	0.9477
de-hi	0.5639	0.8965
fr-hi	0.4729	0.8854
en-th	0.2881	0.4566
en-sw	-0.0424	0.8592
en-ms	0.4297	0.8648

Table 11: Pearson correlation coefficient of 1 dimension PCA results in Qwen1.5-1.8B layer 18.

E   Full Results of SNLI and PAWS

We report the complete results of SNLI and PAWS on 20 different languages in Table 12 and 13 separately. Similar to the sentiment analysis task, we can see that models instruction-tuned on multilingual translation data significantly outperform the base model, which confirms that our findings have good generalization across different tasks.

(a) English Before   (b) Chinese Before   (c) Japanese Before

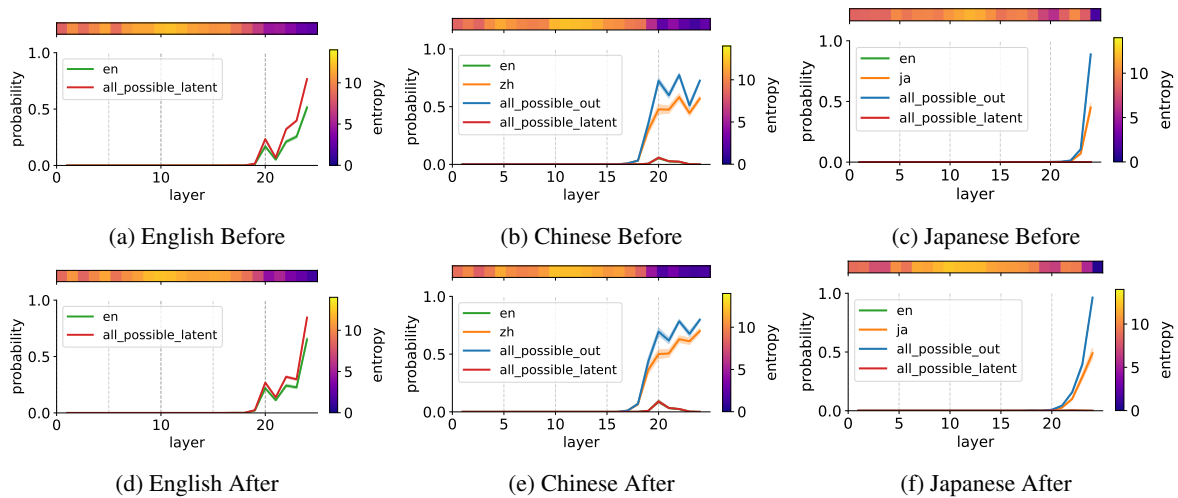(d) English After   (e) Chinese After   (f) Japanese After

Figure 3: **Logit lens on Qwen1.5-1.8B in English, Chinese, and Japanese scenarios (languages not in training data).** The horizontal axes is the layer num and the vertical axes is the probability. "en" (Green) means the latent English output corresponding to the correct answer in the target language. "en/zh/ja" (Orange) means the correct answer in the target language. "all_possible_out" (Blue) means the probability of all possible outputs in the target language. "all_possible_latent" (Red) means all possible outputs in English.

(a) Layer 6 Before   (b) Layer 12 Before   (c) Layer 18 Before   (d) Layer 24 Before

(e) Layer 6 After   (f) Layer 12 After   (g) Layer 18 After   (h) Layer 24 After
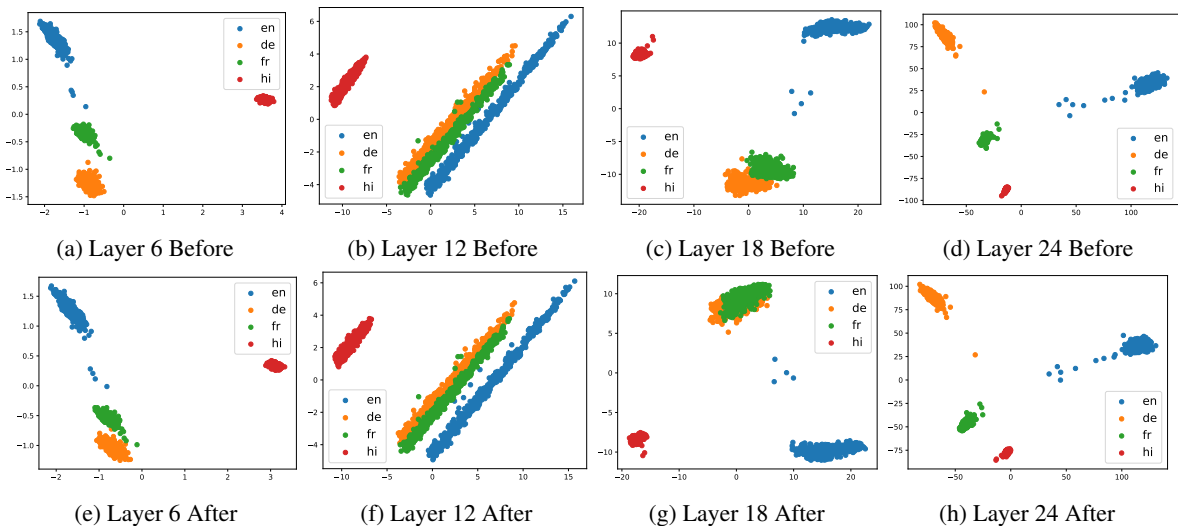
Figure 4: **PCA (Principal Component Analysis) on Qwen1.5-1.8B in English, German, French and Hindi scenarios.** Before means the base model. After means the trained model. All logits are mapped into the two-dimensional representation. Each point in the plot corresponds to one instance. **There is also a similar latent representation pattern for different languages in the intermediate layers while logit lens can't reveal it.**

8050

Qwen1.5-14B	en	zh	de	fr	es	it	nl	ja	ru	sv
base	84.50	83.50	74.17	75.17	81.17	78.67	78.17	51.17	76.83	76.17
zh/es ⇒ en	**92.50**	84.67	82.67	82.83	**85.50**	83.83	84.67	57.00	**82.67**	**84.33**
zh/de ⇒ en	91.83	84.50	**83.67**	**84.50**	85.00	83.67	84.50	**57.17**	81.67	**84.33**
zh/it ⇒ en	91.67	83.83	80.83	82.67	84.00	80.00	83.50	55.83	81.50	83.33
sw/hi ⇒ en	91.33	**85.67**	80.67	80.83	83.50	81.50	82.33	55.00	79.50	82.83

Qwen1.5-14B	sl	pl	bg	no	ms	is	hi	th	sw	bn
base	63.17	67.83	64.33	**43.00**	75.00	48.17	61.00	69.67	45.83	41.33
zh/es ⇒ en	66.00	76.83	76.33	37.50	**80.67**	**57.67**	71.33	**75.00**	**58.33**	40.33
zh/de ⇒ en	**66.83**	77.00	**78.00**	35.33	80.50	57.50	**73.00**	**75.00**	**58.33**	**43.33**
zh/it ⇒ en	65.67	**77.33**	76.17	36.67	79.00	56.00	70.50	73.67	55.33	41.33
sw/hi ⇒ en	63.00	76.67	72.17	39.67	80.33	54.17	67.67	74.67	56.83	41.33

Table 12: Accuracy of Qwen1.5-14B base model and trained models on the SNLI. We report all of the results on 20 languages. The accuracy of randomly choosing is 33.33%. We highlight the best results for every language.

Qwen1.5-14B	en	zh	de	fr	es	it	nl	ja	ru	sv
base	85.4	78.4	74.0	82.2	78.8	79.4	78.6	75.0	73.0	76.6
zh/es ⇒ en	**87.8**	**83.2**	81.2	**86.0**	**85.8**	83.8	**86.0**	**77.0**	82.6	84.4
zh/de ⇒ en	87.0	80.8	**84.8**	84.8	84.6	82.2	**86.0**	**77.0**	**83.0**	84.6
zh/it ⇒ en	87.4	80.8	80.4	85.0	83.8	**85.2**	85.2	**77.0**	82.2	**85.2**
sw/hi ⇒ en	87.2	78.4	79.6	81.8	84.6	83.2	84.4	76.6	81.0	85.0

Qwen1.5-14B	sl	pl	bg	no	ms	is	hi	th	sw	bn
base	70.2	79.8	75.2	78.6	82.0	67.0	73.2	79.4	75.4	65.0
zh/es ⇒ en	**82.6**	80.8	82.0	86.4	**86.8**	77.4	80.6	**82.0**	79.2	76.2
zh/de ⇒ en	80.8	80.2	82.0	85.8	85.8	77.0	**81.4**	81.4	80.8	74.2
zh/it ⇒ en	81.2	81.6	81.8	86.8	85.6	77.2	79.4	81.2	**81.2**	74.8
sw/hi ⇒ en	82.4	**81.8**	**82.6**	**87.0**	85.8	**78.2**	80.8	80.4	78.2	**77.0**

Table 13: Accuracy of Qwen1.5-14B base model and trained models on the PAWS. We report all of the results on 20 languages. The accuracy of randomly choosing is 50.0%. We highlight the best results for every language.