Enhancing Multilingual Capabilities of Large Language Models through Self-Distillation from Resource-Rich Languages

Anonymous ACL submission

Abstract

While large language models (LLMs) have been pre-trained on multilingual corpora, their performance still lags behind in most languages compared to a few resource-rich languages. One common approach to mitigate this issue is to translate training data from resource-rich languages into other languages and then continue training. However, using the data obtained solely relying on translation while ignoring the original capabilities of LLMs across languages is not always effective, which we show will limit the performance of cross-lingual knowledge transfer. In this work, we propose SDRRL, a method based on Self-Distillation from Resource-Rich Languages that effectively improve multilingual performance by leveraging the internal capabilities of LLMs on resource-rich languages. We evaluate on different LLMs (LLaMA-2 and SeaLLM) and source languages (English and French) across various comprehension and generation tasks, experimental results demonstrate that SDRRL can significantly enhance multilingual capabilities while minimizing the impact on original performance in resource-rich languages.

1 Introduction

Contemporary large language models (LLMs; OpenAI, 2022, 2023; Touvron et al., 2023a,b; Jiang et al., 2023; Google et al., 2023) are predominantly trained on multilingual corpora. However, the language distribution in the data is highly imbalanced. For instance, LLMs like LLaMA-2 (Touvron et al., 2023b), with English as the primary language, have also been trained on Japanese text, yet the quantity of English tokens used during pre-training exceeds that of Japanese by a factor of 897.

The imbalanced data distribution above has led to significant limitations in the capabilities of LLMs across most languages. To enhance the multilingual capabilities, a common approach follows the translating and then supervised fine-tuning

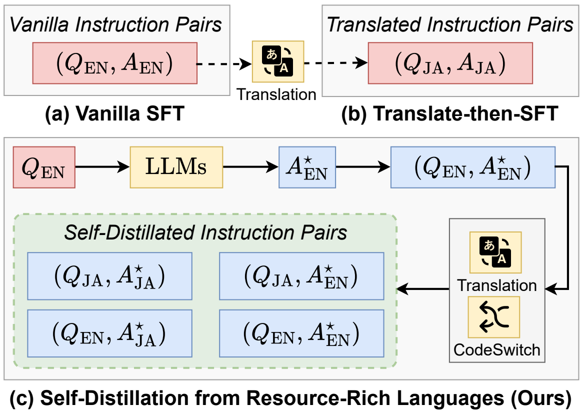

Figure 1: Comparison between vanilla supervised fine-tuning (SFT), translate-then-SFT, and our proposed method. Besides using the translated question-answer pairs in the target language (e.g., Japanese), SDRRL further leverages the generated answer $A^{\star}_{\text{EN}}$ by LLMs in the resource-rich language (e.g., English) and collects self-distillated data (in green box) to help enhance its multilingual capabilities.

(SFT; Ouyang et al., 2022) paradigm, as shown in Figure 1(b). Specifically, training data is translated into the target language using either the model itself or an external machine translation (MT) system before continuing the training process, thereby offering more data in the target language and improving multilingual capabilities.

However, the translate-then-SFT method encounters several challenges: First, the multilingual enhancement gained from translated "question-answer" pairs is limited and may sometimes even degrade the capabilities in the original primary language (Zhu et al., 2024). Second, constrained by the accuracy of machine translation (especially for the low-resource languages), the translated texts used for training can be highly noisy, containing numerous awkward sentences and incorrect content, adversely affecting the quality of the generated text and the multilingual abilities of the LLMs. Therefore, we explore a new question along this trajectory: *Besides translating the training pairs, can we enhance the abilities in other languages by leveraging the original relatively strong capabilities of LLMs in resource-rich language?*

In this paper, we introduce SDRRL, a method that uses Self-Distillation from Resource-Rich Languages) to achieve the goal mentioned above. Specifically, as illustrated in Figure 1(c), SDRRL comprises two parts: (1) *Self-Distillation*: Instead of the ground-truth answer, responses from LLMs in resource-rich languages are collected to construct a transfer set. These are then translated into other languages using machine translation systems and code-switching tools, forming "question-answer" pairs that are semantically identical but linguistically varied, and conducting sentence-level knowledge self-distillation within the same batch. (2) *Incorporating External Parallel Corpus*: We further involve a small amount of machine translation data in the distillation, aiming to align the linguistic representation spaces better and mitigate the negative impact of the noise in machine translation systems on the generative capabilities of LLMs.

Our experiments, based on LLaMA-2-7B (Touvron et al., 2023b) and SeaLLM-7B (Nguyen et al., 2023) with English as the resource-rich language, demonstrate that even with a smaller set of English instruction data as the transfer set, SDRRL can effectively distill English capabilities into 14 other languages, showing effectiveness in both multilingual comprehension and generation tasks. Further analysis indicates that SDRRL helps preserve the original capabilities in high-resource languages and improves the quality of generated responses.

2 Related Work

Multilingual Language Models. Using multilingual data during the pre-training is a common approach to enhance the multilingual capabilities of LLMs (Li et al., 2022; Lample and Conneau, 2019; Workshop et al., 2022; Lin et al., 2022; Xue et al., 2021). Despite being pre-trained and fine-tuned targeting a few resource-rich languages, recent instruction-following LLMs (Touvron et al., 2023b; Jiang et al., 2023; Wang et al., 2023a) have been found to still possess significant multilingual understanding and generation capabilities (Bandarkar et al., 2023; Niklaus et al., 2023). However, limited by the imbalanced training data distribution (Yang et al., 2023), the multilingual capabilities of these popular LLMs lag behind those of languages with abundant resources (Pahune and Chandrasekharan, 2023).

Cross-Lingual Transfer. To enhance the capabilities in languages with scarce resources, one line of work is cross-lingual transfer, where skills learned from one source language can be readily transferred to other languages (Etxaniz et al., 2023; Huang et al., 2023; Ranaldi and Zanzotto, 2023). This has been approached by designing prompts that leverage LLMs to self-translate questions into resource-rich languages (Qin et al., 2023), or by utilizing external machine translation systems for assistance (Zhao et al., 2024). Efforts have also been made to distill synthetic data from high-resource languages to low-resource ones (Chai et al., 2024). Shaham et al. (2024) and Kew et al. (2023) leverage similarities between languages to stimulate capabilities in others. Compared to their work, we focus on proficiency in the resource-rich language and leverage it to improve performance in other languages.

Cross-Lingual Alignment. Another line of work is cross-lingual alignment (Schuster et al., 2019a). Given the scarcity of multilingual data, the construction of alignment data or loss functions of varying granularity can align mid- and low-resource languages with those that are resource-rich. This includes the construction of pre-training tasks using multilingual aligned lexicons (Chi et al., 2021), alignment of word embeddings (Wen-Yi and Mimno, 2023; Schuster et al., 2019b), using aligned data on one side of a problem to improve mathematical reasoning processes (Zhu et al., 2024), and encouraging language switching in chain-of-thought (CoT; Wei et al., 2022) reasoning (Chai et al., 2024). Mao and Yu (2024a) have leveraged the LLM's own capabilities to generate aligned data, while others have constructed it with the aid of external systems (Ranaldi and Pucci, 2023; Chen et al., 2023). Deriving and constructing multilingual supervision signals from existing datasets overlooks the fact that the model's own responses in high-resource languages can also serve as effective supervision signals. We show in our experiments that self-distillation not only improves the LLM's multilingual performance but also helps maintain the performance in the original resource-rich languages.

Knowledge Distillation. Knowledge distillation (Hinton et al., 2015) is a widely used method for transferring knowledge (Gou et al., 2021). In the text generation domain, sequence-level knowledge distillation (Kim and Rush, 2016) has been used as a means of data augmentation in areas such as machine translation (Gordon and Duh, 2019). In particular, *self-distillation* (Zhang et al., 2019, 2022b; Pham et al., 2022) is often utilized to distill knowledge from one component of a model to another (Zhang et al., 2022a), or from one stage of a model to another (Yang et al., 2019). In this work, we apply distilling knowledge between the different linguistic representation spaces within the same LLM to enhance multilingual capabilities.

3 Method

In this section, we first revisit the supervised fine-tuning (SFT) and translate-then-SFT paradigm, subsequently dividing the discussions into two parts of our proposed SDRRL. In the first part, we construct a transfer set using responses in the resource-rich language from LLMs through sentence-level self-distillation. In the second part, we employ parallel translation-based instruction data to further improve multilingual generation capabilities.

3.1 SFT and Translate-then-SFT Paradigm

We consider the given instruction dataset comprised of N entries $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i$ symbolizes the input sentence (question) for the i-th data point, and $\mathbf{y}_i$ signifies the corresponding ground-truth response (answer).

Supervised Fine-Tuning. For a LLM $\mathcal{M}_\theta$ parameterized by a set of parameters θ, which produces a response denoted as $\hat{\mathbf{y}} = \mathcal{M}_\theta(\mathbf{x})$ for the given input question $\mathbf{x}$, the objective of SFT is to align the output sentence $\hat{\mathbf{y}}$ as closely as possible with the ground-truth response $\mathbf{y}$. Specifically, the cross-entropy (CE) loss is employed to assess the discrepancy between the model output $\hat{\mathbf{y}}$ and the ground-truth output $\mathbf{y}$ for a single sample $(\mathbf{x}, \mathbf{y})$, defined as:

$$\ell_{\text{CE}}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{j=1}^{|\mathcal{V}|} y_j \log(\hat{y}_j) \tag{1}$$

where y_j is the one-hot encoding of the ground truth output $\mathbf{y}$ at position j, $\hat{y}_j$ is the probability of the model output $\hat{\mathbf{y}}$ at position j, and $|\mathcal{V}|$ is the size of the vocabulary in the LLM.

For the entire dataset $\mathcal{D}$, the total loss is calculated as the average of all sample losses:

$$\mathcal{L}_{\text{SFT}} = \frac{1}{N}\sum_{i=1}^{N} \ell_{\text{CE}}(\mathbf{y}_i, \mathcal{M}_\theta(\mathbf{x}_i)) \tag{2}$$

Translate-then-SFT. For the translation-then-SFT paradigm, we define the machine translation system as a function $\mathcal{T}$, which accepts text in one language as the source language (`Src`) and outputs equivalent text in the target language (`Tgt`). Using the machine translation system $\mathcal{T}$, each pair $(\mathbf{x}_i, \mathbf{y}_i)$ is translated into the target language, resulting in the translated dataset $\mathcal{D}^{\text{MT}} = \{(\mathbf{x}_i^{\text{MT}}, \mathbf{y}_i^{\text{MT}})\}_{i=1}^N = \{\mathcal{T}(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$.

Similar to Eq. 1, the LLM $\mathcal{M}_\theta$ is then trained on the translated dataset $\mathcal{D}'$, where the loss for a single sample $(\mathbf{x}^{\text{MT}}, \mathbf{y}^{\text{MT}})$ is computed as:

$$\ell_{\text{CE}}(\mathbf{y}^{\text{MT}}, \hat{\mathbf{y}}^{\text{MT}}) = -\sum_{j=1}^{|\mathcal{V}|} y_j^{\text{MT}} \log(\hat{y}_j^{\text{MT}}) \tag{3}$$

where $\hat{\mathbf{y}}^{\text{MT}} = \mathcal{M}_\theta(\mathbf{x}^{\text{MT}})$ is the response of models to the question $\mathbf{x}^{\text{MT}}$ in target language.

3.2 Self-Distillation from Resource-Rich Languages (SDRRL)

LLMs exhibit superior comprehension and generation capabilities in resource-rich languages, which we suppose can be a learning reference for other languages to enhance the multilingual capabilities of LLMs. To achieve this, we propose sentence-level knowledge distillation from resource-rich language responses. The core motivation is that the responses of LLMs in the resource-rich language serve as samples from the resource-rich language representation space. By adding these responses and their translations to the transfer set, the gap for cross-linguistic learning is reduced, facilitating the improvement of multilingual capabilities.

3.2.1 Transfer Set Construction

We construct a transfer set for sentence-level distillation *by collecting LLM responses in the resource-rich language*. For the original instruction dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, LLM $\mathcal{M}_\theta$ generates responses for each question $\mathbf{x}_i$, yielding $\hat{\mathbf{y}}_i = \mathcal{M}_\theta(\mathbf{x}_i)$, then we get the generated dataset $\mathcal{G} = \{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}_{i=1}^N = \{(\mathbf{x}_i, \mathcal{M}_\theta(\mathbf{x}_i))\}_{i=1}^N$. The synthesized transfer set $\mathcal{D}_{\text{synth}}$ is obtained by equally probable random sampling from both datasets $\mathcal{D}$ and $\mathcal{G}$:

$$\mathcal{D}_{\text{synth}} = \text{Sample}(\mathcal{D}) \cup \text{Sample}(\mathcal{G}) \tag{4}$$

3.2.2 Transfer Set Translation

The above constructed transfer set $\mathcal{D}_{\text{synth}}$ contains question $\mathbf{x}_i$, ground-truth answer $\mathbf{y}_i$, and response $\hat{\mathbf{y}}_i$ by LLM $\mathcal{M}_\theta$. We consider translating them into the target language using the machine translation system $\mathcal{T}$, resulting in $\mathbf{x}_i^{\mathrm{MT}} = \mathcal{T}(\mathbf{x}_i), \mathbf{y}_i^{\mathrm{MT}} = \mathcal{T}(\mathbf{y}_i)$, and $\hat{\mathbf{y}}_i^{\mathrm{MT}} = \mathcal{T}(\hat{\mathbf{y}}_i)$. Moreover, we use WMT22-cometkiwi-da (Rei et al., 2022b) as a reference-free metric to assess the translation quality where the translation quality with scores below a threshold $\tau = 0.8$ is rejected.

In particular, four sub-datasets are generated, each containing different language combinations of questions and responses:

- $\mathcal{D}_{\mathrm{LL}}$: Both the questions and responses remain in the resource-rich language, *i.e.*, $\{\mathbf{x}_i, \mathbf{y}_i\}$ or $\{\mathbf{x}_i, \hat{\mathbf{y}}_i\}$.
- $\mathcal{D}_{\mathrm{TL}}$: The questions are translated into the target language, while responses remain in the resource-rich language, *i.e.*, $\{\mathcal{T}(\mathbf{x}_i), \mathbf{y}_i\}$ or $\{\mathcal{T}(\mathbf{x}_i), \hat{\mathbf{y}}_i\}$.
- $\mathcal{D}_{\mathrm{LT}}$: The questions remain in the resource-rich language, while responses are translated into the target language, *i.e.*, $\{\mathbf{x}_i, \mathcal{T}(\mathbf{y}_i)\}$ or $\{\mathbf{x}_i, \mathcal{T}(\hat{\mathbf{y}}_i)\}$.
- $\mathcal{D}_{\mathrm{TT}}$: Both the questions and responses are translated into the target language, *i.e.*, $\{\mathcal{T}(\mathbf{x}_i), \mathcal{T}(\mathbf{y}_i)\}$ or $\{\mathcal{T}(\mathbf{x}_i), \mathcal{T}(\hat{\mathbf{y}}_i)\}$.

This approach, by providing semantically identical but linguistically diverse samples, aids in the implicit alignment of language representation spaces, enhancing unified multilingual performance. Furthermore, $\mathcal{D}_{\mathrm{TL}}$ and $\mathcal{D}_{\mathrm{LT}}$ enhance LLM's cross-linguistic generative capabilities, helping mitigate off-target issues in target language generation.

3.2.3 Applying Code-Switching

Through the aforementioned machine translation process, we achieve alignment in sentence level (*i.e.*, the sentence of question-answer pairs). Additionally, token-level alignment is introduced using a code-switching tool, applied only to the question components $\mathbf{x}_i$ of $\mathcal{D}_{\mathrm{LL}}$, $\mathcal{D}_{\mathrm{TL}}$, $\mathcal{D}_{\mathrm{LT}}$, and $\mathcal{D}_{\mathrm{TT}}$ to increase language diversity without compromising generative capabilities.

Specifically, given $\mathbf{x}_i$ composed of a sequence of tokens $\mathbf{x}_i = x_{i,1}, x_{i,2}, \ldots, x_{i,n}$, where $x_{i,k}$ denotes the k-th token in question $\mathbf{x}_i$ (similarly for $\hat{\mathbf{x}}_i^{\mathrm{MT}}$), the code-switched version $x_{i,k}$ for each token is generated by applying the rule:

$$x_{i,k} = \begin{cases} \mathrm{Dict}(x_{i,k}) & \text{with probability } p; \\ x_{i,k} & \text{with probability } 1-p, \end{cases} \tag{5}$$

where each token $x_{i,k}$ in $\mathbf{x}_i$ is replaced by its corresponding token in the bilingual dictionary for code-switching $\mathrm{Dict}(x_{i,k})$ with a $p = 0.15$ probability if $x_{i,k}$ is found in the bilingual dictionary. Responses, either in the source language $\mathbf{y}_i$ (similarly for $\hat{\mathbf{y}}_i$) or the target language $\mathbf{y}_i^{\mathrm{MT}}$ (similarly for $\hat{\mathbf{y}}_i^{\mathrm{MT}}$), remain unchanged.

3.2.4 Incorporating External Parallel Corpus

The Template for Constructing $\mathcal{D}_{\text{mt}}$ and $\mathcal{D}_{\text{comp}}$
Construct Data for Machine Translation Question: Translate the following sentence from *English* to *Indonesian*. *The quick brown fox jumps over the lazy dog.* Answer: *Sang rubah cokelat cepat melompati anjing malas.* # Construct Data for Sentence Completion Question: Complete the following sentence in *Indonesian* according to its context. *Sang rubah cokelat cepat* Answer: *Sang rubah cokelat cepat melompati anjing malas.*

Table 1: The template for constructing $\mathcal{D}_{\text{mt}}$ and $\mathcal{D}_{\text{comp}}$ with Indonesian-English as an example. $\mathcal{D}_{\text{mt}}$ includes bidirectional translations. $\mathcal{D}_{\text{comp}}$ contains only the target language sentences, which are split at random positions.

The target language sequences $\hat{\mathbf{y}}_i$ synthesized by the external machine translation system $\mathcal{T}$ may contain low-quality translations, thereby introducing a significant amount of noise into the knowledge distillation transfer dataset $\mathcal{D}_{\text{synth}}$.To mitigate the impact of noise on the multilingual generative capabilities of LLMs, we leverage a tiny external parallel corpus $\mathcal{P} = \{(\mathbf{s}_i, \mathbf{t}_i)_{i=1}^{L}\}$ between the resource-rich language `Src` and the target language `Tgt`. Based on the templates in Table 1, we can construct two parts of instruction data: machine translation task instructions ($\mathcal{D}_{\text{mt}}$) and sentence completion task instructions ($\mathcal{D}_{\text{comp}}$). By incorporating these two parts, the transfer set includes non-synthetic natural target language texts, which helps improve the generative quality of LLMs in the target language.

3.2.5 Training Objective

The final training dataset $\mathcal{D}_{\text{train}}$ includes $\mathcal{D}_{\mathrm{LL}}$, $\mathcal{D}_{\mathrm{TL}}$, $\mathcal{D}_{\mathrm{LT}}$, $\mathcal{D}_{\mathrm{TT}}$, $\mathcal{D}_{\text{mt}}$, and $\mathcal{D}_{\text{comp}}$. The total loss function is defined as:

$$\mathcal{L}_{\text{SDRRL}} = \sum_{d \in \mathcal{U}} \frac{1}{|\mathcal{D}_d|} \sum_{\{\mathbf{x},\mathbf{y}\} \in \mathcal{D}_d} \ell_{\text{CE}}(\mathcal{M}_\theta(\mathbf{x}), \mathbf{y}), \tag{6}$$

where $\mathcal{U} = \{\text{LL}, \text{TL}, \text{LT}, \text{TT}, \text{mt}, \text{comp}\}$ and $\mathcal{D}_d$ corresponds to the respective data subset (*e.g.*, $\mathcal{D}_{\text{LL}}$, $\mathcal{D}_{\text{TL}}$, *etc.*).

4 Experiments

4.1 Setup

We use LLaMA-2-7B (Touvron et al., 2023b) as the base model. Drawing from the distribution of language in pre-training corpus, we use English (ENG) as a resource-rich language and conduct experiments on 14 languages: Czech (CES), Danish (DAN), Ukrainian (UKR), Bulgarian (BUL), Finnish (FIN), Hungarian (HUN), Norwegian (NOB), Indonesian (IND), Japanese (JPN), Korean (KOR), Portuguese (POR), Slovenian (SLV), Vietnamese (VIE), and Polish (POL). Stanford Alpaca instruction data (Taori et al., 2023) serve as the base of the transfer set $\mathcal{D}$, providing questions and ground-truth answers in English. For machine translation, we utilize open-source NLLB-200-3.3B (Costa-jussà et al., 2022) model. To improve translation quality, we follow Zeng et al. (2021) to filter low-quality translations and use CLD3 (Ooms, 2024) model to remove off-target translations. We also follow Lin et al. (2021) to construct bilingual dictionaries for code-switching. See appendix A for more details.

Implementation Details Our code is implemented using DeepSpeed (Rasley et al., 2020) on eight NVIDIA A800-SXM4-80GB GPUs. Following Wang et al. (2023a), we set the training duration to four epochs with an automatically calculated learning rate and employ early stopping. Other hyperparameters are set according to Hiyouga (2023).

Baselines. For comparison, we consider the following baseline systems that enhance LLaMA-2's multilingual capabilities using different instruction fine-tuning methods:

- **SFT** (Ouyang et al., 2022): It only involves English instruction datasets in the process of fine-tuning, which is not multilingual-oriented.
- **Translate-then-SFT** (Chen et al., 2023, T-SFT): It uses an external machine translation system to translate English instruction data into non-English languages and construct multilingual data for instruction fine-tuning.
- **Cross-Lingual Instruction Tuning** (CIT; Li et al., 2023a): It constructs cross-lingual instructions for fine-tuning, imposing models to respond in the target language given the source language as context.
- **Cross-Lingual Chain-of-Thought Reasoning** (XCOT; Chai et al., 2024): It applies code-switching to multilingual instruction training data, using high-resource instruction data to supervise the training of low-resource languages with cross-lingual distillation.

Datasets. We evaluate the multilingual capabilities of LLMs on four representative datasets:

- **BELEBELE** (Bandarkar et al., 2023): A widely used language understanding dataset covering 122 languages, where each question, linked to a short passage, has four multiple-choice answers. This dataset proves challenging for state-of-the-art LLMs. Accuracy is reported in our experiments.
- **FLORES** (Goyal et al., 2022): A benchmark for machine translation with parallel text from Wikipedia for 204 languages, making up over 40,000 directions. We evaluate the bidirectional translation results between the target language and English, reporting scores using SacreBLEU (Post, 2018) and COMET score using WMT22-comet-da model (Rei et al., 2022a).
- **XL-SUM** (Hasan et al., 2021): A multilingual abstractive summarization benchmark for 44 languages, comprising multiple long news texts requiring summarization into a single sentence. ROUGE-1 and ROUGE-L F1 scores are reported.
- **MKQA** (Longpre et al., 2020): An open-domain question-answering dataset across 26 diverse languages, providing multiple possible short answers as ground truth for each question. We use the official evaluation script and report token overlapped F1 scores.

4.2 Main Results

Table 2 shows the experimental results of the multilingual understanding task. Tables 3, Table 4 and

	CES	DAN	UKR	BUL	FIN	HUN	NOB	IND	JPN	KOR	POR	SLV	VIE	POL	AVG.
Performance on Target Language															
SFT	49.33	48.33	46.67	49.11	39.78	43.22	49.22	46.15	42.01	37.99	55.98	42.79	42.91	44.69	45.58
T-SFT	48.22	51.67	47.11	51.22	47.11	45.67	51.33	49.72	41.56	43.69	56.20	46.03	47.60	48.72	48.28
CIT	50.11	53.44	47.22	51.44	48.00	45.67	53.33	49.94	43.24	46.26	56.65	46.70	45.59	49.72	49.09
XCOT	51.56	54.22	47.83	52.78	47.00	45.67	51.33	49.16	43.02	46.15	56.42	46.48	46.82	48.49	49.07
SDRRL	**52.11**	**55.00**	**48.33**	**54.00**	**49.56**	**46.44**	**53.89**	**52.40**	**45.81**	**46.82**	**57.88**	**47.26**	**48.38**	**50.17**	**50.58**
Performance on English Language															
SFT	65.39	65.39	65.39	65.39	65.39	65.39	65.39	65.39	**65.39**	65.39	65.39	65.39	65.39	65.39	65.39
T-SFT	63.91	65.25	66.03	65.25	65.70	65.36	65.25	65.70	61.01	60.45	63.80	65.47	65.47	65.92	64.61
CIT	63.46	65.47	65.59	64.02	61.23	63.46	64.13	65.92	62.01	63.46	64.02	63.24	62.91	62.91	63.70
XCOT	65.70	65.47	66.15	66.48	65.81	65.70	66.55	64.92	63.24	65.43	62.46	66.50	63.91	66.37	65.34
SDRRL	**66.26**	**65.70**	**67.15**	**66.53**	**65.92**	**66.70**	**66.59**	**67.15**	65.13	**65.45**	**66.48**	**66.59**	**65.57**	**66.82**	**66.29**

Table 2: Results of baselines and our SDRRL on BELEBELE benchmark. In each column, the best result is **in bold** and the second best result is underlined.

	CES	DAN	UKR	BUL	FIN	HUN	NOB	IND	JPN	KOR	POR	SLV	VIE	POL	AVG.
BLEU scores on X-to-English Tasks															
SFT	34.66	42.57	34.17	33.91	26.76	28.15	38.34	20.78	7.56	3.15	33.25	11.94	16.01	15.31	24.75
T-SFT	32.63	32.21	31.13	31.05	23.53	24.18	27.44	23.38	7.82	7.68	33.03	14.36	19.63	19.43	23.39
CIT	26.54	29.88	24.25	26.66	21.51	21.24	30.21	29.02	6.00	7.58	34.46	16.57	25.84	19.19	22.78
XCOT	31.52	31.26	29.90	31.05	24.37	23.60	32.50	27.33	8.29	9.23	35.86	17.82	25.46	19.40	24.83
SDRRL	**36.38**	**45.71**	**35.33**	**37.49**	**30.80**	**31.62**	**40.88**	**30.93**	**15.42**	**12.20**	**39.81**	**21.15**	**28.68**	**22.52**	**30.64**
BLEU scores on English-to-X Tasks															
SFT	13.00	21.91	11.18	12.98	8.39	9.07	18.53	34.54	17.03	18.15	43.06	28.46	25.06	27.65	20.64
T-SFT	22.68	27.78	23.11	27.59	15.31	16.96	25.60	31.79	19.52	18.11	39.75	26.17	25.09	26.04	24.68
CIT	22.03	28.57	19.92	26.85	14.54	17.46	25.97	29.46	13.81	15.33	35.24	22.60	22.33	22.84	22.64
XCOT	23.11	32.20	21.97	27.33	15.80	17.38	25.96	30.33	9.31	15.13	38.04	26.56	25.43	26.03	23.90
SDRRL	**27.91**	**39.00**	**27.25**	**33.93**	**20.88**	**22.09**	**29.64**	**35.32**	**20.51**	**20.47**	**43.36**	**30.09**	**29.87**	**27.86**	**29.16**
COMET scores on X-to-English Tasks															
SFT	85.35	87.60	84.58	84.97	85.69	84.40	86.35	73.54	63.41	45.44	78.91	80.98	63.43	68.46	76.65
T-SFT	84.71	84.26	83.33	83.82	83.78	82.02	83.31	78.94	78.39	72.95	80.38	81.82	73.81	79.06	80.76
CIT	81.71	82.84	80.06	81.72	82.14	82.29	83.37	84.62	76.16	73.88	83.71	76.38	80.60	78.97	80.60
XCOT	84.40	84.47	83.11	83.90	84.67	81.96	84.68	83.50	78.83	75.66	83.23	76.48	79.46	78.75	81.65
SDRRL	**86.04**	**88.51**	**84.82**	**86.08**	**86.98**	**85.70**	**87.15**	**89.46**	**83.33**	**79.02**	**85.15**	**84.02**	**81.43**	**83.08**	**85.06**
COMET scores on English-to-X Tasks															
SFT	57.19	70.93	55.25	54.99	60.29	53.94	71.97	83.82	82.46	82.14	84.57	55.96	80.78	82.14	69.75
T-SFT	78.94	81.34	79.92	81.43	78.53	76.01	82.69	84.90	82.76	80.58	86.42	69.06	81.62	82.86	80.50
CIT	79.87	82.47	78.63	81.70	78.39	76.18	83.19	84.18	78.15	78.45	85.12	80.12	80.77	80.17	80.53
XCOT	79.22	83.29	79.16	80.86	78.63	75.51	83.30	84.68	74.27	77.31	86.01	78.65	82.24	82.72	80.42
SDRRL	**84.29**	**86.91**	**83.51**	**85.40**	**84.62**	**81.06**	**85.55**	**86.00**	**83.65**	**82.66**	**87.64**	**82.63**	**83.93**	**83.61**	**84.39**

Table 3: Results of baselines and our SDRRL on FLORES benchmark.

Table 5 show the results on multilingual generation tasks. From the experimental results, we can observe that:

(1) **SDRRL effectively enhances performance in the target languages.** Specifically, in every comprehension and generation task, our method surpasses the baselines in almost all target languages. As shown in Table 2, SDRRL improves [erformane in comprehension tasks by approximately +1.5 BLEU score. On the Flores dataset, SDRRL yields up to about +6.0 BLEU score improvement in both directions and about +4.0 COMET score improvement (Table 3). This demonstrates that using proficient responses in resource-rich languages as supervisory signals for knowledge distillation significantly enhances performance in other target languages.

(2) **SDRRL exhibits stronger robustness in generation tasks.** For example, on the XL-SUM dataset (Table 4), which requires the generation of longer texts, the average performance of CIT and XCOT decreased due to the quality of machine-translated texts and pipeline noise, yet SDRRL still achieved about +0.55 ROUGE-L F1 score improvement. On the FLORES dataset (Table 3), which requires cross-lingual text generation, T-SFT and

	IND	JPN	KOR	POR	VIE	UKR	AVG.
	Performance on Target Language (ROUGE-1)						
SFT	20.82	6.17	0.66	23.38	9.30	8.10	11.41
T-SFT	25.61	32.11	7.67	26.68	20.59	14.19	21.14
CIT	24.64	16.11	5.80	26.33	20.55	11.21	17.44
XCOT	22.55	32.39	7.26	26.21	19.84	13.38	20.27
SDRRL	**26.08**	**33.15**	**8.18**	**27.40**	**20.98**	**14.35**	**21.69**
	Performance on Target Language (ROUGE-L)						
SFT	16.03	4.13	0.61	15.84	7.21	6.72	8.42
T-SFT	20.15	**22.83**	6.93	18.41	15.18	11.73	15.87
CIT	19.02	11.22	5.14	18.06	14.91	9.00	12.89
XCOT	17.19	22.32	6.52	18.05	14.57	10.78	14.91
SDRRL	**20.47**	22.81	**7.35**	**19.09**	**15.52**	**11.84**	**16.18**
	Performance on English Language (ROUGE-1)						
SFT	26.35	26.35	26.35	26.35	26.35	26.35	26.35
T-SFT	27.49	26.89	26.68	27.28	26.42	26.75	26.92
CIT	27.84	27.40	26.57	27.39	27.17	24.83	26.87
XCOT	26.44	25.45	25.43	26.78	26.00	25.90	26.00
SDRRL	**28.18**	**27.73**	**27.44**	**27.57**	**27.52**	**27.23**	**27.61**
	Performance on English Language (ROUGE-L)						
SFT	18.68	18.68	18.68	18.68	18.68	18.68	18.68
T-SFT	19.64	19.11	18.94	19.54	18.73	19.01	19.16
CIT	19.93	19.56	18.81	19.56	19.34	17.43	19.11
XCOT	18.63	17.83	17.86	18.91	18.29	18.15	18.28
SDRRL	**20.25**	**19.88**	**19.56**	**19.69**	**19.66**	**19.44**	**19.75**

Table 4: Results of baselines and our SDRRL on XL-SUM benchmark on the target language and English.

	NOB	DAN	FIN	HUN	JPN	KOR	POR	VIE	POL	AVG.
	Performance on Target Language									
SFT	37.30	38.28	37.30	35.21	32.80	33.18	39.29	37.50	37.50	36.48
T-SFT	39.73	39.59	38.95	38.60	33.96	33.90	39.93	38.71	38.14	37.95
CIT	40.18	39.94	37.94	38.40	33.50	**34.24**	39.86	39.94	38.84	38.09
XCOT	39.03	39.28	38.12	35.60	33.07	33.69	39.96	39.49	38.49	37.41
SDRRL	**40.64**	**40.92**	**39.71**	**39.02**	**39.51**	34.06	**41.12**	**40.02**	**39.45**	**39.38**
	Performance on English Language									
SFT	41.62	41.62	41.62	41.62	41.62	41.62	41.62	41.62	41.62	41.62
T-SFT	44.92	42.63	44.24	44.21	41.65	42.11	42.63	42.65	42.81	43.09
CIT	44.09	43.86	43.55	44.12	42.83	43.29	42.51	42.52	43.41	43.35
XCOT	43.23	43.16	43.53	43.06	42.59	42.58	43.39	42.58	43.29	43.05
SDRRL	**45.42**	**45.33**	**45.47**	**44.78**	**43.26**	**43.58**	**43.99**	**45.77**	**44.71**	**44.70**

Table 5: Results of baselines and our SDRRL on MKQA dataset on the target language and English.

CIT lead to decrease of -1.36 and -2.08 BLEU scores, respectively, while our method improves by +5.88 BLEU scores. This suggests that adding machine-translated data constructed instructions to the self-distillation process effectively improves multilingual generation and mitigates the negative impact of low-quality translated texts.

(3) **SDRRL can maintain the original strong capabilities in English.** The results show that it is more challenging to retain the original English capabilities for languages with unique alphabets (*e.g.*, Japanese and Korean). For example, in the Japanese comprehension task (Table 2), all baseline methods lead to a performance drop in English compared to SFT, while only our method successfully preserving the original English capabilities.

		NLU AVG.		NLG AVG.	
		TAR.	ENG	TAR.	ENG
1	Full Method	**50.58**	**66.29**	**28.24**	**31.69**
2	- $\mathcal{D}_{\rm TL}$ and $\mathcal{D}_{\rm LT}$	49.56	65.93	26.15	30.55
3	- $\mathcal{D}_{\rm synth}$ + $\mathcal{D}$	48.59	65.10	25.16	30.10
4	- $\mathcal{D}_{\rm mt}$ and $\mathcal{D}_{\rm comp}$	50.41	66.01	26.61	30.19
5	- Code Switching	50.37	65.94	27.13	30.69
6	Only $\mathcal{D}_{\rm mt}$ and $\mathcal{D}_{\rm comp}$	41.25	61.61	17.89	22.28

Table 6: Ablation study. Average scores of target language (TAR.) and English (ENG) on natural language understanding task (NLU, including BELEBELE) and natural language generation tasks (NLG, including FLORES, XL-SUM ROUGE-L, and MKQA) are reported.

4.3 Ablation Study

We further investigate the effectiveness of each component of our proposed SDRRL. The results are shown in Table 6, where average scores on natural language understanding and generation tasks are reported. Our observations include:

(1) Rows 1 to 5 demonstrate that removing any single component results in performance degradation, affirming the necessity and efficacy of each component in SDRRL.

(2) Insights from row 3 suggest a significant performance decline in both target languages and English when model-generated responses ($\hat{\mathbf{y}}_i$) are removed from $\mathcal{D}_{\rm synth}$, highlighting the effectiveness of utilizing responses in resource-rich languages as additional supervision signals for improving multilingual capabilities. Moreover, row 2 indicates that substituting sentences with their semantic counterparts in different languages also contributes to multilingual performance improvement.

(3) Row 4 and 5 reveal that $\mathcal{D}_{\rm mt}$, $\mathcal{D}_{\rm comp}$, and code-switching provide a limited amount of ground truth. This additional supervisory signal is beneficial for generative tasks and helps improve the quality of responses.

(4) Despite introducing a small amount of parallel data through $\mathcal{D}_{\rm mt}$ and $\mathcal{D}_{\rm comp}$, as shown in row 6, relying solely on these additional data for LLM training supervision leads to severe performance degradation. Compared to row 4, this indicates that these data do not inherently bring positive performance gains but are used to mitigate the deterioration of the LLM's multilingual generative representation space caused by noisy machine-translated text, serving as a regularization mechanism in knowledge distillation.

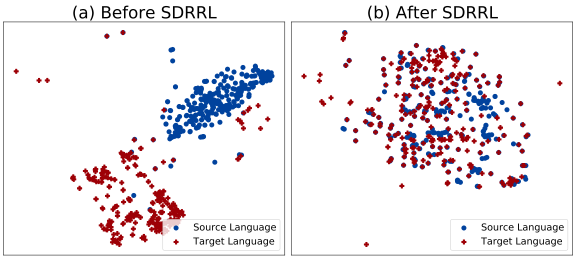

Figure 2: t-SNE visualizations of output representations by LLaMA-2 before and after applying SDRRL. The markers in red and blue represent semantically equivalent instructions in different languages.

4.4 Visualization of Representation Space for Source and Target Langauges

We visualize the sentence representations of input instructions to investigate the effect of SDRRL on the multilingual representation space. Following common practices in sequence classification Li et al. (2023c), we input instructions into the LLaMA-2 and use the last hidden states of the last token as the vector representation of the sentence. We then apply t-SNE (Van der Maaten and Hinton, 2008) to reduce the 4096-dim representations to 2-dim for visualization.

As shown in Figure 2, after applying SDRRL, the representations of semantically equivalent instructions in the source and target languages are drawn closer together. This implies that SDRRL has improved the multilingual representation space by aligning the representation space of the target language closer to that of the resource-rich, better-modeled source language, thereby enhancing the performance in target languages.

4.5 SeaLLM as Different Backbone Model

By using the responses of LLMs in high-resource languages as the supervisory signal for knowledge distillation, SDRRL is applicable to various LLMs, not limited to LLaMA-2. In this part, we conduct experiments on SeaLLM-7B (Nguyen et al., 2023), a specialized language model optimized for Southeast Asian languages.

As shown in Table 7, SDRRL results in an improvement of +2.39 average scores on the target languages. In English, SDRRL maintains its original performance, while the baselines exhibit a performance drop of at least -2.02 average scores compared to vanilla SFT. This demonstrates the generalizability of SDRRL in different LLMs. See appendix B for detailed results on more datasets.

	BELE.	XL-SUM	FLORES	MKQA	AVG.
Performance on Target Language					
SFT	42.24	16.48	18.45	38.86	29.01
T-SFT	42.77	15.32	16.59	43.40	29.52
CIT	42.53	15.75	20.49	43.70	30.62
XCOT	41.19	15.79	17.21	42.04	29.06
SDRRL	**43.67**	**17.89**	**25.86**	**44.63**	**33.01**
Performance on English Language					
SFT	60.19	15.25	28.49	39.62	35.89
T-SFT	58.70	15.63	23.72	37.43	33.87
CIT	58.66	15.42	18.31	36.67	32.27
XCOT	57.73	14.90	23.96	37.94	33.63
SDRRL	**60.67**	**16.24**	**29.47**	**40.32**	**36.68**

Table 7: Results of baselines and our SDRRL on SeaLLM. The average scores across various datasets are reported, and full results are available in appendix B.

4.6 Further Analysis

Non-English Source Languages. SDRRL is also capable of transferring multilingual performance using other source languages in high-resource. In appendix C, we opt for experiments with French instead of English. Experimental results reveal that, despite the LLM and the machine translation system exhibiting stronger performance in English, SDRRL still achieves positive distillation gains with French as the source language.

Case Study. In appendix D, we provide several case studies to offer deeper insights into the impact of SDRRL on the generation capabilities of LLMs. It is observed that the SDRRL process is able to alleviate off-target issues in the target language, reduce grammatical errors and hallucinations, and enhance the fluency of the output text.

5 Conclusion and Future Work

We introduce Self-Distillation from Resource-Rich Languages (SDRRL) to enhance the multilingual capabilities of LLMs. SDRRL uses the model itself to generate high-quality responses in resource-rich source languages and their target language counterparts as supervision signals for knowledge distillation, aiming to align additional target languages with resource-rich languages. We conduct comprehensive experiments across 16 languages on LLaMA-2 and SeaLLM. The results demonstrate that, compared to various baselines, our method significantly enhances the performance of target languages while preserving the capabilities of source languages. This highlights the multilingual potential of LLMs and illuminates paths for further research into multilingual LLMs.

Limitations

Firstly, within our method pipeline, some components are interchangeable. For example, our approach relies on external machine translation systems to provide translations in the target language, while future research could explore self-translation with LLMs that achieve great low-resource translation capabilities, thereby simplifying the process. Additionally, our method uses a small amount of machine-translated parallel corpus to construct the transfer set, but employing monolingual texts in the target language represents a promising research direction. Secondly, our experiments are conducted with only a single source language and target language. Subsequent research could investigate using a mix of multiple languages as both source and target languages and explore the mutual influences between different languages to further enhance the multilingual capabilities of LLMs. Thirdly, our method does not involve engineering on the architecture of LLMs. For specific extremely low-resource languages, modifying the architecture and introducing additional data, such as vocabulary expansion or continuing pre-training, might be beneficial in enhancing multilingual performance.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning.

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

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in bert for text generation.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment.

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english?

Gemini Team Google, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Mitchell A. Gordon and Kevin Duh. 2019. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

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

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Hiyouga. 2023. Llama factory. `https://github.com/hiyouga/LLaMA-Factory`.

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.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt.

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*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023a. Align after pre-train: Improving multilingual generative models with cross-lingual alignment.

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

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu lee Wang, Qing Li, and Xiaoqin Zhong. 2023c. Label supervised llama finetuning.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2021. Pre-training multilingual neural machine translation by leveraging alignment information.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Yukun Ma, Trung Hieu Nguyen, and Bin Ma. 2022. Cpt: Cross-modal prefix-tuning for speech-to-text translation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6217–6221.

Zhuoyuan Mao and Yen Yu. 2024a. Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages.

Zhuoyuan Mao and Yen Yu. 2024b. Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms – large language models for southeast asia.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Jeroen Ooms. 2024. *cld3: Google's Compact Language Detector 3*. R package version 1.6.0.

OpenAI. 2022. ChatGPT. `https://openai.com/chatgpt`.

OpenAI. 2023. GPT-4 technical report.

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

Saurabh Pahune and Manoj Chandrasekharan. 2023. Several categories of large language models (llms): A short survey. *International Journal for Research in Applied Science and Engineering Technology*, 11(7):615–633.

Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. 2022. Revisiting self-distillation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Leonardo Ranaldi and Giulia Pucci. 2023. Does the english matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183.

Leonardo Ranaldi and Fabio Massimo Zanzotto. 2023. Empowering multi-step reasoning across languages via tree-of-thoughts.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Iman Saberi, Fatemeh Fard, and Fuxiang Chen. 2024. Advfusion: Multilingual adapter-based knowledge transfer for code summarization.

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

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

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

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

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023b. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Andrea Wen-Yi and David Mimno. 2023. Hyperpolyglot llms: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. 2019. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. WeChat neural machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 243–254, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation.

L. Zhang, C. Bao, and K. Ma. 2022a. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis; Machine Intelligence*, 44(08):4388–4403.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. 2022b. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation.

Yuanchi Zhang, Peng Li, Maosong Sun, and Yang Liu. 2023. Continual knowledge distillation for neural machine translation.

Yuanchi Zhang and Yang Liu. 2021. Directquote: A dataset for direct quotation extraction and attribution in news articles. *arXiv preprint arXiv:2110.07827*.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer.

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

A Implementation Details

The signature of SacreBLEU we use in this work is “nrefs:1 | case:mixed | eff:no | tok:flores200 | smooth:exp | version:2.0.0”. The Stanford Alpaca dataset comprises 52,002 entries, licensed under the CC BY-NC 4.0 agreement. For each target language, machine translation parallel corpora are sampled from Opus100 (Zhang et al., 2020), consisting of 1,000 entries. When utilizing NLLB for machine translation, we set the beam size to 4, with the remaining configurations adopting the default parameters from Huggingface Transformers (Wolf et al., 2020). In the reimplementation of baselines, the same machine translation system is employed to provide multilingual alignment data. For the 16 languages involved in our experiments, XL-SUM and MKQA datasets have not covered all of them. During the evaluation of MKQA, questions lacking ground truth are skipped.

B More Detailed Results on SeaLLM

We conduct experiments on three common Southeast Asian languages: Indonesian (IND), Thai (THA), and Khmer (KHM). As shown in Table 8, 9, 10, and 11, SDRRL still outperforms the baselines, demonstrating the generalizability of SDRRL in different LLMs.

	IND	KHM	THA	AVG.
Performance on Target Language				
SFT	47.71	32.56	46.44	42.24
T-SFT	48.31	32.89	46.77	42.77
CIT	48.83	32.56	46.20	42.53
XCOT	45.81	32.22	45.55	41.19
SDRRL	**50.39**	**33.67**	**46.96**	**43.67**
Performance on English Language				
SFT	60.56	60.56	**59.46**	60.19
T-SFT	60.78	57.89	57.43	58.70
CIT	58.55	58.10	59.33	58.66
XCOT	57.77	57.99	57.43	57.73
SDRRL	**61.68**	**60.89**	59.44	**60.67**

Table 8: Results of baselines and our SDRRL on BELEBELE benchmark using SeaLLM.

	Ind	Tha	Avg.
Performance on Target Language (ROUGE-1)			
SFT	21.91	24.65	23.28
T-SFT	21.07	21.26	21.17
CIT	21.19	23.93	22.56
XCOT	22.20	21.85	22.02
SDRRL	**23.62**	**25.78**	**24.70**
Performance on Target Language (ROUGE-L)			
SFT	16.46	16.50	16.48
T-SFT	16.23	14.40	15.32
CIT	15.84	15.66	15.75
XCOT	16.93	14.65	15.79
SDRRL	**18.06**	**17.73**	**17.89**
Performance on English Language (ROUGE-1)			
SFT	21.39	22.85	22.12
T-SFT	21.93	23.17	22.55
CIT	21.65	23.07	22.36
XCOT	21.27	21.99	21.63
SDRRL	**23.47**	**23.55**	**23.01**
Performance on English Language (ROUGE-L)			
SFT	14.79	15.71	15.25
T-SFT	15.19	16.07	15.63
CIT	14.91	15.92	15.42
XCOT	14.66	15.14	14.90
SDRRL	**16.34**	**16.15**	**16.24**

Table 9: Results of baselines and our SDRRL on XL-SUM benchmark on the target language using SeaLLM.

	Ind	Tha	Thm	Avg.
xx→en (BLEU)				
SFT	36.75	20.93	20.22	28.49
T-SFT	32.23	14.41	15.21	23.72
CIT	22.52	15.84	14.10	18.31
XCOT	33.20	16.48	14.71	23.96
SDRRL	**38.30**	**21.76**	**20.64**	**29.47**
en→xx (BLEU)				
SFT	30.26	16.53	6.64	18.45
T-SFT	28.29	13.10	4.88	16.59
CIT	31.21	18.15	9.76	20.49
XCOT	29.15	14.28	5.26	17.21
SDRRL	**36.28**	**24.02**	**15.43**	**25.86**
xx→en (COMET)				
SFT	86.94	82.89	80.07	83.51
T-SFT	84.49	74.61	71.29	77.89
CIT	80.78	78.87	76.18	78.48
COT	85.69	77.57	72.14	78.91
SDRRL	**87.39**	**83.07**	**80.63**	**84.01**
en→xx (COMET)				
SFT	86.78	73.22	64.05	75.41
T-SFT	85.44	66.95	59.09	72.26
CIT	85.80	74.42	69.60	77.70
COT	85.23	68.26	62.24	73.74
SDRRL	**88.70**	**79.03**	**75.97**	**82.34**

Table 10: Results of baselines and our SDRRL on FLORES benchmark using SeaLLM.

C Experiments with Non-English Language as the Source Language

SDRRL aims to transfer the proficiency of LLMs from resource-rich languages to another target language, indicating that the source language is not limited to English. In this section, we use French instead of English to generate responses and replicate experiments conducted in Indonesian, Japanese, and Korean languages. Table 12 presents the average knowledge distillation gains from SDRRL over SFT when employing English and French as the source languages, respectively. Due to the inferior performance of LLaMA-2-7B in French (Bandarkar et al., 2023) compared to English and the decreased effectiveness of the external translation system NLLB when French is used as the source language (Costa-jussà et al., 2022), the knowledge distillation gains generated with French as the source language are smaller.

	Tha	Khm	Avg.
Performance on Target Language			
SFT	40.68	37.04	38.86
T-SFT	48.32	38.48	43.40
CIT	48.38	39.01	43.70
XCOT	45.07	39.00	42.04
SDRRL	**49.44**	**39.81**	**44.63**
Performance on English Language			
SFT	39.62	39.62	39.62
T-SFT	37.92	36.94	37.43
CIT	37.64	35.69	36.67
XCOT	38.40	37.48	37.94
SDRRL	**40.66**	**39.97**	**40.32**

Table 11: Results of baselines and our SDRRL on MKQA dataset using SeaLLM.

	NLU Avg.	NLG Avg.
English	**+6.29**	**+5.31**
French	+2.94	+1.77

Table 12: Distillation gains from SDRRL with English or French as the source language. Average scores on the natural language understanding task (NLU, including BELEBELE) and natural language generation tasks (NLG, including FLORES, XL-SUM ROUGE-L, and MKQA) are reported.

D Case Study

As shown in Table 13, we provide several cases of SDRRL in several languages. In the case of Hungarian, LLM encounters severe off-target issues, where the response is in English and is inconsistent with the input language Hungarian, which is a frequent problem during multilingual generation. After SDRRL, the severe off-target issue has been effectively mitigated. In the Slovenian case, LLM produces hallucinations when answering factual

Hungarian	**Question**: Mit jelent a DNA? **Before SDRRL:** A DNA (deoxyribonucleic acid) is the hereditary material present in all living organisms. It is a long double-stranded helix made up of four nitrogenous bases; adenine (A), guanine (G), cytosine (C) and thymine (T). **After SDRRL:** A DNA a dezoxiribonukleinsav rövidítése, egy molekula, amely két szálból áll, amelyek egymáshoz kötődnek, és hidrogénkötések tartják össze őket. Ez a genetikai anyag, amely az élőlények fejlődését és működését szabályozza, és a sejtek fehérje- és szénhidrátszintézisét irányító utasításokat tartalmazza.
Slovenian	**Question**: Opišite življenje in vladanje kralja Karla II. **Before SDRRL:** Kralj Karel II. (1638-1700) je bil francoski krščanski kralj, ki je vladal od leta 1685 do svoje smrti. Ko je postal kralj, je zavzel podeželje in se lotil reform, katerek cilj je bil izboljšati upravo in omogočiti zaščito družbe. Karel II. je bil tudi pomorski kralj, ki si je meril svet🌍. Under his rule, the French navy became one of the most powerful in the world. Karel II. je bil tudi velik pokrovitelj umetnosti in književnosti. Že stoletje po njegovem vrnitemenu je postal Karel II. znan kot eno največjih francoskih vladarjev zadnjega obdobja. **After SDRRL:** Kralj Karel II. je bil angleški kralj od leta 1660 do svoje smrti leta 1685. Bil je sin Karla I., ki je bil odstranjen z oblasti in usmrčen leta 1649. Po angleški državljanski vojni in obdobju republike je bil Karel II. ponovno kronan za kralja Anglije leta 1660, kar je označilo začetek Kraljeve obnove. Bil je znan kot podpornik umetnosti in kulture, ki je spodbujal delo številnih umetnikov in pisateljev. Njegovo obdobje vladanja je bilo zaznamovano z obnovitvenimi projekti, vključno z razvojem Londona po Velikem požaru leta 1666. Karel II. je umrl leta 1685 brez legitimnih moških potomcev, kar je dolgoročno vplivalo na vprašanje nasledstva. Njegova smrt ni neposredno vplivala na konec Kraljeve obnove ali angleško državljansko vojno, ki se je končala pred njegovim vladanjem, niti na začetek angleške suverenosti v Severni Ameriki.
Bulgarian	**Question**: Моля, обяснете историята на ледниковия период. **Before SDRRL:** ледников период е период в ерата, в която Земята е bila предимно покрита с лед. това се дължи на увеличение на азъйствата на Северния полюс, които се оттикват водата в океаните и причинява замразяването на течности и земя. това води до увеличение на масата лед bglobus и снижаваниеа нива на температурата на аеросферата. това връвля banother glacial period or ice age. **After SDRRL:** Ледниковият период е период от време в климатичната история на Земята, през който климатът става много студен и полярните ледени шапки се разширяват, покривайки голяма част от сушата. Това се случва, когато нивото на атмосферната температура значително се понижи, което позволява на водата в океаните да се превърне в лед и води до увеличаване на обема на леда и връщането на вода към морския лед. Ледниковите периоди обикновено са свързани с забележително намаляване на глобалните температури, което се случва в продължение на десетилетия или стотици години. Няколко фактора могат да доведат до ледников период, включително промени в орбитата на Земята, намаляване на вулканичните изригвания и промени в океанските течения. Най-скорошният ледников период, известен като последната ледникова епоха, приключи преди около 10 000 години, когато климатът започна да се затопля и ледниците по света започнаха да се топят.

Table 13: Several cases before and after the application of SDRRL. Following SDRRL, the responses from LLMs become smoother, more comprehensive, and contain fewer errors (colored in red).

questions about history, leading to factual inaccuracies, noise tokens like emojis, and off-target English phrases. For example, The time frame mentioned (1685-1700) and the description does not match any king named Charles. It seems there's a mix-up with historical figures. After SDRRL, the hallucination issue has been alleviated, and the generated content becomes more detailed, refined, and fluid. In the Bulgarian scenario, the response contains several grammatical errors, such as "ледников", "оттикват" and "снижаваниеа". In this case, the SDRRL process enhances the clarity and natural flow of the output text while also eliminating grammatical errors in Bulgarian. See appendix E for statistical results regarding off-target issues.

E Off-Target Issue Analysis

We delve deeper into the effectiveness of SDRRL in alleviating off-target language issues during LLM responses. Weevaluate the responses of LLaMA-2 on Dolly (Conover et al., 2023) and its multilingual extension, Bactrain-X (Li et al., 2023b). As depicted in Figure 3, we showcase the variation in the off-target occurrence rates across each target language throughout the SDRRL process. This indicates that SDRRL plays a constructive role in mitigating off-target issues, ensuring consistency

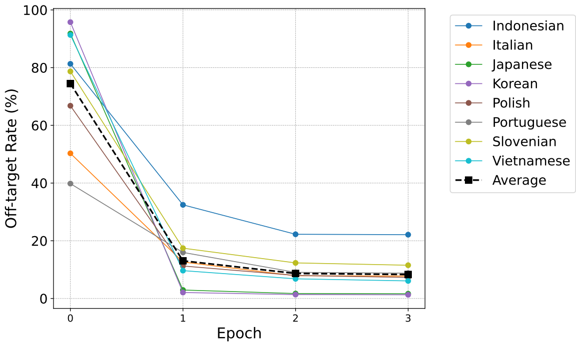

Figure 3: The occurrence rate of off-target issues in various languages during the SDRRL process.

between the input and the response languages.

F Potential Risks of Our Method

Because our method involves distilling knowledge from other target languages towards high-resource languages to achieve cross-linguistic alignment, it may lead to cultural unfairness for mid- and low-resource languages. For instance, after aligning to English using SDRRL, responses of LLMs in African languages may also adhere to the cultural practices and social norms of English.