Prefix Text as a Yarn:
Eliciting Non-English Alignment in Foundation Language Model

Runzhe Zhan♣* **Xinyi Yang**♣ **Derek F. Wong**♣✉ **Lidia S. Chao**♣ **Yue Zhang**♠✉

♣NLP²CT Lab, Department of Computer and Information Science, University of Macau
♠School of Engineering, Westlake University
nlp2ct.{runzhe, xinyi}@gmail.com, {derekfw, lidiasc}@um.edu.mo
zhangyue@westlake.edu.cn

Abstract

While supervised fine-tuning (SFT) has been a straightforward approach for tailoring the output of foundation large language model (LLM) to specific preferences, concerns have been raised about the depth of this alignment, with some critiques suggesting it is merely "superficial". We critically examine this hypothesis within the scope of cross-lingual generation tasks, proposing that the effectiveness of SFT may be constrained by its reliance on prior tokens to guide cross-lingual generation. Based on this crucial insight, and in response to the challenges posed by the costly and limited availability of non-English data for SFT, we introduce a novel training-free alignment method named PRETTY, which employs minimal task-related prior tokens to bridge the foundation LLM and the SFT LLM, achieving comparable performance without training. Experiments on machine translation and part-of-speech tagging across eight languages demonstrate the efficacy of PRETTY in cross-lingual settings. Remarkably, by initiating the decoding process with only one or two prior tokens, foundation LLMs can achieve performance comparable to their SFT counterparts. This method presents a cost-effective alternative to SFT and advances the democratization of multilingual LLMs[1].

1 Introduction

Supervised fine-tuning (SFT) refines large language models (LLMs) using task-specific instruction data to enhance their capability to follow instructions (Touvron et al., 2023; Peng et al., 2023) and to align their outputs with human preferences and safety considerations (Ouyang et al., 2022; Rafailov et al., 2023; Dong et al., 2023b; Yuan

---

*Work was done during a visit to Westlake University.
✉ Co-corresponding authors.
[1]The code and data are available at: https://github.com/NLP2CT/PrettyAlign

et al., 2023; Lan et al., 2024). This process is often termed "alignment", signifying the tailoring of model outputs to conform to specific downstream requirements. Nevertheless, current research casts doubt on the necessity and potential adverse impacts of SFT. But the alignment achieved through SFT is often considered to be "superficial", with the process potentially repurposing pre-existing knowledge from pre-training to merely reshape outputs to meet specific criteria (Zhou et al., 2023; Lin et al., 2023). It has been observed that even a small-scale SFT training dataset can produce significant alignment effects (Liu et al., 2023; Xia et al., 2024). On the other hand, recent empirical studies (Luo et al., 2023; Dong et al., 2023a) have raised concerns that SFT might hurt the knowledge acquired during its pre-training phase, leading to serious consequences like catastrophic forgetting.

Not only is there no definitive consensus on the necessity of SFT, but the majority of these studies also focus on monolingual tasks. LLMs still encounter challenges in handling complex cross-lingual generation tasks (Schioppa et al., 2023; Wang et al., 2023). Current research on cross-lingual alignment primarily seeks to extrapolate or align English capabilities to other languages using the SFT paradigm (Zhang et al., 2023; Chai et al., 2024; Xu et al., 2024), yet there remains a gap in exploring the specific impacts of SFT-based cross-lingual alignment. Furthermore, given the potential risk of SFT leading to the forgetting of pre-training knowledge, the question of how to achieve cross-lingual alignment without training remains underexplored.

To bridge these gaps, our study conducts an in-depth examination of the impact of SFT on cross-lingual generation. We investigate the influence of SFT on the decoding patterns of foundation models in cross-lingual contexts, hypothesizing that the success of SFT largely hinges on the selection of initial prior tokens that are critical for eliciting task-

12131

Instruction: Translate the following sentence from English to Ukrainian: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.
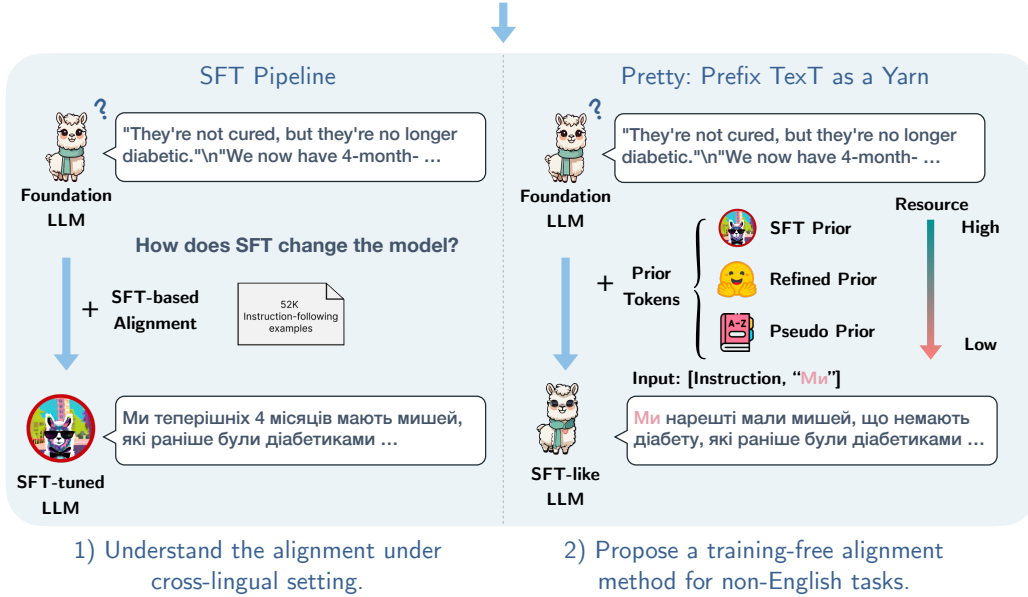
SFT Pipeline

Foundation LLM

"They're not cured, but they're no longer diabetic."\n"We now have 4-month- …

How does SFT change the model?

+ **SFT-based Alignment** — 52K Instruction-following examples

SFT-tuned LLM

Ми теперішніх 4 місяців мають мишей, які раніше були діабетиками …

Pretty: Prefix TexT as a Yarn

Foundation LLM

"They're not cured, but they're no longer diabetic."\n"We now have 4-month- …

+ **Prior Tokens**
- SFT Prior
- Refined Prior
- Pseudo Prior

Resource: High → Low

Input: [Instruction, "Ми"]

SFT-like LLM

Ми нарешті мали мишей, що немають діабету, які раніше були діабетиками …

1) Understand the alignment under cross-lingual setting.

2) Propose a training-free alignment method for non-English tasks.

Figure 1: Illustration of our research question and proposed **P**refix **T**ex**T** as a **Y**arn (PRETTY) framework.

specific generation in the target language. Furthermore, the observed decoding similarities between foundation and SFT models support the extension of the superficial alignment hypothesis to cross-lingual scenarios. Responding to these insights, we introduce a training-free alignment method named "PRETTY" for cross-lingual and non-English tasks. The **Pre**fix **T**ex**T**s act as a **Y**arn (PRETTY) linking the foundation LLM and the SFT LLM, eliciting the foundation LLM to exhibit near-SFT performance levels. Specifically, we augment the original input with a few tokens that serve as decoding priors, and then prompt the foundation LLM to resume decoding based on this modified input. In most cases, only one or two task-related prior tokens are needed, and the method for constructing these prior tokens is flexible across various kinds of language resources, fostering the democratization of multilingual LLMs.

We conducted experiments on machine translation (Goyal et al., 2022), cross-lingual summarization (Bhattacharjee et al., 2023) and non-English part-of-speech (POS) tagging (Liang et al., 2020) tasks across eight languages. These tasks exemplify cross-lingual generation and multilingual language understanding, and they provide ample non-English test data to evaluate effectiveness across varying levels of resource availability. The experimental results demonstrate that PRETTY can ef-

fectively align the foundation model to match SFT model's performance without training, by merely adding two prior tokens in the decoding.

2 Iceberg Model of SFT

2.1 Preliminaries

Pre-training The pre-training (PT) of LLMs is primarily conducted through language modeling tasks on large-scale unlabeled data (Touvron et al., 2023; Achiam et al., 2023). During this phase, given a sequence X_{PT} of length N and a context window k, the optimization objective is maximizing the joint probability P_{LM} as:

$$P_{\text{LM}}(X_{\text{PT}}) = \prod_{i=1}^{N} P(x_i|x_{i-k:i-1}) \qquad (1)$$

which encourages the model to generate text that naturally follows from the preceding context. However, this "text completion" behavior can become a bottleneck when models are prompted to switch languages or follow specific instructions of cross-lingual generation. It is frequently observed that when prompted with English input and instructed to produce text in a different language, as illustrated in the upper example of Figure 1, the foundation model often continues to decode in English.

SFT SFT leverages labeled data pair $(X_{\text{ins.}}, Y)$ to empower models with the ability to follow in-

structions. This stage aims to maximize the probability of the expected answer Y conditioned on the input text $X_{\text{ins.}}$, where $X_{\text{ins.}}$ consists of the task instruction and task input.

$$P_{\text{SFT}}(Y|X_{\text{ins.}}) = \prod_{j=1}^{T} P(y_j|y_{1:j-1}, X_{\text{ins.}}) \quad (2)$$

SFT is crucial for aligning foundation models to perform task-specific instructions, effectively transforming a general-purpose LLM into an instruction-following assistant. However, data quality, training costs, and the imbalance of multilingual data hinder the democratization of assistant LLM. As mentioned before, SFT may be harmful to pre-training knowledge. Thus, it is meaningful and important to understand the underlying mechanism of SFT-based alignment and propose a more efficient alignment method.

2.2 Beneath the SFT-based Alignment

Prior Knowledge Hypothesis It is worth noting that pre-training corpora also contain sequences that naturally express task-specific information, which imparts certain capabilities to the foundation LLMs. For example, the presence of semantically equivalent expressions in the pre-training text may enable LLM acquire machine translation ability during pre-training stage (Radford et al., 2019).

Despite its extensive prior knowledge, the foundation LLM still struggles with complex cross-lingual generation tasks. Beyond existing studies, we provide more concrete insights into this issue by prompting foundation LLMs with various instructions (Bawden and Yvon, 2023). Notably, only 31.8% of these prompts successfully elicit translation capability from the foundation LLMs[2].

This deficiency may stem from two main factors: First, the proportion of text with the aforementioned characteristics in the pre-training corpus X_{PT} is still relatively small, and most of it is far from resembling human instruction text $X_{\text{ins.}}$. Consequently, the model is more likely to predict tokens suitable for completing formal texts than those required for task-specific instructions. As a result, the foundation LLM often fails to produce tokens $y \in Y_{1:T}$ in the intended target language. Secondly, the predominance of English in the pre-training data skews the token generation probabilities of foundation LLM. Given a cross-lingual context, the model favors predicting tokens in English,
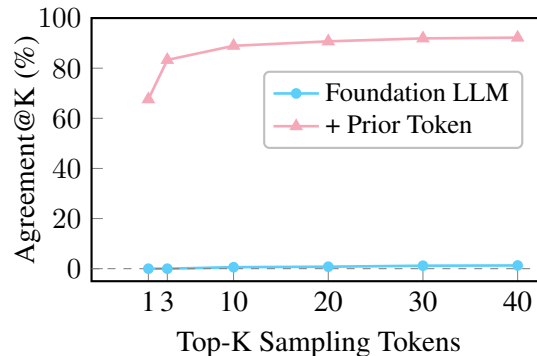

Figure 2: The agreement between the SFT model and the foundation model in terms of the selection of the next token. Once the Prior Token is provided, the token chosen by the SFT model is also can be found within the Top-K candidate words of foundation model.

while the token probabilities for other languages remain comparatively low. For example, English data comprises up to 90% of the Llama2 pre-training data (Touvron et al., 2023), which may lead models to generate text with an English-centric bias.

The above hypothesis might be reasonable when we revisit Equation (1) and Equation (2). The probability $P_{\text{LM}}(X_{\text{PT}})$ of the next token prediction for the foundation model is conditioned on the distribution of the pre-training text X_{PT}. SFT narrows the probability space for token selection, adjusting the parameters to better align with the distribution, *i.e.*, the probability $P_{\text{SFT}}(y|X_{\text{ins.}})$ is conditioned on the distribution of the instruction text $X_{\text{ins.}}$.

Experimental Settings To validate the aforementioned hypothesis, we selected the representative cross-lingual task of machine translation as our analytical testbed. The main research method involved quantifying the differences and similarities in the decision space and token selection behavior between the foundation LLM and the SFT-aligned LLM. For the model selection, we chose the foundation Llama2 7B model and conducted supervised fine-tuning on it using the Alpaca dataset[3](Taori et al., 2023). The optimization was carried out using a cosine learning rate scheduler, with the maximum learning rate set to $2e-5$ and a warmup ratio of 0.03. Training was performed on two Nvidia-H800 GPUs using LoRA parameter-efficient fine-tuning (Hu et al., 2022) technique, with a cumulative batch size of 64. Other hyper-parameters follow those of the original Alpaca settings.

---

[2]For detailed information, please refer to Appendix C.3.
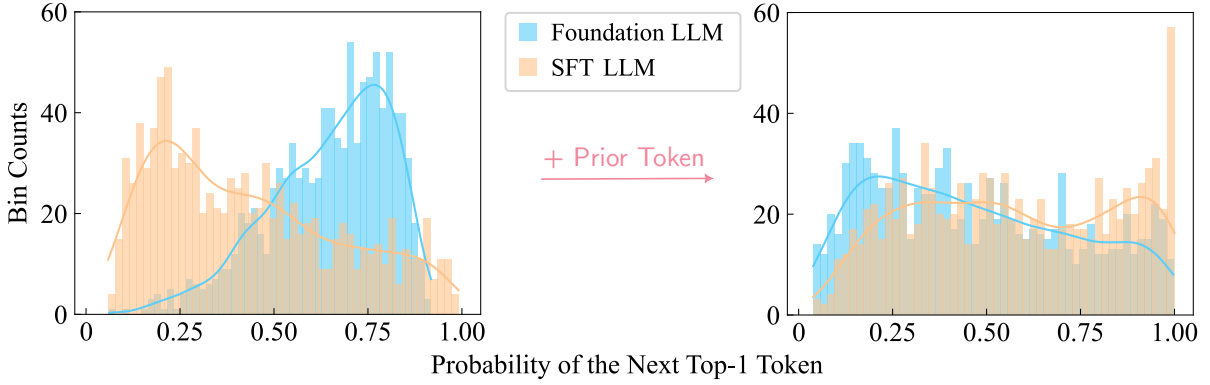
[3]https://github.com/tatsu-lab/stanford_alpaca

Figure 3: The probability distribution of tokens selected by various models. Incorporation of a Prior Token causes the decision probabilities of both models to converge across all data instances.
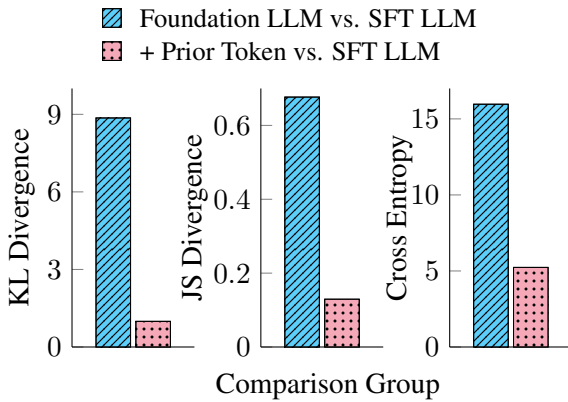

Figure 4: The divergence in probability distributions across the entire vocabulary during decoding. Prior Token significantly reduces the discrepancy between the foundation model and the SFT model.

A Prior Token Elicits Silent Majority Inspired by the categorization of token shifts by Lin et al. (2023), we propose to quantify the agreement of token selection between foundation LLM θ_{PT} and SFT LLM θ_{SFT}. Given the same prefix input $\hat{X}$, we aim to measure whether the next token selected by the SFT LLM, y_{SFT}, is among the top-K tokens, $\mathbf{y}_{\text{PT}}$, with the highest probabilities in the decision space of the foundation LLM, which can be formally expressed as follows:

$$y_{\text{SFT}} = \underset{y \in V}{\arg\max}\ P(y|\hat{X}; \theta_{\text{SFT}})$$

$$\mathbf{y}_{\text{PT}} = \{y|\underset{y \in V}{\arg\,\text{top}K}\ P(y|\hat{X}; \theta_{\text{PT}})\}$$

$$\text{Aggrement}_K = \frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{y_{\text{SFT}}\in\mathbf{y}_{\text{PT}}} \qquad (3)$$

where V is the vocabulary shared by two models, and L is the length of the dataset.

We compare the agreement of the token selection made by the models under the **same** prefix text $\hat{X}$ in two different experimental setups. The first setup uses the instruction text as the prefix, *i.e.*, $\hat{X} = X_{\text{ins.}}$; the second takes the first token decoded by the SFT model as a prior token, appending it to the original instruction prefix, *i.e.*, $\hat{X} = \left[X_{\text{ins.}}, y_{\text{SFT}}^{(1)}\right]$. For the SFT model, the second setup is equivalent to continuing its own decoding behavior, whereas for the foundation model, it becomes decoding with the addition of a prior token.

Figure 2 illustrates the agreement between the foundation model's predictions and those of the SFT model regarding the selection of the next token, given an identical text prefix. Across the entire translation data, it is observed that after incorporating merely one prior token, the foundation model exhibits a high degree of agreement with the SFT model in terms of token selection. This demonstrates that the alignment effect of SFT in cross-lingual generation tasks is also somewhat superficial. Even in instances where the token with the highest probability differs between the two models, 90.8% of the tokens chosen by the SFT model are present within the "*silent majority*" in the decision space of the foundation model, specifically, among the top 20 most probable token choices.

Lens of Distribution Instead of focusing on the coverage of token selection outcomes, we also observe the decision dynamics and similarities from the perspective of the overall probability distribution, with the data settings consistent with the previous setup. First, as shown in Figure 3, after adding a prior token, the probability of the next tokens chosen by both models have closely aligned distributions. The reason that the foundation model

12134

exhibits a high probability given the instruction text as a prefix lies in a preference for choosing to continue the instruction text rather than completing the cross-linguistic semantic transformation. Additionally, we quantify the distribution disparities between the two models through the probability distribution of the vocabulary. The disparity metrics used include Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence, and cross-entropy (Kullback, 1997). As depicted in Figure 4, the disparity of decision space of the foundation model significantly decreases after adding the prior token, aligning more closely with the SFT model.

These findings indicate that such prior tokens serve a dual function: they not only steer the foundation model towards generating tokens pertinent to cross-lingual generation but also modulate the decision space to align more closely with the task-specific distribution.

3 Pretty: Prefix TexT as a Yarn

3.1 Motivation

The observations discussed earlier confirm that SFT effectively narrows the decision space of the foundation model during text generation that is conditioned on instruction text. The disparity in token selection between the foundation LLM and the SFT LLM, however, might not be reduced by a training-based transfer methodology. By appending a prior token into the instruction text, the choices of the next token between the two models tend to become largely consistent, and in the vast majority of cases, the tokens chosen by SFT model are also found within the high-probability candidate words of foundation model. These phenomena show that the alignment elicited by SFT is somewhat superficial in cross-lingual generation tasks and motivate us to propose a training-free alignment method by leveraging these prior tokens.

3.2 Formulation

Upon revisiting Equation (1) and Equation (2), the goal of proposing a training-free approach is to enable the conditional decoding probability of foundation model to approximate those of SFT model. Therefore, ideally, the selected prior tokens $X_{\text{pri.}} = \{x_{\text{pri.}}\}$ may satisfy the following criteria:

$$
\begin{aligned}
& P(\mathbf{y}_{\text{PT}} | \left[X_{\text{ins.}}, X_{\text{pri.}} \right]; \theta_{\text{PT}}) \\
& \approx P(\mathbf{y}_{\text{SFT}} | X_{\text{ins.}}; \theta_{\text{SFT}})
\end{aligned}
\tag{4}
$$

where $\mathbf{y}_{\text{PT}}$ and $\mathbf{y}_{\text{SFT}}$ represent the outputs of the foundation and the SFT models, respectively. It is important to note that a single prior token may not serve as an optimal solution due to its non-derivable characteristic. Hence, we extend our methodological approach to include appending multiple prior tokens, grouping them to form a prefix text.

3.3 Construction of Prior Tokens

To ensure that the proposed method is applicable to a wide array of languages, we propose three construction strategies based on the availability of language resources, aiming to guarantee the universality of our approach.

SFT Prior represents an ideal scenario where the first few tokens generated by a SFT model are used as priors. This method is theoretically rational when the SFT model is derived from the same foundation model because it directly approximates Equation (4) by sampling $x_{\text{pri.}} \sim \{y_{\text{SFT}}\}$. In practical applications, this might be suitable for high-resource languages due to the imbalanced language capabilities of other languages. Additionally, SFT could potentially degrade the knowledge and abilities that the foundation model has already acquired. In such cases, using prior tokens from the SFT model can contribute to generating better results. This situation will be discussed further in the subsequent section.

Refined Prior is more readily accessible for most languages and tasks. We can utilize the output tokens generated by a smaller model trained for specific downstream tasks and use them as prior tokens to achieve weak-to-strong generalization (Burns et al., 2023).

Pseudo Prior For extremely low-resource language pairs, where there is no labeled data for downstream tasks, both SFT and Refined priors are difficult to obtain. For cross-lingual tasks, we can create pseudo labels in target language as prior tokens. For instance, in machine translation tasks, we might use bilingual dictionaries to acquire pseudo prior tokens. However, the quality and accuracy of pseudo labels remain uncertain, and the extent of their impact on the generative performance of the foundation LLM is not yet clear. We will explore this problem further in the context of experimental results discussed later in the paper.

4 Experiments

We examine the effectiveness of our proposed training-free alignment method on two distinct tasks: machine translation, cross-lingual summarization and non-English POS tagging. Machine translation serves as a prototypical cross-lingual generation task, entailing the transformation of a sequence from a source language to a target language (Bahdanau et al., 2015; Vaswani et al., 2017; Zhan et al., 2023; Pang et al., 2024). As for cross-lingual summarization, it requires the model to generate a summary of an article in a different language (Bhattacharjee et al., 2023; Chen et al., 2023). Although POS tagging (Manning, 2011; Nivre et al., 2017; Chiche and Yitagesu, 2022) primarily assesses the model's ability to understand monolingual text, we include it as multilingual experiments to show the universality of our methods.

4.1 Experimental Settings

Data We use Flores-101 (Goyal et al., 2022), CrossSum (Bhattacharjee et al., 2023) as benchmarks for machine translation and cross-lingual summarization tasks, respectively. For POS tagging tasks, we choose the POS test split from the XGLUE benchmark (Liang et al., 2020), which is derived from the Universal Dependencies Treebank v2.5. To investigate the performance across various resource languages, we carefully selected eight languages based on the pre-training data proportions disclosed in the Llama2 technical report (Touvron et al., 2023). These languages are French, German, Chinese, Russian, Ukrainian, Portuguese, Hindi and Arabic. Among these, the first four languages account for more than 0.1% of the pre-training data of Llama2, while Ukrainian and Portuguese fall below 0.1%, Hindi and Arabic is below 0.05%. For the Llama2 model, we can categorize these three types of languages into high-resource languages, low-resource languages, and extremely low-resource languages, respectively.

Models and Baselines The settings of Llama2 foundation model and the SFT model are consistent with those described in Section 2.1. To further demonstrate the generality of our proposed method, we incorporated the Mistral-7B LLM family (Jiang et al., 2023) into our experiments, covering both out-of-the-box SFT and foundation models.

In the machine translation task, the Llama2 foundation model does not tend to generate translations when given explicit translation instructions.

While this is a normal phenomenon according to our previous discussion, to ensure a fair comparison, we also searched for a better prompts for the foundation model. This prompting approach is referred to as "Llama2-7B$_{\text{PROMPTING}}$" in subsequent sections. For POS tagging, we experimented with various instructions and selected one that consistently prompts both the foundation model and the SFT model to reliably generate classification results in text. Although we report the zero-shot performance for the aforementioned tasks, we found that even out-of-the-box SFT models cannot produce stable output for cross-lingual summarization task. Hence, we prepend a constant demonstration before the input to also assess the effectiveness of our proposed method under the in-context learning paradigm (Dong et al., 2023c).

Sources of Prior Token The sources of crafting prior tokens include:

- **SFT Prior**: We took the first k tokens of output produced by SFT model as the prior tokens. For multiple SFT models, we select the model that demonstrates better performance.

- **Refined Prior**: We use downstream task models with smaller parameter sizes as the source of refined priors. For the different tasks, we utilized the distilled 600M variant of NLLB-200 translation model[4](Costa-jussà et al., 2022), mT5 cross-lingual summarization model[5] and the Unicoder-NLU model[6](Huang et al., 2019), respectively.

- **Pseudo Prior**: The pseudo prior is applied to two cross-lingual tasks since it can utilize cross-lingual language resources. We create pseudo prior tokens for machine translation task by referencing dictionary [7] entries. For cross-lingual summarization, we initially extract keywords from each passage using KeyBERT (Grootendorst, 2020) and then perform word-by-word translation. However, not all initial sentence tokens will be covered by the dictionary. To handle such instances, a back-off strategy is implemented, where the target language equivalent of the first available dictionary token is used as the prior token.

---

[4]https://huggingface.co/facebook/nllb-200-distilled-600M
[5]https://hf.co/csebuetnlp/mT5_m2m_crossSum
[6]https://github.com/microsoft/Unicoder/
[7]Please refer to Appendix C.4 for dictionary information.

Models	En-Zh		En-Uk		Zh-En		Uk-En		Avg.		%SFT.
English-Centric											
	spBL.	CoM.	spBL.	CoM.	spBL.	CoM.	spBL.	CoM.	spBL.	CoM.	All
Llama2-7B											
Llama2-7B-Alpaca	13.6	<u>80.9</u>	24.0	83.3	23.5	<u>85.1</u>	34.4	85.5	23.9	83.7	-
Llama2-7B-Chat	7.8	67.2	18.1	71.0	18.5	81.3	30.4	83.3	18.7	75.7	-
Llama2-7B$_{\text{PROMPTING}}$	5.9	64.1	11.0	60.9	**24.3**	84.8	34.2	85.0	18.9	73.7	80.4
Llama2-7B	7.7	72.0	0.2	32.4	12.0	74.4	9.3	59.2	7.3	59.5	52.5
+PRETTY (SFT Prior)	13.3	80.0	23.0	83.1	23.7	84.9	33.6	85.3	23.4	83.3	98.8
+PRETTY (Pseudo Prior)	12.0	75.7	18.1	74.1	16.9	80.3	27.2	78.3	18.6	77.1	85.4
+PRETTY (Refined Prior)	**<u>14.2</u>**	80.5	**<u>24.1</u>**	**83.8**	24.0	**84.9**	**<u>34.6</u>**	**<u>85.6</u>**	**<u>24.2</u>**	**<u>83.7</u>**	**100.9**
Mistral-7B											
Mistral-7B-Instruct	6.6	64.6	20.3	78.2	20.5	<u>83.2</u>	<u>32.9</u>	<u>84.8</u>	20.1	77.7	-
Mistral-7B	1.2	42.6	0.3	30.8	19.9	77.1	21.5	69.4	10.7	55.0	46.2
+PRETTY (SFT Prior)	13.8	78.1	23.1	79.2	20.0	82.3	32.1	83.3	22.3	80.7	117.2
+PRETTY (Pseudo Prior)	13.3	75.8	20.1	75.7	16.5	79.7	24.9	77.3	18.7	77.1	107.2
+PRETTY (Refined Prior)	**<u>15.9</u>**	**<u>81.3</u>**	**<u>24.9</u>**	**82.9**	**21.5**	**83.0**	**32.3**	**83.9**	**<u>23.7</u>**	**<u>82.7</u>**	**124.6**

Models	De-Fr		Fr-De		Zh-Pt		Pt-Zh		Avg.		%SFT.
Non-English-Centric											
	spBL.	CoM.	spBL.	CoM.	spBL.	CoM.	spBL.	CoM.	spBL.	CoM.	All
Llama2-7B											
Llama2-7B-Alpaca	29.8	<u>81.5</u>	<u>24.1</u>	<u>80.9</u>	16.6	<u>81.4</u>	11.3	78.6	<u>20.5</u>	80.6	-
Llama2-7B-Chat	6.2	68.0	7.3	64.5	3.0	67.8	6.2	66.6	5.7	66.7	-
Llama2-7B$_{\text{PROMPTING}}$	22.2	77.4	15.4	73.3	14.4	78.9	4.4	64.1	14.1	73.4	78.5
Llama2-7B	1.0	51.1	3.2	54.0	0.9	61.4	7.3	70.0	3.1	59.1	47.6
+PRETTY (SFT Prior)	28.2	80.6	23.0	80.4	16.3	81.1	10.5	77.4	19.5	79.9	97.2
+PRETTY (Pseudo Prior)	18.3	68.9	17.3	72.2	11.6	70.4	5.0	65.6	13.1	69.3	73.9
+PRETTY (Refined Prior)	**<u>29.1</u>**	**81.4**	**22.9**	**80.4**	**<u>17.1</u>**	**81.1**	**<u>12.2</u>**	**<u>79.4</u>**	**20.3**	**<u>80.6</u>**	**100.4**
Mistral-7B											
Mistral-7B-Instruct	22.1	76.1	20.4	75.9	10.5	74.8	3.3	60.2	14.1	71.8	-
Mistral-7B	1.2	46.1	1.6	40.6	1.0	52.8	0.4	43.6	1.1	45.8	36.5
+PRETTY (SFT Prior)	20.1	73.3	20.7	75.1	11.0	74.7	6.8	67.3	14.7	72.6	113.8
+PRETTY (Pseudo Prior)	18.1	66.4	17.3	70.4	5.9	65.6	3.7	59.4	11.3	65.5	87.7
+PRETTY (Refined Prior)	**<u>28.3</u>**	**<u>78.8</u>**	**<u>22.3</u>**	**<u>78.5</u>**	**<u>14.2</u>**	**<u>78.6</u>**	**<u>13.6</u>**	**<u>80.6</u>**	**<u>19.6</u>**	**<u>79.1</u>**	**153.8**

Table 1: Translation performance of different models on Flores-101 subsets. **Bold** values indicate that the best performance among foundation models. The overall best results are <u>underlined</u>. "%SFT." denotes the relative performance compared to the best SFT model of each family.

For two cross-lingual task, the first $k = 2$ tokens are chosen as the prior tokens. This helps to avoid inadequate guidance from single non-informative tokens like punctuation or numbers. In the case of the pseudo prior, due to the back-off strategy, only one token is used for fair comparison. For POS tagging task, the strategy is more straightforward with only the first $k = 1$ label considered as the prior token.

4.2 Evaluation

To ensure the integrity of the output data from all models, we standardized the output by cleaning it in accordance with the specific output style of each model. Subsequently, we conducted a manual inspection to guarantee that only the required labels were retained.

Task-specific Metrics We use two metrics to evaluate the performance of translation quality: spBLEU[8] (Goyal et al., 2022) and COMET[9](Rei et al., 2020). We employed the ROUGE (Lin, 2004) and LaSE (Bhattacharjee et al., 2023) metrics for the evaluation of summarization quality. For the POS tagging task, we report both the precision score and F_1 score.

Relative Performance We further compute the ratio of the performance scores of the foundation model to the scores of the SFT model with the application of different strategies. This ratio serves

---
[8]https://github.com/mjpost/sacrebleu/
[9]https://github.com/Unbabel/COMET

Models	En-Zh			En-Hi			Uk-Pt			Ar-Ru			Avg.			%SFT.
	R2	RL	LS	R2	RL	LS	R2	RL	LS	R2	RL	LS	R2	RL	LS	All
Llama2-7B w/ Constant 1-Shot Demonstration																
Llama2-7B-Alpaca	7.0	12.4	11.9	1.7	10.7	17.3	1.5	6.1	5.8	0.1	0.5	1.3	2.6	7.4	9.1	-
Llama2-7B-Chat	6.3	11.6	8.7	1.5	11.7	27.1	2.5	8.3	7.1	0.0	0.3	0.2	2.6	8.0	10.7	-
Llama2-7B	9.3	16.6	29.2	1.6	10.2	15.3	0.8	4.0	1.9	0.6	4.1	15.5	3.1	7.6	12.1	262.0
+PRETTY (SFT Prior)	7.4	13.9	25.9	1.9	9.7	12.9	1.9	6.7	9.8	0.1	0.4	0.8	2.7	6.7	9.8	106.3
+PRETTY (Pseudo Prior)	8.0	14.5	29.1	1.4	9.9	14.5	2.5	9.1	13.6	1.2	5.9	23.5	3.3	8.5	15.4	387.5
+PRETTY (Refined Prior)	11.2	19.0	32.6	1.6	10.8	15.9	3.4	10.5	11.3	1.5	7.9	30.1	4.4	10.5	17.5	490.6
Mistral-7B w/ Constant 1-Shot Demonstration																
Mistral-7B-Instruct	5.9	12.2	17.2	1.0	10.3	23.4	1.5	6.2	17.7	0.4	2.6	12.8	2.2	7.8	17.8	-
Mistral-7B	12.3	20.9	44.5	1.6	10.6	17.6	4.8	12.9	27.7	1.8	6.5	23.3	5.1	11.2	21.6	206.1
+PRETTY (SFT Prior)	9.7	17.6	40.7	1.4	10.0	17.0	2.3	7.9	17.5	0.2	1.1	3.2	3.4	8.0	15.0	114.5
+PRETTY (Pseudo Prior)	9.9	17.5	41.0	1.4	9.9	17.4	3.1	11.6	35.1	1.7	7.9	32.9	4.0	10.2	23.5	195.8
+PRETTY (Refined Prior)	15.0	24.1	49.6	1.8	11.3	19.7	5.5	16.5	46.9	2.6	10.9	42.0	6.2	13.8	29.7	275.6

Table 2: Summarization performance of different models on CrossSum subsets. "R2/L" and "LS" refer to the ROUGE and LaSE score, respectively. **Bold** values indicate that the best performance among foundation models. The overall best results are underlined. "%SFT." denotes the relative performance compared to the best SFT model.

Models	Fr		Zh		Pt		Ru		Ar		Avg.	%SFT.
	Prec.	F_1	Prec.	F_1	Prec.	F_1	Prec.	F_1	Prec.	F_1	Prec.	All
Llama2-7B-Alpaca	48.2	42.8	38.6	36.3	40.7	35.9	42.3	36.7	34.4	30.8	38.7	-
Llama2-7B	45.0	37.9	39.8	36.2	39.8	33.2	42.5	33.8	36.5	32.1	37.7	97.4
+PRETTY (SFT Prior)	54.8	50.0	38.0	33.5	49.1	45.3	49.7	44.1	35.1	31.1	43.1	111
+PRETTY (Refined Prior)	59.3	54.8	43.0	38.8	54.5	50.6	55.3	49.2	44.0	39.6	48.9	126

Table 3: POS tagging performance of different Llama2 models on XGLUE subsets. **Bold** values indicate that the best performance among foundation models. The overall best results are underlined. "%SFT." denotes the relative performance compared to Alpaca model.

as a metric for assessing the extent to which the foundation model approximates the SFT model's performance when different strategies are applied.

4.3 Main Results

Machine Translation As shown in Table 1, for the machine translation task, we use up to two prior tokens as decoding guidance, allowing the base model to achieve performance comparable to that of a model after SFT. Moreover, in some language pairs, the translation performance outperforms SFT model when guided by Refined Prior tokens from a smaller model. For Llama2 model family, the prior tokens provided by the SFT model, although slightly less effective, still allow the foundation model to achieve 98% of the performance of SFT model. On the other hand, the use of pseudo labels derived from a dictionary exhibits the least effectiveness, yet this strategy still surpasses the results achieved through costly prompt engineering.

Cross-lingual Summarization The results presented in Table 2 indicate that the foundation model exhibited superior performance compared

to the SFT model in this in-context learning scenario. For prior-guided decoding, the performance of the foundation model was degraded when using prefix tokens from the SFT model, and the small performance gap in this setting suggests that the alignment achieved by the SFT model is relatively "superficial". Notably, the performance of Llama2 foundation model significantly improved when other priors were provided, even when using translated keywords as pseudo labels.

Non-English POS tagging The performance results of POS tagging task are presented in Table 3. These results align with the insights gleaned from the machine translation task, specifically regarding the strategy of prior token construction. Notably, for POS tagging task, the performance of SFT model on most language pairs falls short of the foundation model, suggesting that SFT detrimentally affect the knowledge learned at the pretraining stage. Encouragingly, when the foundation model empowered by auxiliary prior token surpasses the performance of SFT model as well as the prompting results of itself, highlighting the poten-

tial of our proposed method in mitigating the catastrophic forgetting problem associated with SFT.

5 Analysis and Discussion

5.1 Quality of Prior Tokens

To investigate the quality of prior tokens from different sources and how they impact the final performance, we further analyze why the prior tokens given by the SFT model are less effective than those from external auxiliary models in POS tagging task. Unlike the machine translation task, the positional result for the POS task is definite, so we are able to verify whether it corresponds to a ground truth label. The results in Table 4 confirm two points. First, even if the prior tokens provided by the SFT model are of low quality, the foundation model does not suffer from severe error propagation. Secondly, the final performance of proposed method is still associated with the quality of prior tokens. This suggests that prior tokens closely aligned with the ground truth can steer the foundation model towards a more accurate decision trajectory, thereby yielding superior performance.

	Fr	**Zh**	**Pt**	**Ru**	**Ar**
SFT Prior	18.3	18.3	3.74	16.3	12.1
Refined Prior	**88.9**	**88.9**	**88.54**	**87.7**	**79.6**

Table 4: Accuracy of prior tokens used in POS tagging task. SFT prior tokens are of inferior quality.

5.2 Choice of Prior Tokens

Based on the findings from the previous section, if incorrect labels used as prior tokens can still elicit the ability of foundation model, then could random prior tokens in the target language trigger cross-lingual generative capabilities? To investigate this, we attempted to use random tokens of different parts of speech as the prior tokens in the English-Chinese machine translation task. For instance, "Modal Prior" refers to the use of randomly picked modal verb in Chinese as the initial token.

The results shown in Table 5 indicate that the model could not be aligned to a better decision trajectory by these random prior tokens, whether they were function words or tokens with actual meaning. This supports the validity of our proposed methods for constructing prior tokens and also supplements previous findings. From this, we can summarize some rules about prior tokens: they can be of low

quality but should not be completely unrelated to the target sequence.

	spBLEU	**COMET**	**BLEU**
Llama2-7B	7.7	**72.01**	**16.1**
+ Modal Prior	**8.0**	68.29	16.0
+ Adverb Prior	6.4	63.72	13.1
+ Random Prior	6.2	57.11	11.5

Table 5: Comparison of translation performance using three types of random prior tokens.

5.3 Number of Prior Tokens

Figure 5 depicts the relationship between the number of preceding tokens provided and the resulting changes in translation performance. It becomes apparent that performance generally improves with the addition of more tokens. Additionally, we note that introducing two prior tokens appears to be a performance inflection point, which may be due to instances where the initial token is a punctuation mark or a number.
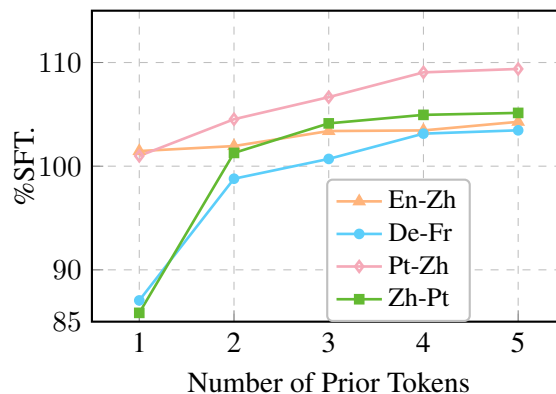

Figure 5: Impact of incrementally adding refined prior tokens on performance across Flores-101 subsets.

6 Conclusions

In this paper, we investigate and analyze the decision-making discrepancies between the foundation model and the SFT model within cross-lingual generation contexts. Drawing from our analysis, we introduce a novel cross-lingual alignment method that requires no additional training and is resource-efficient. The proposed method aligns the foundation LLM to perform comparably with the SFT model solely by utilizing prefix text as priors during generation. In the future, we aim to broaden our research to encompass additional alignment scenarios, such as those involving reinforcement learning from human feedback.

Limitations

The primary limitations of our study stem from the scope of model validation. Our research is limited to 7B models. Future endeavors should aim to extend the validation to a broader scope of models and incorporate various parameter scales to support the universality of our findings. Furthermore, the availability of language resources is still a practical problem, particularly for low-resource languages where access to Prior Token and Refined Token sources is limited. Despite these challenges, our experimental results indicate that Pseudo Prior tokens still exhibits promising potential. It is important to note, however, that the development of pseudo tags may require a dedicated investigation into the linguistic rules specific to each downstream task. This process is time-intensive and resource-demanding.

Acknowledgements

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170. European Association for Machine Translation.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564. Association for Computational Linguistics.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *ArXiv preprint*, abs/2312.09390.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *ArXiv preprint*, abs/2401.07037.

Yulong Chen, Huajian Zhang, Yijie Zhou, Xuefeng Bai, Yueguan Wang, Ming Zhong, Jianhao Yan, Yafu Li, Judy Li, Xianchao Zhu, and Yue Zhang. 2023. Revisiting cross-lingual summarization: A corpus-based study and a new benchmark with improved annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9332–9351. Association for Computational Linguistics.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

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*, abs/2207.04672.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023a. How abilities in large language models are affected by supervised fine-tuning data composition. *ArXiv preprint*, abs/2310.05492.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023b. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv preprint*, abs/2304.06767.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023c. A survey on in-context learning.

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.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1813–1829, Singapore. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Kaixin Lan, Tao Fang, Derek F. Wong, Yabo Xu, Lidia S. Chao, and Cecilia G. Zhao. 2024. Focus: Forging originality through contrastive use in self-plagiarism for language models. In *Findings of ACL 2024*.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *ArXiv preprint*, abs/2402.10110.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2023. Rain: Your language models can align themselves without finetuning. In *The Twelfth International Conference on Learning Representations*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *ArXiv preprint*, abs/2312.01552.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *ArXiv preprint*, abs/2312.15685.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv preprint*, abs/2308.08747.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

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.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Xin He, Wanshun Chen, and Longyue Wang. 2024. Anchor-based large language models. In *Findings of ACL 2024*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *ArXiv preprint*, abs/2304.03277.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv preprint*, abs/2305.18290.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Andrea Schioppa, Xavier Garcia, and Orhan Firat. 2023. Cross-lingual supervision improves large language models pre-training. *ArXiv preprint*, abs/2305.11778.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*, volume 36, pages 2511–2565. Curran Associates, Inc.

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, 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*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23. Association for Computational Linguistics.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *ArXiv preprint*, abs/2308.12067.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *ArXiv preprint*, abs/2402.04333.

Haoyun Xu, Runzhe Zhan, Derek F. Wong, and Lidia S. Chao. 2024. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model. *ArXiv preprint*, abs/2403.11621.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *ArXiv preprint*, abs/2304.05302.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 807–820. Association for Computational Linguistics.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *ArXiv preprint*, abs/2305.11206.

A Related Work

Data-Efficient Alignment The "*superficial alignment hypothesis*" proposed by Zhou et al. (2023) claims that SFT is a process that repurposes pre-training knowledge and reshapes outputs to satisfy specific criteria. Consequently, various studies have explored data-efficient alignment in the context of SFT (Liu et al., 2023). Among these, model-centric approaches (Kung et al., 2023; Xia et al., 2024; Li et al., 2024) leverage either the LLM itself or external models as scorers to select high-quality SFT data from the data store. The heuristic methods (Lu et al., 2023; Wei et al., 2023) carefully choose desired SFT data based on a set of scorers or metrics. The goal of data-efficient alignment methods is to reduce the number of SFT samples. However, for low-resource languages or tasks, the quality and scale of the data pool may not be sufficient to perform effective data selection.

Self-Alignment in LLM According to superficial alignment hypothesis, pre-training data may encompass task-specific knowledge, making it feasible to align LLMs with themselves. Sun et al. (2023) presented an alignment pipeline that minimizes the cost of human annotations, demonstrating the possibility of "self-alignment". Li et al. (2023) introduced a multi-round rewindable inference framework for identifying the optimal generation results of unaligned LLM, showcasing the extensive potential of the LLM's decision space. Lin et al. (2023) further revealed the connections between base LLM and SFT-ed LLM on general-purpose tasks and introduced an in-context alignment approach for base LLM without the need for re-training. These studies primarily focus on monolingual tasks, overlooking the impact of SFT on the effectiveness of non-English and cross-lingual alignment. This area remains ripe for exploration.

B Latest Benchmark

The composition of the pre-training corpus for Llama2 is indeed unclear as Meta AI has not disclosed specifics in their technical report (Touvron et al., 2023). Consequently, we cannot confirm the inclusion of Flores-101 datasets in the pre-training phase. To mitigate this uncertainty, we conducted additional experiments using the WMT23 machine translation benchmark. Since Llama2 was released on July 18, 2023, and the WMT23 News Shared Task submission[10] deadline was July 20, 2023, with the reference translations released on October 10, 2023, it is highly unlikely that WMT23 data influenced Llama2's pre-training.

For these experiments, we used conventional BLEU as our evaluation metric because the Flores-101 benchmark introduces a specialized BLEU score variant known as spBLEU, which diverges from the conventional BLEU by adopting the sentencepiece model (Kudo and Richardson, 2018) for tokenization. As shown in Table 6, the experimental results indicate a consistent trend and support similar conclusions to those obtained from previous benchmarks.

C Detailed Setup

C.1 Llama2 Alpaca

To avoid the need of fine-tuning for Llama2 family, we initially considered using Llama2-Chat. However, it is important to note that Llama2-Chat is a version that has been fine-tuned using both SFT and Reinforcement Learning with Human Feedback (RLHF; Christiano et al. 2017), as it has been aligned with human preferences for helpfulness and safety. In practice, we observed that the Llama2-Chat model sometimes overly prioritizes safety, leading to instances where it refrains from generating responses due to its strict safety protocols. Consequently, we opted for a SFT model that was fine-tuned using Alpaca dataset to ensure more consistent outputs and used it as a source of obtaining SFT priors.

Hyper-parameters The hyper-parameter settings are presented in Table 7. This experiment was conducted using the DeepZero 2.

C.2 Prompting Strategies

Machine Translation

- **Normal**: *Translate the following sentences from {Source Language} to {Target Language}. {Source Text}*

- **Prompting**: *{Source Language}: {Source Text} \n{Target Language} Translation:*

Cross-lingual Summary

- *Please write a concise summary of the text, return your responses with one line that cover*

Models	En-Uk		En-Zh		En-He		Avg.		%SFT.
	BL.	CoM.	BL.	CoM.	BL.	CoM.	BL.	CoM.	All
Llama2-7B									
Llama2-7B-Alpaca	<u>12.8</u>	<u>75.0</u>	<u>25.9</u>	<u>77.6</u>	2.3	46.3	<u>13.7</u>	<u>53.1</u>	-
Llama2-7B	0.2	34.8	10.1	65.3	0.3	41.2	3.5	36.2	59.5
+PRETTY (SFT Prior)	12.1	70.3	24.7	76.4	1.3	43.4	12.7	50.7	94.9
+PRETTY (Pseudo Prior)	9.4	63.9	22.8	72.4	0.7	40.7	11	47	86.8
+PRETTY (Refined Prior)	**12.5**	**72.3**	**24.7**	**76.8**	**<u>2.3</u>**	**<u>48.3</u>**	**13.2**	**52.6**	**98.5**

Table 6: Translation performance of different models on WMT23 subsets. **Bold** values indicate that the best performance among foundation models. The overall best results are <u>underlined</u>. "%SFT." denotes the relative performance compared to Alpaca model.

Hyper-Parameters	Settings
Learning Rate	2e-5
LR Scheduler	Cosine
Epochs	3
Seed	42
Warmup Ratio	0.03
Weight Decay	0
Cumulative Batch Size	64
LoRA Alpha	64
LoRA Rank	128
LoRA Dropout	0.05
Trainable Modules	q_proj,v_proj,k_proj,o_proj, gate_proj,down_proj,up_proj

Table 7: Settings used for fine-tuning Llama2-7B foundation model with Alpaca data.

the key points of the text.\n Input: {Source Text} \n Output: {Target Language} Summary:

POS tagging

- *Please provide the POS tags for each word in the input sentence. The input will be a list of words from the sentence. The output format should be a JSON array, where each key-value pair includes a word from the input list and its corresponding POS tag from the set of labels: [{PoS Tags}]\nNote: Your response should include only a JSON array, presented in the order that the words appear in the input sentence. Each key-value pair should contain the word and its POS label.*

C.3 Success Rate of Vanilla Prompting

We randomly selected 500 sentences from the Flores-101 benchmark for four translation directions and randomly used ten diverse prompts for the corresponding translations. In addition to prompts proposed by Bawden and Yvon (2023), the used prompts are as follows:

- *Translate the following sentences from {Source Language} to {Target Language}.*

- *Can you help me translate this sentence from {Source Language} to {Target Language}?*

- *What is the translation of sentence in {Target Language}? {Source Text}*

- *Infill in {Target Language}: {Source Text}*

- *Please provide machine translation text for me.\n{Source Language}: {Source Text}\n {Target Language}:*

To determine whether a translation was generated (regardless of its translation quality), we employed automated tools[11] and then checked the results by human. We then calculated the success rate of translation requests. The results, which will be detailed in table below, indicate that foundation LLMs exhibit considerable instability in generating the target language when presented with different prompts. In most cases, the LLMs failed to produce the desired language output.

	En-Zh	En-Uk	Zh-Pt	De-Fr
Success Rate ↑	28.6%	21.6%	39.6%	37.4%

Table 8: Success Rate of translation prompting with different instructions.

---

[11] https://github.com/Mimino666/langdetect

C.4 Pseudo Labels

We collected lexical data from various sources, including Wiktionary[12], MUSE[13](Lample et al., 2018), and PanLex[14], and merge their entries into one dictionary. Table 9 presents the detailed information of compiled dictionaries.

Languages	# Entries	Source
En-Zh	31,443	Wiktionary, MUSE
Zh-En	13,667	MUSE
En-Uk	32,685	MUSE
Uk-En	34,888	MUSE
De-Fr	61,527	Wiktionary, MUSE
Pt-Zh	322,987	PanLex

Table 9: The detailed statistical information of the dictionary used for constructing Pseudo Priors.

---

[12]https://www.wiktionary.org/
[13]https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries
[14]https://panlex.org/source-list/