Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts

Anonymous ACL submission

Abstract

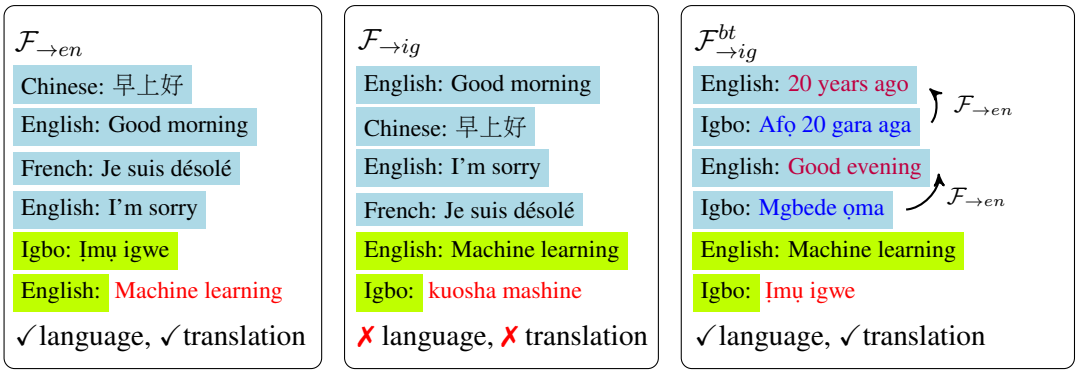
Large language models (LLMs) are known to perform tasks by simply observing few exemplars. Moreover, competent generative capabilities of LLMs are observed mostly in high-resource languages, while their performances among under-represented languages fall behind due to pre-training data imbalance. To elicit LLMs' ability onto low-resource languages without any supervised data, we propose to assemble synthetic exemplars from a diverse set of high-resource languages. These prompts can directly induce generative capabilities in low-resource languages and serve as intra-lingual exemplars to even improve tasks in these languages. Our zero-shot/unsupervised prompting method performs on par with supervised few-shot learning in LLMs of different sizes for translations between English and 34 Indic and African languages, and surpasses supervised prompting in non-English tasks. The method also significantly improves low-resource performances in many other intra-lingual tasks like summarization (XLSum), question answering (XQUAD & TydiQA) and conversational instruction following (Sea-Bench).

1 Introduction

Recent scaling effort in foundation large language models (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Touvron et al., 2023a) with massive pre-training data has enabled them to learn a broad range of natural language tasks through few-shot in-context learning, where a few input-output exemplars are concatenated to the test input to prompt the model to predict the output and no gradient update of the model is performed. While most LLMs are pre-trained with multilingual corpora in addition to the gigantic English corpus, and have been shown to demonstrate impressive abilities in other languages (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022), they only excel in high-resource languages, such as French. Further, they may still require pivoting the inputs into English, that is, performing tasks in English first before reverting the response back to native outputs (Shi et al., 2022; Huang et al., 2023). Improving LLMs' abilities in extremely low-resource languages can be even more challenging, particularly where the data coverage is less than 0.0001% (Scao et al., 2022). We also found that the models may confusedly respond in a wrong language or struggle with low-resource non-latin scripts due to overly fragmented tokenization, where words are broken into many byte-level tokens.

In this work, we propose Linguistically-Diverse Prompting or LDP (codes will be made available), a technique that promotes an LLM to perform generative tasks in low-resource languages by *demonstrating few-shot exemplars in a diverse set of high-resource languages*. This method works in both unsupervised setup with foundation base LLMs (Scao et al., 2022; Touvron et al., 2023a) and pseudo-zero-shot setup instruction-tuned counterparts (Ouyang et al., 2022; Muennighoff et al., 2022; OpenAI, 2023), by synthetically creating few-shot examples from zero-shot prompting. A manifestation of LDP for unsupervised translation task is shown in Figure 1, where we gather a small set of synthetic $X \rightarrow$En exemplars from a diverse set of high-resource languages using a pre-trained unsupervised MT model (Tran et al., 2020). Figure 2, meanwhile, later explain LDP in other generalized adoptions in many other tasks. Our method is based on the following empirical observations of LLMs: (*i*) in-context exemplars may play a larger role in helping the model *locate* the task in its pre-trained knowledge (Xie et al., 2021), (*ii*) LLMs possess dominant abilities in English while they may lag behind in other lower-resource languages (Ouyang et al., 2022; Touvron et al., 2023a; Huang et al., 2023; Shi et al., 2022).

Our method is shown to perform on par with supervised prompting in unsupervised translation tasks between English and 13 Indic and 21 African



(a) X→En (b) En→X (c) En→X with back-translation

Figure 1: LDP prompting for unsupervised translation. (1a) $\mathcal{F}_{\rightarrow en}$ translates from any language into English by concatenating the fixed linguistically-diverse shots and input text to prompt LLMs to generate the correct translation. (1b) Similarly $\mathcal{F}_{\rightarrow ig}$ translates English into Igbo, but with low accuracy. (1c) $\mathcal{F}^{bt}_{\rightarrow ig}$ translates English to Igbo using synthetic intra-lingual exemplars generated from unlabeled target-language data with $\mathcal{F}_{\rightarrow en}$.

low-resource languages, with BLOOM (Scao et al., 2022) and InstructGPT (text-davinci-003) (Ouyang et al., 2022) models. Furthermore, adapting our method to X→Y non-English directions even outperforms supervised promptings by up to 3 chrF++ in pairs involving low-resource languages. In multilingual summarization tasks (Narayan et al., 2018), our zero-shot LDP method outperforms both basic prompting and other English-pivoting methods by up to 4 ROUGE-L and is generally favored by GPT-4-EVAL (Liu et al., 2023). With GPT-3.5, our method considerably improves zero-shot question answering XQUAD (Artetxe et al., 2019) and TyDiQA (Clark et al., 2020) tasks in 6 languages by up to 14 F1 and matches few-shot prompting performance. Our method can even enable Llama-2 **base** (Touvron et al., 2023b) to perform conversational instruction following tasks and improve the chat model in Sea-Bench (Nguyen et al., 2023) for 2 languages that were not instruction-tuned.

2 Related Work

Large language models (LLMs) display outstanding capabilities because they are pre-trained on massive amounts of internet text data (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Touvron et al., 2023a,b). Without any gradient update, base LLMs are able to perform in-context learning by simply observing a list of high-quality input-output exemplars (Brown et al., 2020; Wei et al., 2023). This technique works across many tasks that involve language understanding, reasoning and generation (Brown et al., 2020; Wei et al., 2022; Shi et al., 2022). Much research has been conducted to understand in-context learning. Some suggest that the models secretly perform gradient descent on the exemplars (Dai et al., 2022), while others demonstrate that most of the knowledge is learned during pre-training, and the exemplars are only to provide evidence for the model to *locate* the intended task via a Bayesian inference process (Xie et al., 2021; Min et al., 2022; Zhou et al., 2023).

Most LLMs are trained with multilingual corpora (Wenzek et al., 2020), even if these make up a tiny fraction of the large English corpora (Radford et al., 2019; Brown et al., 2020). Despite that, LLMs still exhibit strong multilingual capabilities with high-resource languages like French, German and Chinese, often with the help of English-pivoting using supervised translation systems (Shi et al., 2022) or prompting the model to firstly generate intermediate English context (Huang et al., 2023). BLOOM (Scao et al., 2022) is one of the LLMs trained with the most number of languages, whose ROOTS corpus consists of data from 46 languages (Laurençon et al., 2022). The ROOTS corpus includes 34 Indic and African languages regarded as low-resource, with each language having a pre-training coverage of less than 1% in Hindi for the Indic group, to $2e^{-5}$% in Tumbuka for the African group, as shown in Figure 5 in the Appendix. Exclusively only for unsupervised translation tasks, our linguistically-diverse prompting (LDP) strategy is also an English-pivoting method, but it is different from other cross-lingual counterparts (Shi et al., 2022; Huang et al., 2023) in that while others only pivot inputs to English intermediates, we use in-context pairs between En-

glish and a diverse set of high-resource languages to promote the intended task in the target low-resource language. For other intra-lingual tasks like instruction following, where input and output are both expected to be in the same language, our LDP helps prevent English-tuned models Touvron et al. (2023b) from responding with English answer given a non-English query.

Part of our work also intersects with unsupervised multilingual machine translation (UMT), where back-translation is proven to be effective (Edunov et al., 2018; Lample et al., 2018; Conneau and Lample, 2019; Liu et al., 2020; Nguyen et al., 2022b), along with other methods (Tran et al., 2020; Nguyen et al., 2022a). English-pivoting is also prominent in the realm of machine translation, where training models on high-resource En↔X bitext improves lower-resource En↔Y tasks (Garcia et al., 2020, 2021). Analyses of machine translation using LLMs have also been done. Hendy et al. (2023) show that GPT models can perform competitively alongside the best MT models. Zhu et al. (2023) focus on optimizing supervised exemplars selection strategies, while Sia and Duh (2023) discover that using specific coherent prompts for each input helps improve performance. Nonetheless, such studies only consider supervised instruction-tuned models (Ouyang et al., 2022; Muennighoff et al., 2022), which may risk test-set contamination (Muennighoff et al., 2022). Thus, there is still limited research involving low-resource languages in completely zero-shot setups. As such, since low-resource languages may not enjoy the privilege of having large unlabeled data to conduct searching, only random selection is used in this study, while optimal exemplar selection is out of scope.

3 Method

3.1 Linguistically-Diverse Prompting (LDP)

Our method is inspired from two observations: (*i*) LLMs may have already learned most of the task concepts implicitly during pre-training, and that in-context exemplars play a larger role in providing evidence for the model to identify the intended task (Xie et al., 2021; Min et al., 2022; Zhou et al., 2023). (*ii*) LLMs perform generative tasks dominantly well in only a few major languages (*e.g.,* English), whose pre-training data is highly abundant (Brown et al., 2020; Touvron et al., 2023a).

The main principle of LDP is using few-shot exemplars from a diverse set of high-resource languages to elicit generative abilities in low-resource languages. Such idea can be designed and catered differently depending on use cases. In this paper, we present two manifestations of LDP for (*i*) cross-lingual tasks (*e.g.,* translation) where the input and output are in different languages and (*ii*) intra-lingual tasks (*e.g.,* summarization, question answering) where both are in the same language.

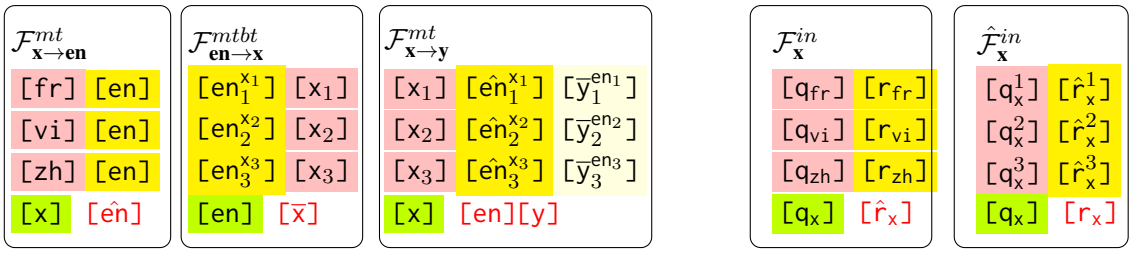
Figure 1 illustrates how our LDP method aims to take advantage of the aforementioned observations in the case of unsupervised translation tasks. Particularly, we prompt the model to identify the task of "translating from *any language* X into E", by demonstrating pairs from "every language" to E. Practically, shown in Figure 1a, we use synthetic pairs from diverse high-resource languages as exemplars to prompt the models to translate the target low-resource language X (*e.g.,* Igbo) into English (En) with high quality. Such diverse set of prompt languages should include various script types ranging from Latin alphabets to logograms. Figure 1b shows that applying the same technique for En→X task may results in incorrect translation. In Figure 1c, however, we leverage LDP to translate unlabeled texts of target X language into En, forming back-translated synthetic pairs to prompt the model to translate from En to X with higher quality. This is because the target-side distribution is now realistic and consistently close to the true target distribution, which has been shown to be crucial for in-context learning (Xie et al., 2021).

3.2 LDP for Cross-lingual Tasks (Translation)

For tasks where the input and output are in different languages, such as translation, we adopt LDP for $X \rightarrow E$, $E \rightarrow X$ and $X \rightarrow Y$ (where $X, Y \neq E$), differently, as shown in Figure 2a, where we assume $E =$ English (En) for better understanding.

$X \rightarrow E$ task. As mentioned above, we first gather n $Z_i{\rightarrow}E$ exemplar pairs (s_{Z_i}, t^i_E) with $Z_i \in \mathcal{Z}$ and $\mathcal{Z}$ being a diverse set of languages with various writing systems, lexical and regional characteristics, such as Chinese (Zh), and $Z_i \notin \{X, E\}$. Such exemplars can be collected from unlabeled data of the respective language Z_i and using unsupervised MT models to translate them into E (for unsupervised tasks), or from labeled few-shot pairs (for zero-shot tasks). From that, we can perform translation of an input s_X into E with an LLM (θ) by conditioning the LDP prompts as:

$$\mathcal{F}^{mt}_{X\rightarrow E}(s_X) \sim p_\theta(y|s_X, s_{Z_1}, t^1_E, .., s_{Z_n}, t^n_E) \quad (1)$$



(a) LDP for translation for $X\to$En, En$\to X$ and $X\to Y$.

(b) LDP for intra-lingual tasks.

Figure 2: Illustrations LDP for $X\to$En, En$\to X$ and $X\to Y$ cross-lingual translation (2a) and general intra-lingual tasks (2b). For $X\to$En, the colored box [z] represents an unlabeled text in language z, [en] represents its corresponding En translation, while [x] stands for the test input in language x and uncolored box [ên] represents model outputs. For En$\to X$, [en^{x}] represents En text translated with $\mathcal{F}^{mt}_{\text{x}\to\text{en}}$. For $X\to Y$, [$\bar{\text{y}}^{\text{en}}$] represents a text in language y translated from En text [$\hat{\text{en}}^{\text{x}}$]. Similarly for intra-lingual tasks like summarization (2b), [$\hat{\text{r}}_{\text{z}}$] represents a response in language z for query [q_{z}].

$E \to X$ task. We leverage $\mathcal{F}^{mt}_{X\to E}$ to build intra-lingual prompts with unlabeled data from the target language X. Specifically, given m unlabeled texts $s^j_X \in \mathcal{D}_X$ with $\mathcal{D}_X$ being a monolingual corpus in language X, we produce synthetic back-translation (BT) target $s^j_E = \mathcal{F}^{mt}_{X\to E}(s^j_X)$. Then, we use the BT synthetic pairs as exemplars for $E \to X$ tasks for a test input s_E:

$$\mathcal{F}^{mtbt}_{E\to X}(s_E) \sim p_\theta(y|s_E, s^1_E, t^1_X, ..., s^m_E, t^m_X) \quad (2)$$

The intra-lingual exemplars with the same language in the target side helps the model locate the intended language to generate more effectively than a standard language tag, as these exemplars show the model *what the intended language looks like*.

Note that we could also use $\mathcal{F}^{mtbt}$ for $X \to E$ ($\mathcal{F}^{mtbt}_{X\to E}$) by simply swapping the direction of the (s^j_E, t^j_X) to (s^j_X, t^j_E). However, we found in the experiments that both $\mathcal{F}^{mt}$ and $\mathcal{F}^{mtbt}$ perform similarly and on par with supervised prompting for the $X \to E$ task, suggesting that we do not need any supervised or unlabeled data to translate any language into English. Furthermore, in Section 4.6, we demonstrate that we can even omit these back-translation exemplars entirely with non-BT $\mathcal{F}^{mt}$ LDP by using native language tags.

$X \to Y$ task. We leverage $\mathcal{F}^{mtbt}_{X\to E}$ and $\mathcal{F}^{mtbt}_{E\to X}$ to build E-pivoting triplets from unlabeled text from the source side. Specifically, given unlabeled text $s^j_X \in \mathcal{D}_X$ in X language, we back-translate them into $s^j_E = \mathcal{F}^{mtbt}_{X\to E}(s^j_X)$, which we then use to produce $s^j_Y = \mathcal{F}^{mtbt}_{E\to Y}(s^j_E)$. This process forms triplets $[s^j_X, s^j_E, s^j_Y]$ to prompt the model to generate intermediate E translation before producing the final result in Y. Formally, given an input s_X, the translation in Y is computed as:

$$\mathcal{F}^{mt}_{X\to Y}(s_X) \sim p_\theta(y|s_X, s^1_X, s^1_E, s^1_Y, ..., s^m_X, s^m_E, s^m_Y) \quad (3)$$

Unsupervised fine-tuning. The $\mathcal{F}^{mt}_{X\to E}$ prompting method also allows us to create larger-scale synthetic X-E data from unlabeled corpora to fine-tune the model for translation tasks without any in-context prompt at inference time. Specifically, we use the `[input]<lang-tag>[output]` template to construct multilingual training samples with the generated data pairs from multiple low-resource languages. We fine-tune the query-key-value linear weights of all attention layers, which account for 20-30% of the total parameters to avoid overfitting.

3.3 LDP for intra-lingual tasks

For intra-lingual tasks, where the input and output are in the same language, such as summarization and question answering, we adopt LDP in zero-shot setups for instruction-tuned models (Ouyang et al., 2022) differently as illustrated in Figure 2b. Formally, given a query q_X in the target language X and n in-domain queries q_{Z_i} with $Z_i \in \mathcal{Z}$ and $\mathcal{Z}$ being a diverse set of high-resource languages, we use standard or augmented zero-shot prompting strategies h (Huang et al., 2023; Wei et al., 2022) to obtain responses $r_{Z_i} = h(q_{Z_I})$. We then use the synthetic query-response pairs (q_{Z_i}, r_{Z_i}) as LDP in-context exemplars to compute the target-language response r_X for q_X as:

$$\mathcal{F}^{in}_X(q_X) \sim p_\theta(y|q_X, q_{Z_1}, r_{Z_1}, .., q_{Z_n}, r_{Z_n}) \quad (4)$$

	Ind-En	En-Ind	Afr-En	En-Afr
Base BLOOM-175B				
Supervised-8-shot	47.31	34.66	28.64	14.93
Unsupervised-LDP	47.62	34.54	28.72	14.57
Base BLOOM-7B				
Supervised-8-shot	39.86	24.02	21.51	11.27
Unsupervised-LDP	39.88	24.41	20.47	12.04
Fine-tune	42.19	32.72	21.14	15.73
Supervised InstructGPT (text-davinci-003)				
Zero-shot	35.37	20.71	27.10	15.45
Supervised-6-shot	37.07	24.74	31.51	19.22
Unsupervised-LDP	38.45	25.17	31.92	19.51
Supervised upperbound				
NLLB-200 distilled	*61.00*	*46.77*	*48.42*	*39.18*

Table 1: Averaged performances of different prompting techniques across various model sizes and types, namely BLOOM (Scao et al., 2022) and InstructGPT text-davinci-003 (Brown et al., 2020; Ouyang et al., 2022), in translation tasks between English (En) and 13 Indic (Ind) and 21 African (Afr) low-resource languages present in the ROOTS corpus (Laurençon et al., 2022). SacreBLEU scores are provided in the Appendix.

Similar to $E \rightarrow X$ translation task, we then use zero-shot $\mathcal{F}^{in}_X$ to generate synthetic intra-lingual prompts from m unlabeled queries $q^j_X \in \mathcal{D}_X$ by producing responses $r^j_X = \mathcal{F}^{in}_X(q^j_X)$ in X language. After that, we compute the final response for the input q_X with $\hat{\mathcal{F}}^{in}_X$ as:

$$\hat{\mathcal{F}}^{in}_X(q_X) \sim p_\theta(y|q_X, q^1_X, r^1_X, .., q^m_X, r^m_X) \quad (5)$$

4 Experiments

In this section, we evaluate our method in various translation (sections 4.1 and 4.2), summarization (4.3), question answering (4.4) and instruction-following (4.5) across different settings and languages. We also conduct extensive analyses to provide further insights into our method (4.6).

4.1 Low-resource ↔ English Translation

As the ROOTS corpus (Laurençon et al., 2022) that BLOOM (Scao et al., 2022) was pre-trained on offers the most diverse language coverage with open-sourced transparency, we tested our methods mainly with the BLOOM model on 13 Indic (Ind) languages and 21 African (Afr) languages present in the ROOTS corpus. We also conduct experiments with supervised InstructGPT (text-davinci-003) (Ouyang et al., 2022) to provide further references. As not much detail about text-davinci-003 has been disclosed, its results are only to compare prompting techniques within the same model and not between models. Specific details about languages and test sets are provided in the Appendix. Following Costa-jussà et al. (2022), we report results in mainly chrF++ (Popović, 2015), which is a universal metric for all languages, while also reporting SacreBLEU (Post, 2018) in the Appendix.

In terms of methodologies, for supervised prompting, we collect as many supervised pairs as the models can fit within their context lengths (8 for BLOOM and 6 for text-davinci-003). We use `<src>[input]\n<tgt>[output]` as the prompt template, where `<src>` and `<tgt>` are the language tag names in English. For our unsupervised linguistically-diverse prompting (LDP) method, we use 4 LDP Z_i↔En pairs from Arabic (Ar), Chinese (Zh), Vietnamese (Vi) and French (Fr) to conduct $X \rightarrow E$ synthetic data generation with $\mathcal{F}^{mt}_{X\rightarrow E}$ before using them as intra-lingual prompts for the target pair with $\mathcal{F}^{mtbt}_{X\leftrightarrow E}$, as explained in Section 3. For LDP, we do not include the language tags in the prompts as they offer no benefit. In our fine-tuning experiment, we use $\mathcal{F}^{mt}_{X\rightarrow E}$ to generate synthetic training data from various unlabeled sources (Wenzek et al., 2020) to fine-tune BLOOM-7B.

Table 1 shows the averaged chrF++ scores for translations between English and 13 Indic and 21 African low-resource languages across different prompting techniques with various models. Noticeably, our unsupervised-LDP method performs on par with supervised prompting across all language groups and LLM models. This indicates that the synthetic prompts generated by our $\mathcal{F}^{mt}_{X\rightarrow E}$ technique are as good as supervised prompts when serving as few-shot exemplars,[1] thanks to the LLMs' outstanding ability in English. Furthermore, fine-tuning a 7B model with data generated by itself helps the model to advance towards the performance of its 175B sibling, especially for En→X direction. This suggests that fine-tuning the model on more low-resource language data improves generative abilities in such languages.

For text-davinci-003, we observe the same pattern when comparing supervised and unsupervised-LDP. It is interesting to see that GPT's scores for Indic languages are lower than BLOOM but higher for African languages, despite the fact that the African languages are likely to have less data coverage. One of the reasons may be the token frag-

[1]The synthetic outputs themselves are still lower-quality than supervised translations or the ground truths.

	High-High		High-Low				Low-Low			
	Vi-Fr	Fr-Vi	Zh-Ne	Ne-Zh	Es-Pa	Pa-Es	Ta-Sw	Sw-Ta	Te-Sw	Sw-Te
Foundation BLOOM-175B										
Supervised-8-shot	52.17	51.50	30.91	17.83	25.67	37.71	31.45	31.81	31.46	25.84
Unsupervised-LDP	52.66	50.24	31.61	18.34	27.85	39.51	34.61	34.47	32.14	30.57
Supervised InstructGPT (text-davinci-003)										
XLT (Huang et al., 2023)	51.16	44.84	28.56	13.26	23.61	34.18	24.20	25.46	24.89	23.48
Unsupervised-LDP	51.19	45.80	28.67	15.80	25.40	35.02	27.24	27.70	28.95	25.12

Table 2: chrF++ translation scores for X→Y non-English tasks across high-high, high-low and low-low groups.

mentation issue which we explain in the Appendix. Similarly, we observe LDP performs competitively with supervised prompting on 20 European languages with LLaMA (Touvron et al., 2023a), which we also detail in Table 9 in the Appendix.

4.2 Non-English-centric Translation

For non-English $X \rightarrow Y$ directions, we compare our unsupervised method $\mathcal{F}^{mt}_{X \rightarrow Y}$ with supervised prompting in three categories: High-High resource languages with Vi and Fr, High-Low resource between Zh, Es, Ne (Nepali) and Pa (Punjabi), and Low-Low resource languages with Sw (Swahili), Ta (Tamil) and Te (Telugu). We use the same model and evaluation pipelines as explained Section 4.1. For this experiment, we evaluate on the FLoRes-200 devtest sets Costa-jussà et al. (2022). As reported in Table 2, our unsupervised LDP technique also performs on par with supervised prompting in High-High Vi-Fr pairs. More interestingly, for High-Low and Low-Low language pairs, our unsupervised method even outperforms supervised prompting for these languages by up to 5 chrF++, largely thanks to the presence of English intermediate translations in the exemplars.

4.3 Zero-shot Summarization

	Es	Id	Sw	So	Mr
Basic	12.7/2.9	12.8/2.5	12.2/2.3	11.5/3.0	4.1/2.9
XLT	17.7/3.9	17.6/3.4	20.5/3.1	18.5/3.9	10.3/3.8
LDP	18.1/4.1	18.6/3.6	21.8/3.3	19.0/4.0	10.0/3.9
LDP+U	**18.1/4.2**	**24.8/3.8**	**23.5/3.3**	**19.3/4.0**	**11.4/3.9**

Table 3: ROUGE-L / GPT-4-EVAL scores (1-5 ratings) of different prompting techniques using InstructGPT text-davinci-003 for zero-shot summarization in high-resource (Es, Id) and low-resource (Sw, So, Mr) in the XL-sum summarization task (Narayan et al., 2018).

We extend our LDP method to multilingual summarization by combining intral-lingual LDP (section 3.3) with cross-lingual prompting (XLT) (Huang et al., 2023) using the supervised text-davinci-003 model. XLT is a recent English-pivoting instruction proposed by Huang et al. (2023). We follow the LDP adoptions for intral-lingual tasks with (LDP+U or $\hat{\mathcal{F}}^{sum}$) and without (LDP or $\mathcal{F}^{sum}$) unlabeled data, as described in Section 3.3. We conduct evaluation on the XL-Sum benchmark (Narayan et al., 2018) in both high-resource (Es-Spanish, Id-Indonesian) and low-resource (Sw, So-Somali, Mr-Marathi) languages. We evaluate the models with ROUGE-L (Lin, 2004) and GPT-4-EVAL (Liu et al., 2023). GPT-4-EVAL is a GPT-4 based metric that recently scores best in human judgement alignment. We compare our methods with XLT, and basic instruction. As shown in Table 3, our LDP methods outperform standard XLT across all languages by up to 7 ROUGE-L and exceeds basic prompting by large margins. Our methods are also consistently preferred by GPT-4-EVAL with higher ratings.

GPT-3.5	XQUAD			TydiQA			
	Ar	Hi	Th	Ar	Bn	Fi	Ru
3-shot	69.9	69.3	53.8	27.7	20.2	34.7	16.8
0-shot	52.9	45.9	26.3	19.1	5.7	21.7	12.3
w/ LDP	69.8	69.0	54.0	23.2	18.9	32.6	17.0

Table 4: Multilingual question answering F1 scores of ChatGPT (GPT 3.5) using different prompting techniques across different languages in the XQUAD and TydiQA benchmarks.

4.4 Zero-shot Question Answering

Our method also works well for multilingual comprehension and world-knowledge question answering with the XQUAD (Artetxe et al., 2019) and no-context TydiQA (Clark et al., 2020) benchmarks respectively. We demonstrate this with ChatGPT-3.5 across Arabic (Ar), Hindi (Hi), Thai (Th), Bengali (Bn), Finnish (Fi) and Russian (Ru). We select supervised exemplars from En, Vi, and Zh as LDP

	Task		Instruct		NatQA	
	Vi	Id	Vi	Id	Vi	Id
ChatGPT (3.5)	7.47	7.85	9.42	9.80	9.05	9.45
LLama2-13B						
-Chat	6.45	5.45	6.15	7.67	4.95	5.65
-Base w/ LDP	3.87	2.61	4.65	7.05	4.80	6.10
-Chat w/ LDP	3.83	6.54	8.57	8.72	4.94	6.85

Table 5: GPT-4 rated LLM-as-a-judge scores (Zheng et al., 2023) of different models and prompting strategies for task-solving (**Task**), instruction-following (**Instruct**) and natural question answering (**NatQA**) categories in the Sea-bench set (Nguyen et al., 2023) for Vi and Id.

	gu	mr	pa	kn	ne	te	ml	ur	ta	bn	hi	##	chrf++
gu	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	19.64
mr	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.1	8.98
pa	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.6	1.01
kn	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	15.27
ne	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.4	0.0	25.28
te	0.0	0.0	0.0	0.1	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.1	19.88

(a) LDP without back-translation $\mathcal{F}^{mt}_{\text{En}\rightarrow X}$.

	gu	mr	pa	kn	ne	te	ml	ur	ta	bn	hi	##	chrf++
gu	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.18
mr	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	18.78
pa	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.35
kn	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.82
ne	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.14
te	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	33.18

(b) LDP with back-translation $\mathcal{F}^{mtbt}_{\text{En}\rightarrow X}$.

Figure 3: Probabilities of whether the BLOOM model generates the right language for En$\rightarrow X$ task using LDP without (3a) and with (3b) intra-lingual BT prompts. Columns indicate the languages the model generates into while rows are the languages it is *supposed* to generate. ## are other languages.

pairs for XQUAD and similarly exemplars from En, Id, Ko-Korean for TydiQA. As shown in Table 4, our method improves zero-shot and rivals 3-shot supervised prompting across various low-resource languages.

4.5 General Instruction Following

Beyond traditional NLP tasks, we also show that our LDP prompts can elicit chatbot-style instruction following abilities in base pre-trained model **without** any supervised fine-tuning, and improve English-tuned models. Specifically, we utilize Sea-bench (Nguyen et al., 2023) - a set of categorized instructions in multiple languages, designed to evaluate models with LLM-as-a-judge recipe (Zheng et al., 2023). We measure GPT-4 rated scores of LLama2-13B base and chat models (Touvron et al., 2023b), using 4 random instructions from En, Zh, Fr, Ru as LDP prompts. As shown in Table 5, our method can invoke relatively good instruction following capability in Vi and Id even with a **base** model. With Llama2-chat, which has undergone supervised finetuning, our method can further improve the performance in various benchmarks for certain under-represented languages.

4.6 Ablation Study

In this section, we conduct various analyses in the unsupervised translation tasks to provide a deeper understanding of our LDP method and the importance of each component, while presenting more experiments in the Appendix.

Generating the Right Language. Figure 3a reveals one reason the models struggle to translate En$\rightarrow X$ when using LDP prompts $\mathcal{F}^{mt}$ (without intra-lingual BT data) is that the target-side distribution contains multiple languages, and the models struggle to recognize unfamiliar language tags, such as Marathi (Mr), and often generate wrong translations in the wrong languages (*e.g.,* Hindi instead of Marathi). Meanwhile, supplying synthetic intra-lingual prompts where the target-side is consistently in the intended language, as shown in Figure 3b with $\mathcal{F}^{mtbt}$, is more important in getting the models to recognize language rather than the language tag. In fact, we found that removing the language tag entirely can help improve the performance slightly.

Impact of Native Language Tag. The reason why we need unlabeled text to create intra-lingual prompts for En$\rightarrow X$ direction is because the models fail to recognize the correct language from the English language tags. A convenient way to eliminate such unlabeled text is to replace English-tag prompts (*e.g.,* "`Spanish:[es-text]\nChinese:[zh-text]`") with native language tags for the target language (*e.g.,* "`Española:[es-text]\n`中 文:[zh-text]"). Such native tags serve as examples of how the intended language looks like. As shown in Table 6a, using LDP with native language tags without using any unlabeled text or intra-lingual back-translation (BT) prompts improves the performance of En$\rightarrow X$ tasks significantly, compared to using English tags. This method even approaches the performance of 8-shot supervised prompting and LDP with unlabeled BT prompts. Combining it with back-translation data (Native-tag + BT) even

BLOOM	Ind-En	En-Ind
Unsupervised LDP		
En-tag	46.96	22.53
En-tag + BT	47.43	34.41
Native-tag	46.90	29.80
Native-tag + BT	47.52	35.22
No-tag	46.81	–
No-tag + BT	47.62	34.54

(a) Different language tags (chrF++).

BLOOM	Indic10-En	En-Indic10
Supervised	46.32	32.44
Unsupervised LDP with $\mathcal{Z}$ =		
Ar,Zh,Vi,Fr (default)	45.53	17.65
Hi,Hi,Hi,Hi (Hindi)	43.27	15.34
Ta,Bn,Hi (Indic)	45.51	16.25
Fr,Es,Pt (European)	45.31	18.98
Vi,Vi,Vi,Vi	44.91	12.94
Zh,Zh,Zh,Zh	44.71	15.78
Ar,Fr,Es,Pt,Vi,Zh,Id	45.50	16.88

(b) Choices of LDP languages (chrF++).

Table 6: (6a): Impact of English tag, native language tags and no language tag for in-context prompts in Indic languages. (6b): Impact of different choices of LDP languages on X→En directions using LDP without back-translation ($\mathcal{F}^{mt}$) across **10** Indic languages excluding Ta, Bn and Hi (Indic10). Note that we use supervised exemplars in Table 6b for analysis purpose.

helps it outperform supervised prompting. In fact, the English tag may confuse the model to an extent that not using the language tag at all (*e.g.,*using "`Input:[input]\nOutput:[output]`") does not hurt the performances.

Choice of LDP languages. Another necessary question to ask is which high-resource languages should be selected as LDP exemplars. Table 6b examines which LDP language choice is optimal. As shown, for 10 Indic low-resource languages, choosing a single related language (Hindi), which is often called cross-lingual prompting (Zhang et al., 2023; Zhu et al., 2023), can be disastrous as the model tends to translate the prompt language rather than the test language. Choosing a single but distant language, like Vi, yields better results, while choosing a wide variety of languages across different regions (*e.g.,* Ar,Zh,Vi,Fr) may be the optimal choice.

Comparison with Unsupervised MT. We also compare our method against the specialized unsupervised MT model CRISS (Tran et al., 2020) on eligible languages (Gu, Ne, Hi). As shown in Table 7, unsupervised LDP prompting with BLOOM significantly outperforms CRISS across all languages, thanks to its larger size and strong English abilities.

	Gu-En		Ne-En		Hi-En	
	→	←	→	←	→	←
CRISS	41.88	32.41	37.64	28.17	51.23	42.29
BLOOM Prompting						
Supervised	51.63	38.23	47.07	35.91	55.18	44.94
LDP	50.09	37.63	48.26	35.76	55.71	45.36

Table 7: Comparison in chrF++ between unsupervised LDP prompting and specialized unsupervised MT CRISS (Tran et al., 2020)

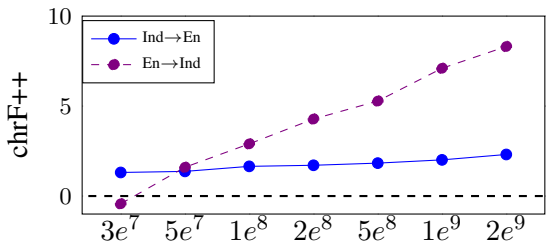


Figure 4: Gains achieved by fine-tuning BLOOM-7B w.r.t numbers of trainable parameters.

Fine-tuning Trainable Parameters. Figure 4 analyzes how LoRA-fine-tuned BLOOM-7B models (Hu et al., 2021) perform in X→En and En→X Indic translation tasks as we increase the trainable parameters from 30M to 2B (full query-key-value weights). As shown, gain margins for X→En are relatively low within 1 chrF++ as we fine-tune more parameters. Meanwhile, we observe a substantial gain of 8.7 chrF++ for En→X task, suggesting that learning to generate an unfamiliar language needs much more parameters, rendering parameter-efficient methods, like LoRA, ineffective.

5 Conclusion

We introduce linguistically-diverse prompting (LDP), which is designed to use synthetic high-quality in-context exemplars from high-resource languages to prompt LLMs to perform generative tasks in low-resource languages. Our unsupervised approach achieves on par with supervised few-shot prompting in English to and from 34 Indic and African translation tasks, even outperforming in non-English-centric directions. Our method also outperforms other English-pivoting techniques in zero-shot multilingual summarization, question answering and instruction following.

6 Limitations and Ethical Impact

Our linguistically-diverse prompting method comes with a few limitations that should be considered when used. First, it is a way to invoke and improve LLM's abilities in low-resource languages, and not necessarily boosting low-resource knowledge beyond the data the model was trained. Second, the presence of texts in the target low-resource languages are often needed in the context for the method to work effectively, thus it does not entirely eliminate the need for unlabeled data in such languages at inference times. Third, like many methods with LLMs, hallucinations may occur with our LDP prompting method.

Regarding ethical impact, we do not foresee any potential ethical issues with our proposed method.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

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*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online. Association for Computational Linguistics.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur Parikh. 2020. A multilingual view of unsupervised machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3160–3170, Online. Association for Computational Linguistics.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Online. Association for Computational Linguistics.

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.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv 2303.16634*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Xuan-Phi Nguyen, Hongyu Gong, Yun Tang, Changhan Wang, Philipp Koehn, and Shafiq Joty. 2022a. Contrastive clustering to mine pseudo parallel data for unsupervised translation. In *International Conference on Learning Representations*.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and AiTi Aw. 2022b. Refining low-resource unsupervised translation by language disentanglement of multilingual translation model. In *Advances in Neural Information Processing Systems*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint*.

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.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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.

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. *arXiv preprint arXiv:2305.03573*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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. *arXiv preprint arXiv:2302.13971*.

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

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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 arXiv:2305.11206*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Example Appendix

A.1 Low-resource Language Details

Table 8 lists the details of each low-resource language in the ROOTS corpus (Laurençon et al., 2022) that we mainly evaluate with the BLOOM model (Scao et al., 2022). Regarding test sets, we primarily choose from the ML50 benchmark (Tang

et al., 2020), which collected test data from various sources, such as WMT (Barrault et al., 2020) and FLoRes (Guzmán et al., 2019; Goyal et al., 2022). For languages absent in ML50, we choose the NLLB-devtest sets (Costa-jussà et al., 2022) as replacement. For non-English $X \rightarrow Y$ tasks, we choose NLLB-devtest for all our evaluation. To limit the API call costs within our budget, we randomly the same 200 samples from each test set for evaluation.

A.2 Experiment Details

Few-shot data sources. For supervised prompting, we collect randomly parallel pairs from the respective valid set for each language. For unlabeled data for our LDP method, we collect and filter data from various sources, as specified in *Unlabeled* column of Table 8. Specifically, the primary unlabeled source is the CC100 corpus (Wenzek et al., 2020; Conneau et al., 2020). For those absent in CC100, we collect data from other sources, such as the ROOTS corpus (Laurençon et al., 2022), MMTAfrica (Emezue and Dossou, 2021) and MAFAND (Adelani et al., 2022). For the remaining languages where we could not find in research repositories, we crawled from several religious and news websites (OUR). The sizes of collected unlabeled texts vary greatly, ranging from a few millions lines for Hindi to less than 1000 lines for Bambara, thus presenting a challenge for data balancing. For LDP non-English high-resource exemplars, we randomly collect a single high-quality sentence of similar lengths from the CC100 corpus for each language and use the unsupervised CRISS model (Tran et al., 2020) to translate them into English.

Unlabeled data filtering To ensure high-quality native texts for unsupervised LDP prompting as well as larger-scale synthetic data creation for fine-tuning, we filter unlabeled texts such that they (*i*) are within 20 to 200 character lengths, (*ii*) do not contain non-conversational artifacts like URLs, brackets, bullet points or excessive numbers, and (*iii*) do not contain more than 20% alphabetical characters for Indic and non-latin characters for African languages. For fine-tuning, we use an upscaling temperature of 25 to smoothen the data mixture imbalance.

Other Details. We evaluate translation tasks with chrF++ (Popović, 2015) and SacreBLEU (Post, 2018). For SacreBLEU, we use the default tokenizer for Latin-based languages, while follow Guzmán et al. (2019); Goyal et al. (2022) to use indic_nlp_library for Indic language tokenization.

For each of the 68 language pairs, we sample randomly and evaluate the same 200 sentences from each test set with the same zero seed to limit the cost of API calls[2]. We conduct full-set evaluations for 4 random languages in each group and observe < 1 chrF++ standard deviation from our 200-sample evaluations.

Low-resource $\leftrightarrow$ En translation. Table 10 presents the full results with chrF++ and SacreBLEU (BLEU) scores for unsupervised translation tasks between English and 13 Indic (Ind) and 21 African (Afr) low-resource languages present in the ROOTS corpus.

A.3 Additional Experiments

Breakdown of $X \rightarrow$En. Similar to the observation for En$\rightarrow X$ in the main paper, Figure 6 shows that LDP performs generally on par with supervised prompting equally across all languages, and that it does not unevenly perform much worse or better in any particular language.

High-resource Translation with Llama LLaMA (Touvron et al., 2023a) is another open-sourced LLM that only supports 20 European high-resource languages. We evaluate LLaMA in translation tasks between English and the remaining 19 languages, which include Hungarian, Danish and Catalan. Specifically, we use CRISS to generate synthetic LDP exemplars from De, Es and Fr, which we then use to prompt LLaMA to translate from and to such languages. As reported in Table 9, we observe similar trends where our LDP method performs competitively with supervised prompting. The overall scores for such languages are also much higher than those of non-Latin languages because LLaMA was also pre-trained with bitexts, though without explicit alignments.

BLOOM vs. InstructGPT. While much evidence show that InstructGPT text-davinci-003 is more superior than the vanilla BLOOM (Scao et al., 2022; Ouyang et al., 2022) in many languages, our experiments with low-resource languages demonstrate it is not always true for low-resource non-Latin languages, as shown in the main paper. Fig-

[2] bigscience/bloom, openai.com.

Indic				African			
Name	**Code**	**Test**	**Unlabeled**	**Name**	**Code**	**Test set**	**Unlabeled**
Assamese	as	NLLB	CC100	Tumbuka	–/tum	NLLB	OUR
Oriya	or	NLLB	ROOTS	Kikuyu	ki/kik	NLLB	OUR
Gujarati	gu	ML50	CC100	Bambara	bm/bam	NLLB	MAFAND
Marathi	mr	ML50	CC100	Akan	ak/aka	NLLB	OUR
Panjabi	pa	NLLB	CC100	Tsonga	ts/tso	NLLB	MMTAfrica
Kannada	kn	NLLB	CC100	Southern Sotho	st/sot	NLLB	OUR
Nepali	ne	ML50	CC100	Chewa	ny/nya	NLLB	MMTAfrica
Telugu	te	ML50	CC100	Tswana	tn/tsn	NLLB	MMTAfrica
Malayalam	ml	ML50	CC100	Lingala	ln/lin	NLLB	MMTAfrica
Urdu	ur	NLLB	CC100	Northern Sotho	–/nso	NLLB	MMTAfrica
Tamil	ta	ML50	CC100	Fon	–/fon	NLLB	MAFAND
Bengali	bn	NLLB	CC100	Rundi	rn/run	NLLB	OUR
Hindi	hi	ML50	CC100	Wolof	wo/wol	NLLB	CC100
			CC100	Luganda	lg/lug	NLLB	CC100
			CC100	Shona	sn/sna	NLLB	CC100
			CC100	Zulu	zu/zul	NLLB	CC100
			CC100	Igbo	ig/ibo	NLLB	CC100
			CC100	Xhosa	xh/xho	NLLB	CC100
			CC100	Kinyarwanda	rw/kin	NLLB	MMTAfrica
			CC100	Yoruba	yo/yor	NLLB	CC100
			CC100	Swahili	sw/swa	NLLB	CC100

Table 8: Low-resource language details and corresponding test sets and unlabeled data sources for $X \leftrightarrow$En translation tasks.

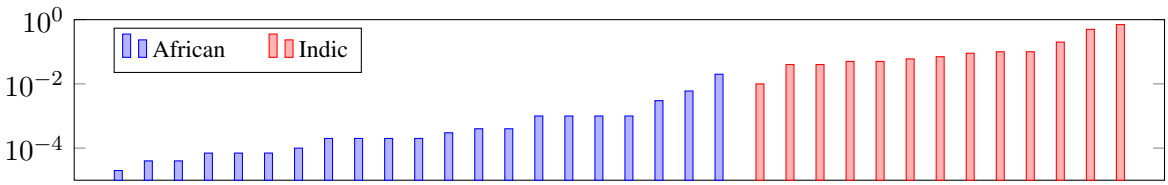


Figure 5: Low-resource language coverage % of the ROOTS corpus (Laurençon et al., 2022) used to train BLOOM. The highest-resource language for Indic and African are Hindi and Swahili. Hindi accounts for 0.7% and the rarest language, Tumbuka, takes up only $2e^{-5}\%$ of the corpus.

LLaMA-30B	**X→En**		**En→X**	
	chrF++	**BLEU**	**chrF++**	**BLEU**
Supervised	61.80	39.51	53.65	28.98
Unsupervised-LDP	61.75	38.83	54.00	29.58

Table 9: Comparison between supervised and unsupervised-LDP prompting with LLaMA-30B model in translation tasks between English (En) and 19 European languages (X). LDP prompts consist of exemplars from high-resource languages seen by CRISS.

ure 8 explains clearly the reason is that GPT's tokenizer is not designed to allocate meaningful sub-word tokens for non-Latin texts, such as Indic lexical items, while significantly favors Latin characters due to the sheer size of Latin texts in its pre-training data. For example of InstructGPT, a 10-token English text can be equivalent to a 160-token Tamil text but only a 28-token Tumbuka text, despite Tumbuka is much more low-resource. This issue is non-existent in BLOOM, as the ratios naturally decrease when data coverages increase. As shown in the table, InstructGPT becomes worse than BLOOM as soon as the ratio between token lengths of target language over English surpass 5 in Indic languages. We refer to this as sub-word token fragmentation, where texts are broken into very long byte-level tokens that exceed the context length and suppress performances.

	Ind-En		En-Ind		Afr-En		En-Afr	
	chrF++	**BLEU**	**chrF++**	**BLEU**	**chrF++**	**BLEU**	**chrF++**	**BLEU**
Foundation BLOOM-175B								
Supervised-8-shot	47.31	22.32	34.66	9.02	28.64	8.35	14.93	2.00
Unsupervised-LDP	47.62	22.38	34.54	8.88	28.72	8.71	14.57	1.89
Foundation BLOOM-7B								
Supervised-8-shot	39.86	14.77	24.02	4.42	21.51	4.33	11.27	0.59
Unsupervised-LDP	39.88	14.96	24.41	4.52	20.47	3.65	12.04	0.62
Fine-tune QKV (2B params)	42.19	17.13	32.72	8.33	21.14	5.15	15.73	2.13
Supervised RLHF InstructGPT (text-davinci-003)								
Zero-shot with instruction	35.37	11.48	20.71	3.88	27.10	8.04	15.45	1.13
Supervised-6-shot	37.07	13.13	24.74	5.21	31.51	10.88	19.22	2.66
Unsupervised-LDP	38.45	14.22	25.17	5.06	31.92	11.12	19.51	2.61
Supervised upperbound								
NLLB-200 distilled	*61.00*	*37.24*	*46.77*	*18.78*	*48.42*	*26.92*	*39.18*	*12.95*

Table 10: Averaged performances of different prompting techniques across various model sizes and types, namely BLOOM (Scao et al., 2022) and InstructGPT text-davinci-003 (Brown et al., 2020; Ouyang et al., 2022), in translation tasks between English (En) and 13 Indic (Ind) and 21 African (Afr) low-resource languages present in the ROOTS corpus (Laurençon et al., 2022).

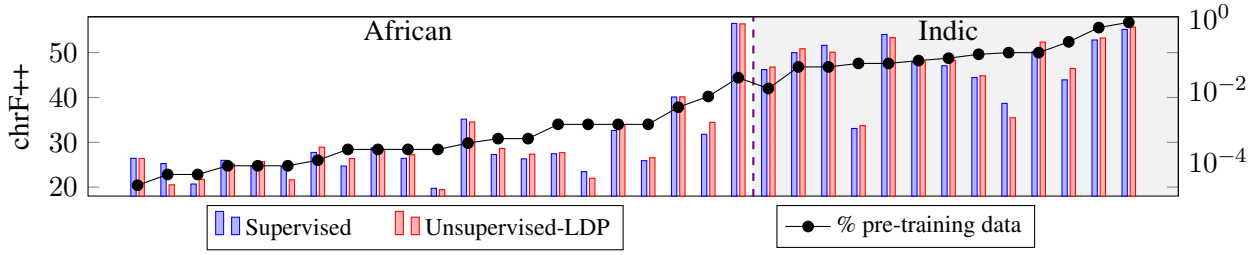


Figure 6: chrF++ scores for translation from each Indic and African language in the ROOTS corpus to English (X→En), using BLOOM. The right y-axis indicates corresponding pre-training coverage of each language at log scale.

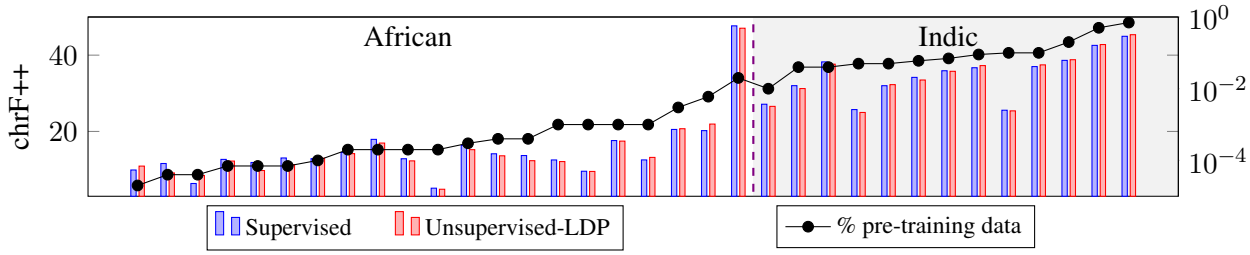


Figure 7: chrF++ scores for translation from English to each Indic and African language in the ROOTS corpus (En→X), using BLOOM. The right y-axis indicates corresponding pre-training coverage of each language at log scale.

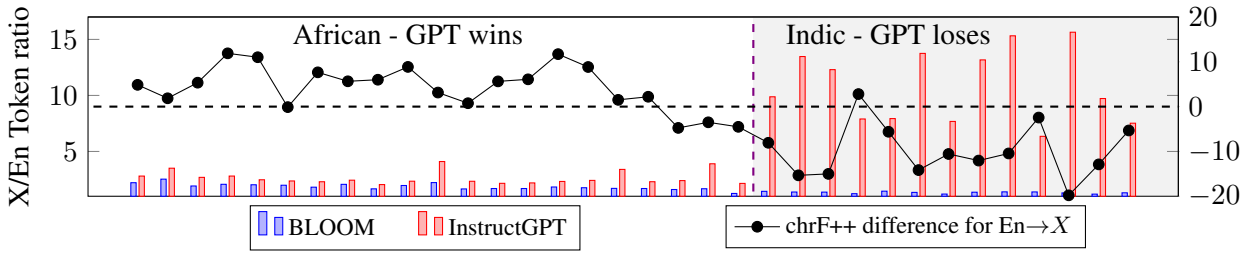


Figure 8: Tokenization issue. Left y-axis bar chart: The average ratios between the token lengths of X-language text over their English counterparts of the same meaning. Right y-axis line chart: chrF++ performance difference between GPT text-davinci-003 and BLOOM for En→X tasks, meaning < 0 indicates GPT is worse than BLOOM.