sPHINX: Sample Efficient Multilingual Instruction Fine-Tuning Through N-shot Guided Prompting

Anonymous ACL submission

Abstract

Despite the remarkable success of LLMs in English, there is a significant gap in performance in non-English languages. In order to address this, we introduce a novel recipe for creating a multilingual synthetic instruction tuning dataset, sPHINX, which is created by selectively translating instruction response pairs from English into 50 languages. We test the effectiveness of sPHINX by using it to fine-tune two state-of-the-art models, MISTRAL-7B and PHI-3-SMALL and then evaluating them across a comprehensive suite of multilingual benchmarks that test reasoning, question answering, reading comprehension and machine translation. Our results show that MISTRAL-7B and PHI-3-SMALL fine-tuned with sPHINX perform better on an average by **5%pt** for both the models when compared to the base variants of these models. We also devise a strategy to incorporate N-shot examples in each fine-tuning sample which further boosts the performance of these models to **9%pt** and **4%pt** respectively compared to vanilla fine-tuning. To show efficacy of our data curation approach, we also directly translate our original dataset to the target languages, and observe an increase of **7%pt** and **4%pt** on both the models respectively. sPHINX outperforms other multilingual instruction tuning datasets in both efficiency and diversity, reducing dataset creation costs. It also maintains strong performance on standard English LLM benchmarks, with minimal regression.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional performance across various tasks in English. However, their performance in some non-English languages remains comparatively limited (Ahuja et al., 2023; Asai et al., 2023). Further, the gap between the performance of Large Language Models (LLMs) and Small Language Models (SLMs) is more pronounced Ahuja et al. (2024) on non-English languages. Cui et al. (2023) and Balachandran (2023) utilize the method of fine-tuning models on datasets focused on particular languages. However, this can lead to catastrophic forgetting, which may negatively impact performance on English(Zhao et al., 2024). Several techniques have been proposed to bridge this gap, such as incorporating better pre-training data in various languages and enhancing base tokenizers (Xu et al., 2024; Dagan et al., 2024). However, most of these changes need to be implemented at the pre-training stage which demands extensive data and computational resources, making it practically unfeasible in many scenarios (Brown et al., 2020). Consequently, the most well-studied technique involves fine-tuning models for specific languages and tasks. Instruction fine-tuning (IFT) has become a popular technique for enhancing the performance of language models in specific languages. This method combines the benefits of both the pre-training, fine-tuning and prompting paradigms (Wei et al., 2021).

Sample diversity is essential for effective instruction tuning in multilingual datasets. Many recent datasets have been generated by translating English content into other languages or by employing self-instruct techniques based on seed prompts (Li et al., 2023; Taori et al., 2023). However, both methods can restrict diversity. Machine translation may result in the loss of semantic nuance (Baroni and Bernardini, 2006), while self-instruct approaches often yield repetitive and homogeneous samples (Wang et al., 2022). This highlights the critical need for datasets that encompass a wide range of diverse samples.

In this paper, we present a novel recipe for creating a multilingual synthetic instruction tuning dataset, sPHINX. It comprises 1.8M instruction-response pairs in 51 languages, derived by *selectively translating* the Orca instruction tuning

dataset (Mukherjee et al., 2023) using GPT-4 (Achiam et al., 2023). We assess the effectiveness of sPHINX by fine-tuning two models — PHI-3-SMALL and MISTRAL-7B — across a range of evaluation benchmarks that test various language model capabilities like reasoning, question answering, and reading comprehension. We compare models fine-tuned on sPHINX with those using other synthetic multilingual instruction tuning datasets like AYA (Üstün et al., 2024), MULTILINGUAL ALPACA (Taori et al., 2023), and BACTRIAN (Li et al., 2023) and observe significant performance gains. We also compare our proposed translation strategy with translating the entire instruction using Translator API, as is done with the popular multilingual synthetic IFT datasets to demonstrate the efficacy of our approach.

The contributions of this paper are as follows:

- We introduce a novel approach to generate synthetic data for multilingual instruction tuning by ***selectively translating*** the Orca dataset with the assistance of GPT-4 (§3.1)
- We conduct an in-depth analysis where we find sPHINX to be more sample efficient and diverse across languages (§3.3)
- We devise ***LA**nguage-Specific **N**-shot **G**uided Instruction fine-tuning (LANG)* strategy for enhancing the multilingual capabilities of LLMs (§4.2)
- We also perform an extensive number of experiments on different multilingual instruction tuning datasets to test generalizability in multilingual settings (§5)

2 Related Work

2.1 Multilingual LLMs

Recently, there has been there has been interest in developing and improving SLMs such as LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), Phi3 (Abdin et al., 2024), and Gemma (Team et al., 2024). In a relatively brief period, researchers have developed numerous monolingual and multilingual versions of these foundational models, either by pre-training from scratch on specific languages or groups of languages, or by further fine-tuning on them (Nguyen et al., 2023; Gala et al., 2024; Balachandran, 2023; Cui et al., 2023; Zhang et al., 2023; Qin et al., 2024). Prior work has shown that smaller models show promising results in high-resource languages but perform worse on low-resource ones (Ahuja et al., 2024) compared to larger models. To address this, researchers have proposed using multilingual corpora for pre-training, expanding the vocabulary (Wang et al., 2019) and continual training (Liu et al., 2023), though this increases training costs significantly. fine-tuning for specific tasks, such as translation, has been explored but often struggles with generalization (Mishra et al., 2021). In contrast, instruction tuning has shown to improve generalization to unseen tasks, enhancing the models' ability to understand and respond to human instructions (Mishra et al., 2021; Ouyang et al., 2022). This makes the models more versatile and capable of handling a wide variety of tasks.

2.2 Multilingual Instruction fine-tuning

Early studies focused on fine-tuning pre-trained models on a variety of languages through data augmentation for a single task (Hu et al., 2020; Longpre et al., 2021; Asai et al., 2022). Currently, the approach has shifted on fine-tuning these models using a wide-variety of tasks (Longpre et al., 2023; Ouyang et al., 2022). Models such as BLOOMZ (Muennighoff et al., 2022) and mT0 (Muennighoff et al., 2022) makes significant strides in improving the multilingual performance of decoder-based models (Ahuja et al., 2023). There have been multiple multilingual instruction datasets and models proposed such as Bactrian (Li et al., 2023), AYA (Üstün et al., 2024), POLYLM (Wei et al., 2023b) after BLOOMZ and mT0. However, these models still do not perform as well as English on other languages, with the gap being particularly large for low-resource languages and languages written in scripts other than the Latin script (Ruder et al., 2021; Ahuja et al., 2023; Asai et al., 2023; Ahuja et al., 2024). In this work, we aim to narrow the performance gap by introducing a strategy for creating datasets for multilingual instruction tuning and recipes for fine-tuning, which we will discuss in the following sections.

2.3 Multilingual Synthetic Data Generation

Most instruction-tuning datasets across multiple languages typically focus on general tasks rather than specific reasoning capabilities. While datasets like Orca (Mukherjee et al., 2023) and Orca 2 (Mitra et al., 2023) exist in English, they highlight a prevalent issue: current methods often prioritize

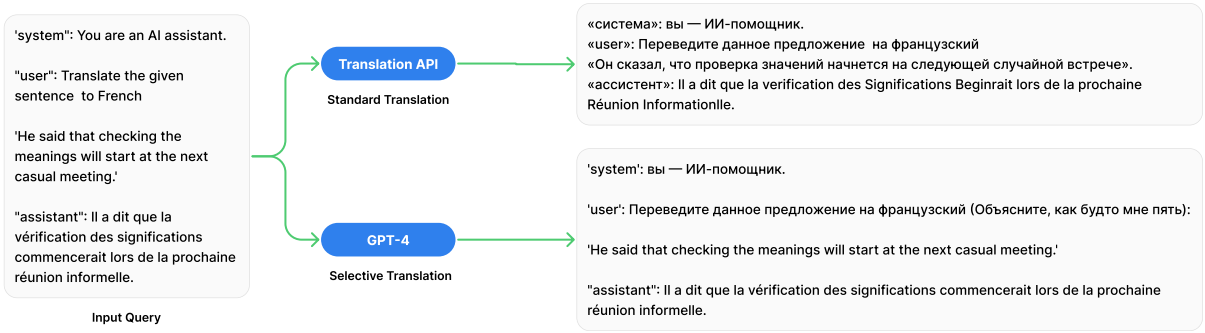

Figure 1: The above figure describes translating using a Translation API vs Selective Translation

style imitation over leveraging the reasoning abilities found in large foundation models (LFMs). The Orca dataset addresses this by imitating rich signals from GPT-4, including explanation traces and step-by-step thought processes (Wei et al., 2023a), guided by assistance from ChatGPT. In order to create multilingual datasets, researchers commonly use translation APIs or LLMs to translate English-specific datasets into target languages. For example, the Bactrian dataset (Li et al., 2023) translates Alpaca and Dolly instructions into 52 languages using the Google Translator API and generates outputs with GPT-3.5 turbo. However, these translated datasets often struggle to encode semantic information effectively (Baroni and Bernardini, 2006). Our dataset approach aims to tackle these challenges by selectively translating only the essential portions of multilingual inputs. This strategy not only preserves semantic information but also accommodates diverse linguistic contexts, thereby enhancing the overall quality and applicability of instruction-tuning datasets across languages.

3 SPHINX Dataset

In this section, we describe our dataset construction methodology (§3.1), dataset filtering and cleaning pipelines (§3.2), and exploratory analysis (§3.3) for determining diversity of our data in comparison to other instruction tuning datasets.

3.1 Dataset Construction

Mukherjee et al. (2023) illustrate how smaller models can replicate the reasoning processes of LLMs and learn from detailed signals like as explanation traces, step-by-step thought processes, and other intricate instructions obtained from data annotated by ChatGPT and GPT-4[1]. Their dataset consists of 5M ChatGPT and 1M GPT-4 generated instruction-response pairs. Inspired by this work, we utilized 1M GPT-4 generated instruction-response pairs of their dataset and construct our dataset along similar lines by *selectively* translating them into 50 different languages with the help of GPT-4. We categorize them into three groups: high-resource, mid-resource, and low-resource languages as outlined in Table 13. For high resource languages, we randomly sample 100k instruction-response pairs from the Orca 1M dataset and generate the responses from GPT-4 using **Selective Translation** as shown in Table 1. Similarly, we leverage the same strategy for medium and low resource languages by sampling 50k and 25k pairs respectively. Although GPT-4 performs competitively with commercial translation systems (Google Translate & Bing Translate) it still lags behind on medium and low resource languages (Jiao et al., 2023; Hendy et al., 2023). Furthermore, as highlighted in Chang et al. (2023), fine-tuning with a large set of samples from medium and low-resource languages might lead to catastrophic forgetting of high-resource languages. Therefore, we deliberately create fewer samples for medium and low-resource languages than for high-resource ones.

A fundamental problem with using an off-the-shelf translation API is the lack of semantic and task awareness, in addition to *translationese* (Baroni and Bernardini, 2006)), which can result in poor quality training data. Consider for example the task of Machine Translation as part of the instruction, wherein the language of the source sentence should be retained. However, an off-the-shelf API, without task awareness, would translate it, resulting in an ill defined instruction. To mitigate this issue, we prompt GPT-4 to selectively translate the instructions, so that the tasks are translated

[1]GPT-4 inference hyper-parameters in Azure OpenAI interface set as: temperature=0.0, top_p=1, frequency_penalty=0, presence_penalty=0, stop=["<|*im_end*|>", "<|*im_start*|>"]

INPUT QUERY	MULTIALPACA DATASET	SELECTIVE TRANSLATION

{'instruction': 'Find the French equivalent of the following phrase.',
'input': 'Wishing you good luck',
'output': 'Je vous souhaite bonne chance'}
	{'instruction': 'निम्नलिखित वाक्यांश के फ्रेंच समकक्ष का पता लगाएं।',
	'input': 'आपको शुभकामनाएं',
	'output': 'Vous avez mes meilleurs vœux.'}
		{
		"system": "निम्नलिखित वाक्यांश का फ्रेंच समकक्ष खोजें।",
		"human": "Wishing you a good luck",
		"assistant": "Je vous souhaite bonne chance"
		}

{'instruction': 'You are provided with a sentence, edit it in a way that it becomes grammatically correct.',
'input': 'The wind is blowing fast and furious',
'output': 'The wind is blowing fast and furiously.'}
	{'instruction': 'आपको एक वाक्य प्रदान किया जाता है, इसे इस तरह संपादित करें कि यह व्याकरणिक रूप से सही हो जाए।',
	'input': 'हवा तेज और उग्र चल रही है',
	'id': 'alpaca-9380',
	'output': 'तेज और उग्र हवा चल रही है।'}
		{
		"system": "आपको एक वाक्य दिया गया है, इसे इस प्रकार संपादित करें कि यह व्याकरणिक रूप से सही हो जाए।",
		"human": "The wind is blowing fast and furious",
		"assistant": "The wind is blowing fast and furiously."
		}

Figure 2: Some examples of input queries and its counterpart existing in the hindi version of the MULTIALPACA dataset and if it was generated using the Selective Translation strategy

into the appropriate language without changing the semantic meaning. Figure 2 illustrates this with concrete examples. The first example demonstrates the aforementioned translation inconsistency issue for an instruction asking for a French equivalent of an English phrase. The second example demonstrates the consequence of direct translations in the MULTIALPACA dataset: wherein the translation of the task input results in the task being ill-defined based on the instructions. As demonstrated, our proposed **Selective Translation** method is able to keep the semantic information of the task intact while translating the instructions. For the exact prompt, please refer to Figure 5 in the Appendix.

3.2 Dataset Filtering and Cleaning

After generating these translations, we analyze how many instruction-assistant pairs were unsatisfactorily translated, such as incomplete or missing translations, by identifying the English content within the instances. We use the NLTK[2] library to identify the fraction of English words in the sample. The NLTK corpus contains approximately 240,000 English words, making it suitable for our use case. We manually analyzed various proportions of English content in samples of each language and found that the samples were satisfactory when the English content was within 90% for all languages. If the English content exceeded 90% in a language that uses a different script, those samples were removed from our dataset as shown in our data filtering algorithm here. Finally, we manually inspected a small sample of the dataset to ensure the quality of the translations, and found them to be of good quality.

After filtering, the final dataset comprised of 1.8 million samples in 51 languages, divided into three subsets: Train, Test, and Few-shot. We partitioned each language's dataset to ensure the Test and Few-shot sets contained 2000 and 1000 samples respectively, with the Train set consisting the remaining data. This approach guarantees consistent distributions across languages in the Test and Few-shot sets, ensuring equitable representation regardless of training distribution. The train, test, and few-shot sets are in the ratio of 92 : 5.3 : 2.7.

3.3 Data Analysis

A significant issue we hypothesize—supported by preliminary analysis of prior Multilingual Instruction Following Task (IFT) datasets—is the lack of diversity in the data samples. To further investigate this, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) (Cai and Ma, 2022) to visualize the structure of both prior multilingual IFT datasets and SPHINX. We used the `paraphrase-multilingual-MiniLM-L12-v2` model from HuggingFace's sentence-transformers library (Reimers and Gurevych, 2020, 2019)[3] to generate the embeddings. The t-SNE hyperparameters were set as follows: learning rate of 200, perplexity of 30, and number of iterations (n_iter) at 1000. We visualized 1000 samples evenly distributed across 10 languages.

As demonstrated in Figure 3, SPHINX shows significantly greater diversity compared to prior

[2] https://www.nltk.org/

[3] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

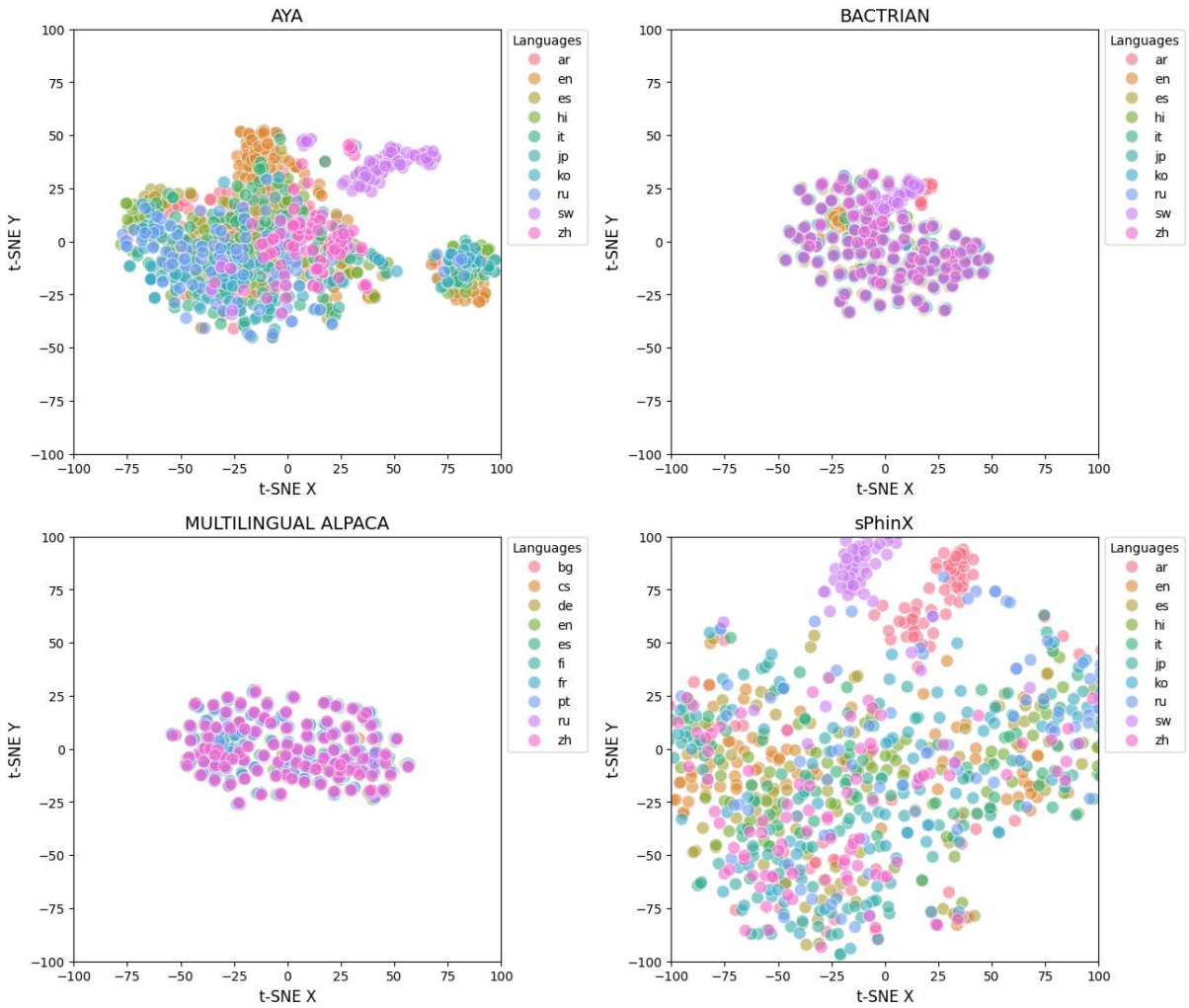

Figure 3: t-SNE Visualisation of 1000 samples equally distributed in 10 languages

datasets. Notably, both BACTRIAN and MULTILINGUAL ALPACA lack diversity, likely due to their generation method, where samples are largely derived from translating a fixed set of prompts from English into various languages. This approach limits the variability across languages, as the same core set of tasks is simply mirrored across different languages.

In contrast, sPHINX was designed to increase diversity by performing Selective Translation. Specifically, we applied selective translation on a subset of 1 million original data samples, translating approximately 10% for high-resource languages, 5% for mid-resource languages, and 2.5% for low-resource languages. This method ensures that the translated data is not uniform across languages, leading to a more diverse set of samples across different languages. By incorporating a more diverse seed set (Mukherjee et al., 2023) and using this Selective Translation strategy, we are able to address the lack of diversity present in prior datasets.

Additionally, sPHINX includes a substantial amount of code-switched data due to the nature of Selective Translation, which further enhances the linguistic diversity. While the AYA dataset shows more diversity than other prior datasets, it still demonstrates limited variation across task instructions.

4 Experiments

4.1 Setup

Base Models: We use MISTRAL-7B [4] (Jiang et al., 2023) and PHI-3-SMALL (Abdin et al., 2024) base model variants and instruction fine-tune them.

Datasets: Apart from the sPHINX dataset, we use BACTRIAN (Li et al., 2023), MULTIALPACA (Wei et al., 2023b) and AYA (Singh et al., 2024) instruc-

[4]We specifically use the v1.0 base model from here.

tion datasets for comparative evaluation. We also utilize the Azure Translator API[5] (SPHINX-T) to translate the original dataset into all our target languages, demonstrating the effectiveness of our **Selective Translation** approach.§3.1.

- BACTRIAN (Li et al., 2023) is a machine translated dataset of the original alpaca-52k (Taori et al., 2023) and dolly-15k (Conover et al., 2023) datasets into 52 languages. The instructions for this dataset were translated using a Translation API and then GPT-3.5-Turbo was prompted to generate outputs. We fine-tune our models on the complete dataset consisting of 3.4M instances.
- MULTIALPACA (Wei et al., 2023b) is a self-instruct dataset that translates seed instructions from English to 11 languages, using GPT-3.5-Turbo for response generation. We fine-tune our models on the full dataset, which contains 500k datapoints.
- AYA (Singh et al., 2024) contains human-curated prompt-completion pairs in 65 languages, along with 44 monolingual and multilingual instruction datasets and 19 translated datasets across 114 languages, totaling around 513M instances. To ensure parity with the SPHINX dataset, we sampled it down to 2.7M instances, ensuring equal representation for each language in our subset.

Evaluation: We evaluate[6] our fine-tuned models along with the available base and Instruction fine-tuned model variants of MISTRAL-7B and PHI-3-SMALL (IFT[7]) on 4 discriminative tasks; XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2022), XWinograd (Muennighoff et al., 2023), (Tikhonov and Ryabinin, 2021), Belebele (Bandarkar et al., 2023), and 2 generative tasks; XQuAD (Artetxe et al., 2020) and Translation (Bojar et al., 2014, 2016; Tom et al., 2023) using the language model evaluation harness (Gao et al., 2023).

- **XCOPA**: A causal commonsense reasoning dataset in 11 languages, evaluated in a 4-shot prompt setting.
- **XStoryCloze**: A professionally translated version of the English StoryCloze dataset (Mostafazadeh et al., 2017) in 10 languages, evaluated in a 4-shot prompt setting.
- **Belebele**: A parallel reading comprehension dataset across 122 languages, with evaluation on a subset of 14 languages in a 0-shot prompt setting.
- **XQuAD**: A QA dataset consisting of professional translations of a subset of SQuAD into 10 languages, evaluated in a 3-shot prompt setting due to context window limitations.
- **XWinograd**: A collection of Winograd Schemas in six languages for cross-lingual commonsense reasoning, evaluated in a 0-shot setting.
- **Translation**: We utilize a subset of WMT14, WMT16 and WMT23 of language pairs (7 languages), with evaluation in a 4-shot setting.

4.2 Fine-Tuning Methodology

Inspired from Longpre et al. (2023)'s strategies to instruction tune a model, we devise *Language-Specific N-shot Guided Instruction fine-tuning (LANG)*. This method aims to improve the model's ability to follow instructions by augmenting training examples with additional context from a set of few-shot examples in the same language.

For each training example, we begin by sampling a number of few-shot examples, which are instruction-response pairs in the same language. The number of few-shot examples N is determined probabilistically, with a 30% chance of selecting no few-shot examples, a 20% chance of selecting one, and gradually lower probabilities for higher numbers of few-shots. The maximum number of few-shot examples we sample is six, due to constraints imposed by the model's context length (8192 tokens) and the typically higher tokenization length in languages other than English.

Once the number of few-shot examples is determined, they are prepended to the main training example, forming an augmented input. This augmented input is then fed into the model for instruction tuning. The purpose of this approach is to expose the model to additional examples of different tasks, helping it generalize better to new tasks in the same language.

We performed experiments to analyze how the model performs on each dataset when fine-tuned

[5] https://azure.microsoft.com/en-us/products/ai-services/ai-translator

[6] Evaluation prompts and other details in Appendix §A.1

[7] We take the MISTRAL-7B instruction-tuned variant from here and PHI-3-SMALL variant from here.

using the *LANG* strategy detailed in the next section (§5). Additionally, we fine-tuned the models on the SPHINX dataset without using *LANG* to provide a baseline for comparison. To assess the effectiveness of each instruction tuning dataset on an equal scale, we conducted a comparative analysis of model performance on different benchmarks, fine-tuning each model on approximately 8 billion tokens per dataset using the *LANG* strategy.

This fine-tuning strategy is consistently applied across datasets for both the PHI-3-SMALL and MISTRAL-7B base models. A comparison of token lengths across different datasets is provided in Table 4, showing the average token lengths as tokenized by the PHI-3-SMALL model.

5 Results

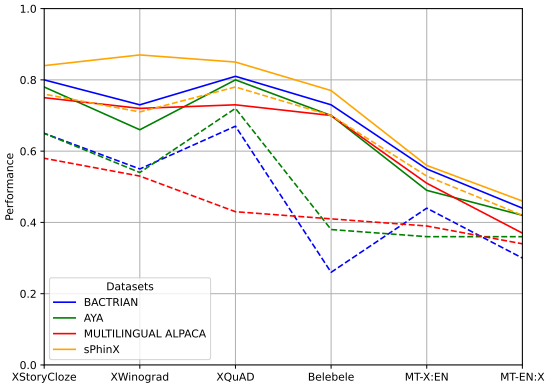

Figure 4: Performance of MISTRAL-7B and PHI-3-SMALL when instruction-tuned on 8B tokens across various datasets on different benchmarks.

We evaluate reasoning, question answering, translation and reading comprehension abilities of the PHI-3-SMALL and MISTRAL-7B models, instruction-tuned on different multilingual datasets, using various benchmarks and find that fine-tuning on SPHINX provides an improvement of **5%pt**[8] on both the models. (Refer to SPHINX-0S in Table 1 for overall results and to Appendix §A.4 for language-wise results). Additionally as observed in Figure 4, the SPHINX dataset significantly enhances the multilingual performance of the PHI-3-SMALL and MISTRAL-7B model compared to other datasets even when fine-tuned on equal number of tokens. Furthermore, during instruction tuning on 8B tokens, the models encountered fewer training samples for SPHINX due to its higher average token length per sample, as illustrated in Table 4.

[8]A percentage point (pt) is the unit for expressing the absolute difference between two percentage values

Model	XC	XS	XW	XQ	BL	MT^1	MT^2
MISTRAL-7B							
Base Model	0.63	0.68	0.52	0.74	0.24	0.54	0.42
IFT	0.62	0.73	0.54	0.60	0.47	0.49	0.39
M-ALPACA	0.55	0.59	0.51	0.46	0.41	0.41	0.39
AYA	0.68	0.71	0.54	0.66	0.38	0.39	0.37
BACTRIAN	0.54	0.67	0.54	0.69	0.26	0.45	0.34
SPHINX-T	0.61	0.78	0.57	0.78	0.67	0.49	0.38
SPHINX-0S	0.58	0.58	0.68	0.69	0.67	0.49	0.42
SPHINX	**0.68**	**0.81**	**0.71**	**0.80**	**0.71**	**0.55**	**0.46**
PHI-3-SMALL							
Base Model	0.64	0.78	0.75	0.78	0.65	0.54	0.42
IFT	0.68	0.79	0.78	0.75	0.70	0.54	0.46
M-ALPACA	0.68	0.79	0.81	0.77	0.75	0.45	0.39
AYA	0.65	0.79	0.69	0.83	0.72	0.41	0.40
BACTRIAN	0.71	0.82	0.73	0.85	0.77	0.54	0.40
SPHINX-T	0.70	0.80	0.77	0.78	0.75	0.55	0.44
SPHINX-0S	0.71	0.81	0.80	0.82	**0.79**	**0.56**	0.45
SPHINX	**0.72**	**0.84**	**0.87**	**0.87**	**0.79**	**0.56**	**0.46**

Table 1: Performance of MISTRAL-7B and PHI-3-SMALL instruction-tuned on various datasets. Abbreviations: XC - XCOPA (Acc.,4-shot), XS - XStoryCloze (Acc.,4-shot) , XW - XWinograd (Acc., 0-shot), XQ - XQuAD (F1,3-shot), BL - Belebele (Acc., 0-shot). MT^1- Translation for x:en (ChrF, 4-shot), MT^2 - Translation for en:x direction (ChrF, 4-shot). The best performing dataset for each model is indicated in bold, and the overall best performing model is indicated with an underline.

6 Ablations

6.1 Improvements from *LANG*

To demonstrate the effectiveness of our *LANG* strategy, we also instruction-tuned the models on SPHINX without any pre-pended shots, referring to this as SPHINX-0S. As shown in Table 1 (with detailed results in Appendix §A.4), models fine-tuned on SPHINX especially MISTRAL-7B exhibits superior performance compared to its counterparts fine-tuned on other datasets across all benchmarks. Moreover, fine-tuning both MISTRAL-7B and the PHI-3-SMALL on SPHINX using the *LANG* strategy further boosts the performance by **7%pt** and **4%pt** respectively as compared to the vanilla fine-tuned model (SPHINX -0S).

Furthermore, employing the *LANG* strategy leads to additional performance improvements indicating that *LANG* can effectively enhance the multilingual capabilities of LLMs. From the detailed results in Appendix §A.4, we observe no performance regression on high resource languages which normally occurs due to catastrophic forgetting (Chang et al., 2023).

We also observe significant performance improvements in medium and low-resource languages such as Arabic, Hindi, Thai, Turkish, Tamil, and Telugu, further showcasing the effectiveness of our dataset and the *LANG* fine-tuning strategy (Appendix §A.4).

6.2 Comparisons with the API translated dataset

Due to the code-mixed nature of the instruction along with CoT reasoning explanations, a single sample of sPHINX is notably richer as compared to its counterparts from the other datasets. This can be observed in the Table 1 for sPHINX-T wherein the sPHINX trained models with the *LANG* strategy outperform the directly translated dataset baselines by **9%pt** and **4%pt** for both MISTRAL-7B and PHI-3-SMALL respectively when compared to the sPHINX-T baselines. Consequently, even with fewer samples (keeping the number of the tokens same), models trained on sPHINX achieve better performance; thereby demonstrating the per-sample efficiency of sPHINX

6.3 Regression Analysis on Standard LLM Benchmarks

It is well-studied that training on multiple languages cause regression on the performance on English language due to catastrophic forgetting (Chang et al., 2023). We test this by checking for performance of the PHI-3-SMALL model fine-tuned with sPHINX on English in the multilingual benchmarks we evaluate ((Appendix §A.4) and on popular English-only benchmarks.

We find that the PHI-3-SMALL sPHINX model maintains its performance in English on the multilingual benchmarks and is also consistently able to maintain performance on standard English benchmarks such as MMLU (5-shot) Hendrycks et al. (2021), MedQA (2-shot) Jin et al. (2021), Arc-C (10-shot), Arc-E (10-shot) Clark et al. (2018), PiQA (5-shot) Bisk et al. (2020), WinoGrande (5-shot) Sakaguchi et al. (2021), OpenBookQA (10-shot) Mihaylov et al. (2018), BoolQ (2-shot) Clark et al. (2019) and CommonSenseQA (10-shot) Talmor et al. (2018) (Table 12). We notice some regression in the GSM-8k (8-shot, CoT) Cobbe et al. (2021) benchmark. This indicates that gains in multilingual performance caused by sPHINX do not come at the cost of regression in English performance.

7 Conclusion

In this paper, we demonstrated how instruction tuning MISTRAL-7B and PHI-3-SMALL on sPHINX effectively improve their multilingual capabilities. We observed that instruction tuning the models using the sPHINX dataset leads to consistent gains by **5%pt** for both MISTRAL-7B and PHI-3-SMALL when compared to their corresponding base models. Moreover, sPHINX exhibits greater sample efficiency and diversity compared to other multilingual instruction tuning datasets.

Additionally, we proposed a method for further enhancing the model's performance by utilizing *LANG*, which supplements the training examples with N samples from a few-shot set, providing the model with additional context to aid in its learning process. This further boosts the performance for both MISTRAL-7B and PHI-3-SMALL by **9%pt** and **4%pt** respectively when compared to the vanilla fine-tuned model with sPHINXẆe also observed an increment of **7%pt** and **4%pt** on both MISTRAL-7B and PHI-3-SMALL fine-tuned using the *LANG* strategy when compared to the sPHINX-T baseline which involved directly translating the source dataset. Models instruction-tuned on sPHINX, exhibit enhanced performance even in languages they have not previously encountered during training. Finally, we evaluate the performance of the PHI-3-SMALL model fine-tuned on sPHINX and find that it is able to maintain English performance, suggesting that gains in multilingual performance do not come at the cost of English performance while fine-tuning with sPHINX.

8 Future Work

We have conducted all experiments using 7B base models in full parameter fine-tuning settings. It would be interesting to study the same effect by adaptive learning using LoRA (Hu et al., 2022) or PEFT (Mangrulkar et al., 2022). We believe that our strategies could also be effective for models with fewer parameters, leading to a notable improvement in multilingual performance.

Our *LANG* strategy involves utilizing N examples from the same language. Other ideas that can be explored are incorporating N examples from the same language script to increase the diversity of the sample. This approach could be particularly beneficial for enhancing the performance of models in low-resource languages.

Limitations

Our study has several limitations that can be considered in future research. Firstly, we conducted an extensive series of experiments, utilizing significant GPU resources and substantial time for model fine-tuning. Due to these resource-intensive processes, it may be difficult to apply our strategies to fully fine-tune a model. Besides, our study is confined to 7B models, explicitly excluding larger models. Despite this limitation, we believe our methodologies are broadly applicable for fine-tuning smaller datasets using techniques like LoRA and PEFT. Secondly, our fine-tuning dataset focuses on reasoning tasks and excludes some low-resource languages. We evaluated the models' performance against these reasoning benchmarks. However, we did not benchmark our models on generative tasks such as summarization, nor did we evaluate models on hallucination, toxicity, or fairness.

Ethics Statement

Despite our rigorous efforts to ensure that our dataset is free from discriminatory, biased, or false information, there remains a possibility that these problems are present, particularly in multilingual contexts. Hence, it is possible that these issues might propagate to our fine-tuned models as well. We are committed to mitigate such risks and strongly advocate for the responsible use of recipes and prevent any unintended negative consequences.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

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 arXiv:2303.08774*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Akari Asai, Shayne Longpre, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H Clark, and Eunsol Choi. 2022. Mia 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 108–120.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *arXiv preprint arXiv:2311.05845*.

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

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. 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.

T Tony Cai and Rong Ma. 2022. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, 23(301):1–54.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint arXiv:2311.09205*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

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

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

Gautier Dagan, Gabriele Synnaeve, and Baptiste Rozière. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. *arXiv preprint arXiv:2402.01035*.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan, et al. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv:2401.15006*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junpeng Liu, Kaiyu Huang, Hao Yu, Jiuyi Li, Jinsong Su, and Degen Huang. 2023. Continual learning for multilingual neural machine translation via dual importance-based model division. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12011–12027, Singapore. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

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*.

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.

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

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning.

Kocmi Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *WMT23-Eighth Conference on Machine Translation*, pages 198–216.

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

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Hai Wang, Dian Yu, Kai Sun, Janshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual models with vocabulary expansion. *arXiv preprint arXiv:1909.12440*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models.

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

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

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

A Appendix

A.1 Prompt Templates

Figure 5 is the template for **Selective Translation** that was used to generate the synthetic data. Our reference dataset is in English and the `{language}` is the target language to generate the data in. Figure 6, 7, 8,9 and 10 are the prompts used to evaluate XQuAD, XstoryCloze, Xwinograd, XCOPA and Belebele respectively.

A.2 Data Filtering

Algorithm 1: Data Filtering Algorithm

Input: listOfSentences: list of strings

```
Function dataFilter(listOfSentences):
    englishWords ← set of English words from NLTK;
    foreach sentence in listOfSentences do
        cleanedSentence ← replace all punctuations, digits, and single characters with a single space;
        cleanedSentence ← replace all sequences of whitespace with a single space;
        wordCount ← 0;
        foreach word in cleanedSentence do
            if word.lower() in englishWords then
                wordCount ← wordCount + 1;
        content ← wordCount / len(cleanedSentence);
        if content > 0.90 then
            remove sentence from listOfSentences
```

A.3 Hyperparameters and Training Setup

We used 5 nodes with each node containing 8 A100 GPUs with 80GB VRAM. These nodes communicated with each other using InfiniBand [9]. We use DeepSpeed (Rasley et al., 2020) to do distributed fine-tuning over these GPUs. We use the same hyperparameters (Table 2) to fine-tune both MISTRAL-7B and PHI-3-SMALL models.

A.4 Detailed Results

Tables 5, 6, 7, 8, 9, 10 and 11 show the granular results on our models and dataset.

[9] https://network.nvidia.com/pdf/whitepapers/IB_Intro_WP_190.pdf

Hyperparameter	Value
Batch Size	512
Context length	8192
Learning Rate	10^{-5}
Scheduler	Cosine
Epochs	10
Weight Decay	0.1
Optimizer	AdamW

Table 2: Hyperparameters for model fine-tuning

N	$p(N)$	N	$p(N)$
0	0.3	4	0.1
1	0.2	5	0.1
2	0.1	6	0.1
3	0.1		

Table 3: Probabilities of selecting number of shots in the *LANG* strategy

Dataset	Average Token Length/Sample
AYA	2240
BACTRIAN	2465
MULTILINGUAL ALPACA	1620
SPHINX-0S	544
SPHINX	3100

Table 4: Average Token Length in each dataset

Language	en	fr	jp	pt	ru	zh	avg
MISTRAL-7B							
Base Model	0.52	0.47	0.52	0.54	0.54	0.50	0.52
IFT	0.61	0.57	0.57	0.57	0.60	0.56	0.58
M-ALPACA	0.61	0.57	0.57	0.57	0.60	0.56	0.58
AYA	0.55	0.56	0.54	0.54	0.56	0.54	0.55
BACTRIAN	0.61	0.57	0.57	0.57	0.60	0.56	0.58
SPHINX-T	0.58	0.61	0.57	0.53	0.54	0.57	0.57
SPHINX-0S	0.75	0.65	0.68	0.67	0.66	0.65	0.68
SPHINX	**0.80**	**0.69**	**0.72**	**0.70**	**0.67**	**0.67**	**0.71**
PHI-3-SMALL							
Base Model	0.86	0.67	0.73	0.77	0.74	0.72	0.75
IFT	0.86	0.78	0.72	0.78	0.77	0.75	0.78
M-ALPACA	0.87	0.76	0.75	0.78	0.76	0.71	0.81
AYA	0.79	0.61	0.67	0.70	0.70	0.66	0.69
BACTRIAN	0.83	0.72	0.71	0.75	0.70	0.68	0.73
SPHINX-T	0.87	0.74	0.74	0.77	0.77	0.72	0.77
SPHINX-0S	0.88	0.75	0.78	**0.79**	0.81	0.76	0.80
SPHINX	**0.89**	**0.76**	**0.79**	**0.79**	**0.82**	**0.77**	**0.84**

Table 5: Language-wise performance of instruction-tuned MISTRAL-7B and PHI-3-SMALL models evaluated on XWinograd (0-shot). Metric: Accuracy. The best performing IFT dataset for each model is indicated in bold, and the overall best performing IFT model is indicated with an underline.

Language	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
MISTRAL-7B													
Base Model	0.62	0.81	0.64	0.89	0.86	0.65	0.82	0.71	0.59	0.68	0.79	0.72	0.73
IFT	0.42	0.68	0.33	0.92	0.66	0.5	0.71	0.61	0.38	0.63	0.71	0.68	0.60
M-ALPACA	0.10	0.75	0.15	0.86	0.82	0.12	0.62	0.68	0.12	0.38	0.52	0.46	0.46
AYA	0.33	0.73	0.65	0.85	0.80	0.63	0.75	0.67	0.57	0.61	0.75	0.59	0.66
BACTRIAN	0.67	0.76	0.26	0.85	0.86	0.74	0.77	0.71	0.59	0.69	0.77	0.65	0.69
SPHINX-T	0.71	0.83	0.75	0.92	0.89	0.77	0.84	0.75	0.60	**0.77**	0.86	0.63	0.78
SPHINX-0S	0.54	0.76	0.70	0.88	0.84	0.69	0.77	0.66	0.52	0.64	0.71	0.60	0.69
SPHINX	**0.74**	**0.87**	**0.77**	**0.93**	**0.90**	**0.79**	**0.86**	**0.77**	**0.63**	**0.77**	**0.88**	0.73	**0.80**
PHI-3-SMALL													
Base Model	0.68	0.90	0.77	0.93	0.91	0.61	0.84	0.80	0.55	0.73	0.86	0.69	0.78
IFT	0.71	0.88	0.73	0.92	0.91	0.64	0.84	0.80	0.44	0.70	0.67	0.76	0.75
M-ALPACA	0.55	0.92	0.74	0.96	0.94	0.68	0.87	0.85	0.50	0.73	0.88	0.66	0.77
AYA	0.61	0.89	0.84	0.94	0.93	0.80	0.89	0.82	0.73	0.83	0.91	0.79	0.83
BACTRIAN	0.81	0.92	0.81	0.95	0.95	0.80	0.90	0.84	0.72	0.82	0.91	0.79	0.85
SPHINX-T	0.80	0.91	0.82	0.95	0.94	0.80	0.90	0.83	0.69	0.81	0.91	0.73	0.78
SPHINX-0S	0.75	0.89	0.81	0.94	0.94	0.75	0.87	0.79	0.63	0.77	0.88	0.78	0.82
SPHINX	**0.84**	**0.93**	**0.87**	**0.96**	**0.96**	**0.81**	**0.91**	**0.86**	**0.73**	**0.84**	**0.92**	**0.81**	**0.87**

Table 6: Granular results for XQuAD (3-shot) on our model. Metric: F1. The best performing IFT dataset for each model is indicated in bold, and the overall best performing IFT model is indicated with an underline.

Language	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	en	avg
MISTRAL-7B													
Base Model	0.54	0.51	0.72	0.81	0.49	0.52	0.50	0.53	0.58	0.62	0.78	0.93	0.63
IFT	0.52	0.52	0.69	0.79	0.50	0.51	0.50	0.54	0.57	0.63	0.75	0.90	0.62
M-ALPACA	0.51	0.50	0.52	0.63	0.50	0.50	0.50	0.51	0.51	0.49	0.65	0.74	0.55
AYA	0.57	0.54	0.64	0.67	0.53	0.56	0.57	0.62	0.56	0.61	0.64	0.78	0.61
BACTRIAN	0.52	0.50	0.53	0.60	0.49	0.51	0.50	0.51	0.51	0.52	0.52	0.71	0.54
SPHINX-T	0.57	0.50	0.64	0.72	0.50	0.50	0.58	0.57	0.57	0.62	0.71	0.83	0.61
SPHINX-0S	0.54	0.5	0.58	0.63	0.51	0.55	0.52	0.52	0.54	0.57	0.64	0.8	0.58
SPHINX	**0.64**	**0.54**	**0.73**	**0.80**	**0.53**	**0.61**	**0.59**	**0.63**	**0.67**	**0.66**	**0.80**	**0.91**	**0.68**
PHI-3-SMALL													
Base Model	0.55	0.51	0.80	0.93	0.52	0.54	0.46	0.56	0.61	0.66	0.86	0.98	0.64
IFT	0.55	0.57	0.81	0.93	0.53	0.58	0.48	0.60	0.62	0.69	0.88	0.96	0.68
M-ALPACA	0.53	0.54	0.80	0.92	0.49	0.54	0.51	0.59	0.64	0.68	0.87	0.99	0.68
AYA	0.60	0.55	0.72	0.83	0.52	0.55	0.52	0.62	0.59	0.69	0.75	0.89	0.65
BACTRIAN	**0.62**	0.56	0.83	0.91	0.52	**0.60**	0.52	0.66	0.65	0.71	0.86	0.98	0.70
SPHINX-T	0.55	0.58	0.83	0.93	0.51	0.57	0.56	0.66	0.68	0.68	0.87	0.98	0.70
SPHINX-0S	0.59	0.58	0.84	0.93	0.50	0.60	0.54	0.63	0.68	**0.72**	0.89	0.96	0.71
SPHINX	0.59	**0.60**	**0.85**	**0.94**	**0.52**	0.57	**0.58**	**0.68**	**0.69**	0.71	**0.90**	**0.99**	**0.72**

Table 7: Granular results for XCOPA (4-shot) on our model. Metric: Accuracy. The best performing IFT dataset for each model is indicated in bold, and the overall best performing IFT model is indicated with an underline.

Language	ar	en	es	eu	hi	id	my	ru	sw	te	zh	avg
MISTRAL-7B												
Base Model	0.65	0.89	0.83	0.56	0.62	0.76	0.52	0.81	0.56	0.52	0.80	0.68
IFT	0.70	0.95	0.92	0.54	0.69	0.79	0.57	0.90	0.58	0.54	0.88	0.73
M-ALPACA	0.53	0.73	0.70	0.51	0.51	0.57	0.50	0.66	0.52	0.52	0.71	0.59
AYA	0.64	0.86	0.81	0.56	0.71	0.73	0.60	0.82	0.67	0.60	0.81	0.71
BACTRIAN	0.69	0.82	0.74	0.52	0.59	0.76	0.54	0.73	0.62	0.61	0.76	0.67
SPHINX-T	0.78	0.92	0.87	0.56	0.80	0.81	0.66	0.86	0.74	0.70	0.86	0.78
SPHINX-0S	0.57	0.66	0.64	0.47	0.56	0.61	0.50	0.62	0.56	0.52	0.69	0.58
SPHINX	**0.83**	**0.96**	**0.94**	**0.57**	**0.84**	**0.87**	**0.67**	**0.91**	**0.80**	**0.69**	**0.94**	**0.81**
PHI-3-SMALL												
Base Model	0.80	0.98	0.96	0.61	0.72	0.92	0.53	0.96	0.61	0.55	0.94	0.78
IFT	0.81	0.98	0.96	0.61	0.75	0.92	0.56	0.96	0.61	0.53	0.94	0.79
M-ALPACA	0.81	0.98	0.98	0.58	0.76	0.93	0.52	0.97	0.64	0.54	0.96	0.79
AYA	0.77	0.98	0.97	0.57	0.77	0.93	0.53	0.96	0.74	0.56	0.94	0.79
BACTRIAN	0.83	0.98	0.98	0.61	0.83	0.94	0.54	0.97	**0.79**	0.63	0.94	0.82
SPHINX-T	0.84	0.98	0.98	0.60	0.80	0.95	0.52	0.96	0.72	0.60	0.85	0.80
SPHINX-0S	0.84	0.98	0.97	**0.64**	0.77	0.95	0.52	0.96	0.74	0.57	0.95	0.81
SPHINX	**0.86**	**0.99**	**0.99**	0.61	**0.82**	**0.96**	**0.54**	**0.98**	0.74	**0.61**	**0.97**	**0.82**

Table 8: Granular results for XStoryCloze (4-shot) on our model. Metric: Accuracy. The best performing IFT dataset for each model is indicated in bold, and the overall best performing IFT model is indicated with an underline.

Language	ar	de	es	en	fi	fr	hi	it	jp	ko	ta	te	vi	zh	avg
MISTRAL-7B															
Base Model	0.25	0.23	0.23	0.24	0.23	0.23	0.26	0.24	0.26	0.23	0.23	0.25	0.26	0.25	0.24
IFT	0.32	0.60	0.62	0.74	0.36	0.62	0.32	0.61	0.43	0.47	0.27	0.27	0.39	0.58	0.47
M-ALPACA	0.32	0.50	0.53	0.56	0.45	0.51	0.27	0.51	0.40	0.41	0.26	0.26	0.33	0.48	0.41
AYA	0.34	0.43	0.43	0.48	0.38	0.47	0.35	0.44	0.4	0.36	0.27	0.25	0.37	0.42	0.38
BACTRIAN	0.24	0.27	0.25	0.25	0.26	0.27	0.24	0.28	0.26	0.26	0.23	0.23	0.34	0.28	0.26
SPHINX-T	0.66	0.74	0.71	0.81	0.66	0.76	0.57	0.73	0.66	0.68	0.54	0.46	0.66	0.71	0.68
SPHINX-0S	0.64	0.75	**0.75**	0.82	0.66	0.79	0.53	0.73	0.69	0.66	0.48	0.44	0.66	0.75	0.67
SPHINX	**0.69**	**0.80**	0.69	**0.87**	**0.71**	**0.82**	**0.60**	**0.79**	**0.73**	**0.73**	**0.56**	**0.48**	**0.70**	**0.80**	**0.71**
PHI-3-SMALL															
Base Model	0.54	0.87	0.85	0.92	0.58	0.86	0.41	0.86	0.70	0.58	0.26	0.30	0.62	0.82	0.65
IFT	0.63	0.89	0.88	0.93	0.63	0.88	0.48	0.88	0.77	0.68	0.32	0.32	0.68	0.85	0.70
M-ALPACA	0.65	0.92	0.90	0.94	0.74	0.91	0.54	0.90	0.80	0.70	0.47	**0.45**	0.72	0.84	0.75
Aya	0.58	0.86	0.85	0.91	0.65	0.87	0.50	0.86	0.76	0.67	0.37	0.35	0.69	0.84	0.70
BACTRIAN	0.67	0.88	0.88	0.92	0.70	0.88	0.51	0.86	0.77	0.70	0.37	0.37	0.74	0.86	0.72
SPHINX-T	0.71	0.90	0.90	0.93	0.73	0.90	0.56	0.89	0.82	0.75	0.42	0.38	0.75	0.87	0.75
SPHINX-0S	0.73	0.91	0.90	0.93	0.75	0.92	0.57	0.91	0.82	**0.82**	0.45	0.40	0.76	**0.89**	0.77
SPHINX	**0.74**	**0.93**	**0.91**	**0.94**	**0.77**	**0.93**	**0.58**	**0.92**	**0.84**	0.76	**0.46**	0.40	**0.78**	**0.89**	**0.79**

Table 9: Granular results for Belebele (0-shot) on our model. Metric: Accuracy. The best performing IFT dataset for each model is indicated in bold, and the overall best performing IFT model is indicated with an underline.

Language	ar	fr	de	ro	ja	ru	zh	avg
MISTRAL-7B								
Base Model	0.48	0.63	0.65	0.56	0.43	0.54	0.48	0.54
IFT	0.37	0.61	0.61	0.58	0.28	0.5	0.49	0.49
M-ALPACA	0.34	0.51	0.51	0.46	0.29	0.36	0.41	0.41
AYA	0.30	0.50	0.51	0.46	0.24	0.35	0.41	0.39
BACTRIAN	0.40	0.55	0.52	0.51	0.34	0.44	0.42	0.45
SPHINX-T	0.39	0.60	0.60	0.54	0.39	0.49	0.48	0.49
SPHINX-0S	0.40	0.57	0.61	0.53	0.40	0.51	0.44	0.49
SPHINX	**0.45**	**0.63**	**0.64**	**0.59**	**0.45**	**0.55**	**0.52**	**0.54**
PHI-3-SMALL								
Base Model	0.48	0.61	0.65	0.57	0.45	0.52	0.53	0.54
IFT	0.43	0.62	0.63	0.52	0.41	0.49	0.50	0.54
M-ALPACA	0.44	0.60	0.61	0.56	0.32	0.44	0.19	0.45
AYA	0.44	0.56	0.57	0.54	0.12	0.45	0.17	0.41
BACTRIAN	0.48	0.63	0.65	0.59	0.45	0.52	0.52	0.54
SPHINX-T	0.48	0.64	0.65	0.59	0.46	0.53	0.52	0.55
SPHINX-0S	**0.49**	0.63	0.65	0.59	**0.46**	**0.54**	**0.53**	**0.56**
SPHINX	**0.49**	**0.64**	**0.66**	**0.60**	**0.46**	**0.54**	**0.53**	**0.56**

Table 10: Granular results for Translation for language to English direction (4-shot) on our model. Metric: ChrF. The best performing dataset for each model is indicated in bold, and the overall best performing model is indicated with an underline.

Language	ar	fr	de	ro	ja	ru	zh	avg
MISTRAL-7B								
Base Model	0.29	0.60	0.54	0.52	0.21	0.34	0.47	0.42
IFT	0.16	0.58	0.57	0.48	0.17	0.29	0.43	0.38
M-ALPACA	0.15	0.62	0.66	0.40	0.12	0.48	0.44	0.41
AYA	0.12	0.54	0.63	0.48	0.16	**0.39**	0.42	0.39
BACTRIAN	0.14	0.51	0.49	0.44	0.10	0.35	0.40	0.38
SPHINX-T	0.31	0.58	0.53	0.47	0.20	0.30	0.26	0.38
SPHINX-0s	0.30	0.60	0.55	0.51	0.22	0.31	0.45	0.42
SPHINX	**0.35**	**0.61**	**0.60**	**0.55**	**0.26**	0.36	**0.49**	**0.46**
PHI-3-SMALL								
Base Model	0.31	0.63	0.60	0.43	0.24	0.30	0.45	0.42
IFT	0.29	0.61	0.58	0.39	0.21	0.27	0.41	0.46
M-ALPACA	0.39	0.60	0.58	0.31	0.11	0.24	0.47	0.38
AYA	0.17	0.62	0.56	0.45	0.22	0.28	0.46	0.39
BACTRIAN	0.30	0.60	0.56	0.47	0.18	0.24	0.44	0.40
SPHINX-T	0.33	0.63	0.60	0.48	0.24	0.31	0.48	0.43
SPHINX-0s	0.32	0.63	**0.61**	0.50	**0.27**	**0.36**	**0.49**	0.45
SPHINX	**0.35**	**0.64**	**0.61**	**0.51**	0.26	0.33	**0.49**	**0.46**

Table 11: Granular results for Translation for English to language direction (4-shot) on our model. Metric: ChrF. The best performing dataset for each model is indicated in bold, and the overall best performing model is indicated with an underline.

Benchmarks	Base Model	SPHINX
MMLU (5-shot)	**0.76**	0.75
HellaSwag (5-shot)	0.81	**0.83**
GSM-8k (8-shot, CoT)	**0.85**	0.77
MedQA (2-shot)	0.64	**0.66**
Arc-C (10-shot)	**0.90**	**0.90**
Arc-E (10-shot)	**0.97**	**0.97**
PIQA (5-shot)	0.84	**0.89**
WinoGrande (5-shot)	0.77	**0.82**
OpenBookQA (10-shot)	0.86	**0.88**
BoolQ (2-shot)	0.82	**0.87**
CommonSenseQA (10-shot)	0.80	**0.81**

Table 12: Performance of the PHI-3-SMALL base model and the SPHINX tuned model on standard English LLM benchmarks.

High-Resource (100k)	Spanish, Chinese Simplified, Japanese French, German, Portuguese, Italian
Mid-Resource (50k)	Dutch, Swedish, Danish Finnish, Russian, Norwegian Korean, Chinese Traditional, Polish Turkish, Arabic, Hebrew Portuguese, Czech, Hungarian
Low-Resource (25k)	Indonesian, Thai, Greek Slovak, Vietnamese, Slovenian Croatian, Romanian, Lithuanian Bulgarian, Serbian, Latvian Ukranian, Estonian, Hindi Burmese, Bengali, Afrikaan Punjabi, Welsh, Icelandic Marathi, Swahili, Nepali Urdu, Telugu, Malayalam Russian, Tamil, Oriya

Table 13: Language distribution and samples across three tiers

```
Please carefully convert a conversation between a human and an AI
assistant from English to language. The dialogue will be presented
in JSON format, where 'system' denotes system instructions,
'human' indicates user queries, and 'assistant' refers to the AI's
response. You should approach this task as if the 'human' original
language is {language}. Translate the 'system' instructions fully
into {language}. For the 'human' input, however, carefully discern
which segments require translation into {language}, while leaving
other parts in their original form.
For instance: 1. If the human contains a mix of languages, only
translate the instruction part.
2. If the task is about language correction do not translate the
target passage.

   For the 'assistant' part, generate the 'assistant' response as
you were prompted with ths newly translated system and assistant
instructions. The outcome should retain the JSON format. Your
response should solely contain the JSON. Do not translate the
JSON keys. {"system": System text here, "human": User text here,
"assistant": Assistant text here }
```

Figure 5: Prompt for Selective Translation using GPT-4

```
The task is to solve reading comprehension problems. You will be
provided questions on a set of passages and you will need to provide
the answer as it appears in the passage. The answer should be in
the same language as the question and the passage.
Context:
{context}
Question:
{question}
Referring to the passage above, the correct answer to the given
question is {answer}
```

Figure 6: XQuAD evaluation prompt

```
{input_sentence_1} {input_sentence_2}
{input_sentence_3} {input_sentence_4}
What is a possible continuation for the story given the following
options?
Option1: {sentence_quiz1} Option1: {sentence_quiz2}
```

Figure 7: XstoryCloze evaluation prompt

```
Select the correct option out of option1 and option2 that will fill
in the _ in the below sentence:
{sentence}
Choices:
-option1: {option1}
-option2: {option2}
```

Figure 8: Xwinograd evaluation prompt

```
The task is to perform open-domain commonsense causal reasoning.
You will be provided a premise and two alternatives, where the
task is to select the alternative that more plausibly has a causal
relation with the premise. Answer as concisely as possible in the
same format as the examples below: Given this premise:
{premise}
What's the best option?
-choice1 : {choice1}
-choice2 : {choice2}
We are looking for{% if question == c̈ause%̈} a cause {% else %} an
effect {% endif %}
```

Figure 9: XCOPA evaluation prompt

```
The task is to perform reading comprehension task. Given the
following passage, query, and answer choices, output only the letter
corresponding to the correct answer. Do not give me any explanations
to your answer. Just a single letter corresponding to the correct
answer will suffice.
Passage: {flores_passage}
Query: {question}
Choices:
A: {mc_answer1}
B: {mc_answer2}
C: {mc_answer3}
D: {mc_answer4}
```

Figure 10: Belebele evaluation prompt

```
Translate the following sentence pairs:
{Source Language}: {Source Phrase} {Target Language}: {Target
Phrase}
```

Figure 11: Translation evaluation prompt

Code	Languages	Script	Data
af	Afrikaan	Latin	20206
ar	Arabic	Arabic	26803
bn	Bengali	Bengal	20165
bg	Bulgarian	Cyrillic	17300
my	Burmese	Burmese	12123
zh-Hans	Chinese_Simplified	Han	100650
zh-Hant	Chinese_Traditional	Hant	32363
hr	Croatian	Latin	17340
cs	Czech	Latin	32711
da	Danish	Latin	36348
nl	Dutch	Latin	36586
en	English	Latin	199900
et	Estonian	Latin	17207
fi	Finnish	Latin	33622
fr	French	Latin	100337
de	German	Latin	100265
el	Greek	Greek	17317
he	Hebrew	Hebrew	24483
hi	Hindi	Devanagari	20240
hu	Hungarian	Latin	31999
is	Icelandic	Latin	20164
id	Indonesian	Latin	17297
it	Italian	Latin	85175
jp	Japanese	Japanese	98366
ko	Korean	Hangul	30890
lv	Latvian	Latin	17247
lt	Lithuanian	Latin	17232
ml	Malayalam	Malayalam	19817
mr	Marathi	Devanagari	20069
ne	Nepali	Devanagari	20092
nb	Norwegian	Latin	36811
or	Oriya	Oriya	19153
pl	Polish	Latin	34711
pt	Portuguese	Latin	37229
pa	Punjabi	Gurmukhi	20026
ro	Romanian	Latin	17149
ru	Russian	Cyrillic	20108
sr	Serbian	Latin	17165
sk	Slovak	Latin	17255
sl	Slovenian	Latin	17300
es	Spanish	Latin	100351
sw	Swahili	Latin	20170
sv	Swedish	Latin	36533
ta	Tamil	Tamil	19807
te	Telugu	Telugu	19947
th	Thai	Thai	17322
tr	Turkish	Latin	34405
uk	Ukrainian	Cyrillic	17282
ur	Urdu	Perso-Arabic	20162
vi	Vietnamese	Latin	17358
cy	Welsh	Latin	20207

Table 14: Language Distribution in Sphinx Dataset