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

Sanchit Ahuja[†*] **Kumar Tanmay**[♡◇*] **Hardik Hansrajbhai Chauhan**[†]
Barun Patra[†] **Kriti Aggarwal**[♣◇] **Luciano Del Corro**[†]
Arindam Mitra[†] **Tejas Indulal Dhamecha**[β◇] **Ahmed Awadallah**[†]
Monojit Choudhury[♠◇] **Vishrav Chaudhary**[α◇△] **Sunayana Sitaram**[†△]
[†]Microsoft Corporation [♡]Harvard University [β]Adobe Research
[♠]MBZUAI University [♣]Hippocratic AI [α]Meta AI
{sanchitahuja205,kr.tanmay147}@gmail.com

Abstract

Despite the remarkable success of large language models (LLMs) in English, a significant performance gap remains in non-English languages. To address this, we introduce a novel approach for strategically constructing a multilingual synthetic instruction tuning dataset, sPhinX . Unlike prior methods that directly translate fixed instruction-response pairs, sPhinX enhances diversity by selectively augmenting English instruction-response pairs with multilingual translations. Additionally, we propose *LANGIT*, a novel N-shot guided fine-tuning strategy, which further enhances model performance by incorporating contextually relevant examples in each training sample. Our ablation study shows that our approach enhances the multilingual capabilities of Mistral-7B and Phi-3-small improving performance by an average of 39.8% and 11.2%, respectively, across multilingual benchmarks in reasoning, question answering, reading comprehension, and machine translation. Moreover, sPhinX maintains strong performance on English LLM benchmarks while exhibiting minimal to no catastrophic forgetting, even when trained on 51 languages.

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., 2024). Further, the gap between the performance of Large Language Models (LLMs) and Small Language Models (SLMs) is more pronounced (Ahuja et al., 2024) in 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 in English (Zhao et al., 2024; Aggarwal et al., 2024). Few techniques have been proposed to bridge this gap, such as incorporating better pre-training data in various languages and improving base tokenizers (Xu et al., 2024; Dagan et al., 2024). However, most of these changes need to be implemented in 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 to enhance 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 limit 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 augmenting the Orca instruction tuning dataset samples(Mukherjee et al., 2023) through *Selective Translated Augmentation* using GPT-4 (Achiam

[*] denotes equal contribution, [△]denotes equal advising, [◇]Work done when the authors were at Microsoft

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 927–946
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

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 across discriminative and generative tasks. 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 across languages. We also compare our proposed translation strategy with translating the entire instruction using Azure 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 *Selective Translated Augmentation* of the Orca dataset with the assistance of GPT-4 (§3.1)
- We devise ***LA**nguage-Specific **N**-shot **G**uided **I**nstruction **T**uning (LANGIT)* strategy for enhancing the multilingual capabilities of LLMs (§4)
- We also conduct extensive instruction tuning experiments on various multilingual instruction tuning datasets to evaluate generalizability in multilingual settings (§6).
- We plan to release a subset of the augmented dataset by applying our strategy to the OpenOrca [1] dataset (Lian et al., 2023) (Open-sPhinX) as well.

2 Related Work

2.1 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 to 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) make 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 (Muennighoff et al., 2023). However, these models still do not perform as well as English in other languages, with the gap being huge 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., 2024; 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.

2.2 Multilingual Synthetic Data Generation

Most instruction-tuning datasets across multiple languages typically focus on general tasks rather than specific reasoning capabilities. Although 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 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. Another example is of this work (Lai et al., 2024), which also utilizes Google Translation API for translating their source datasets. 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. In our work, we also show the pitfalls of training with data generated only using the Translation APIs such Google Translate or Azure Translate.

[1] https://huggingface.co/datasets/Open-Orca/OpenOrca

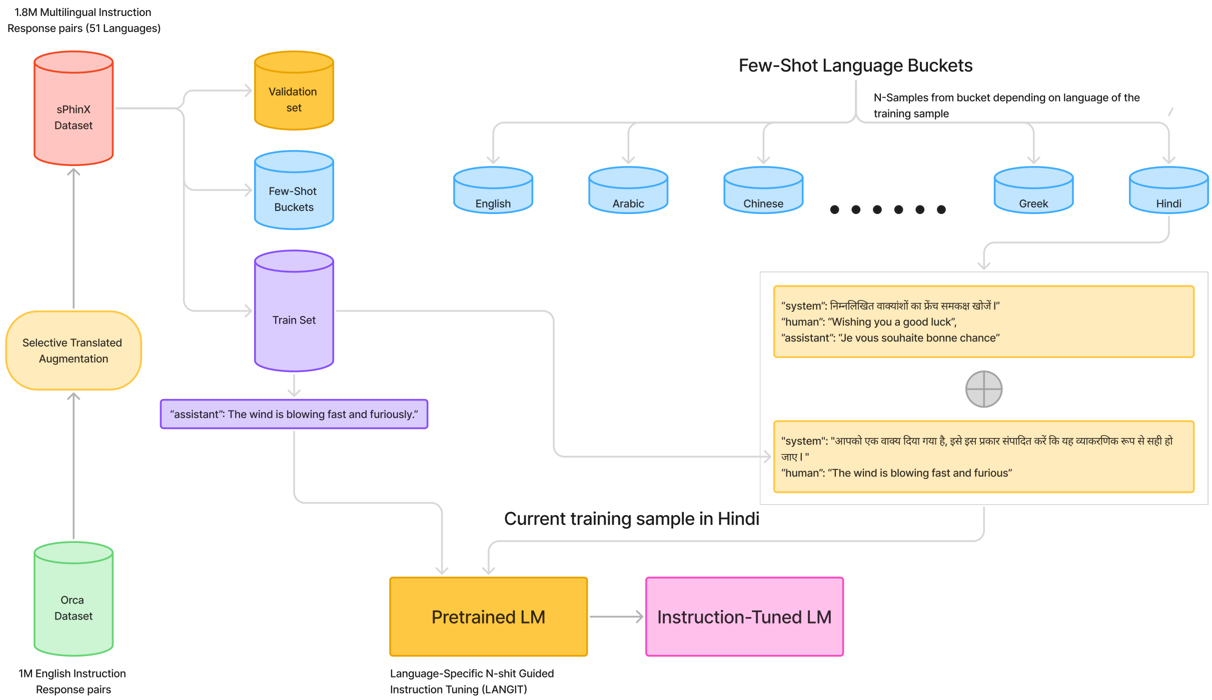

Figure 1: The figure above illustrates pipelines for SPHINX data creation using Selective Translated Augmentation and Multilingual Instruction Tuning using *LANGIT* strategy.

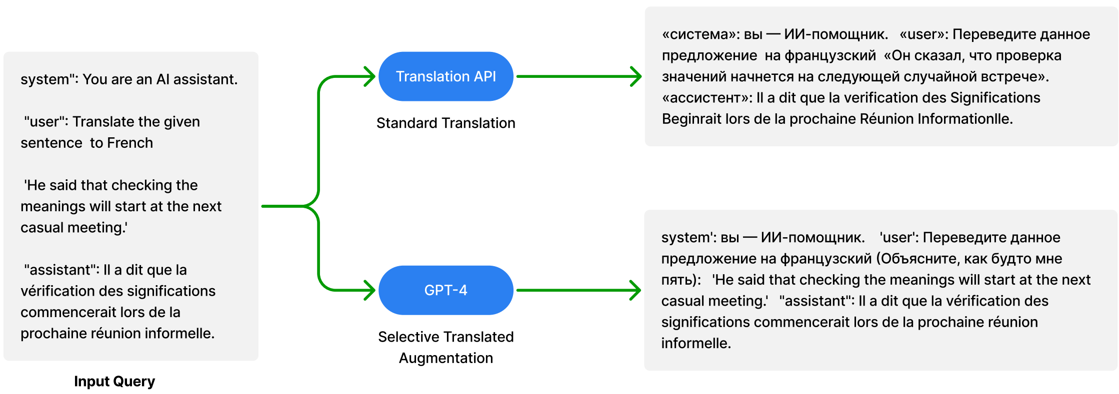

Figure 2: The figure compares the Translation API with Selective Translated Augmentation. The Translation API translates the entire input into Russian, while the Selective strategy localizes only necessary components. Here, system and user prompts are translated, but the input question and assistant's response remain in the original language, preserving structure and intent.

3 SPHINX Dataset

In this section, we describe our dataset construction methodology (§3.1), dataset filtering, and cleaning pipelines (§3.2).

3.1 Dataset Construction

Inspired by (Mukherjee et al., 2023)'s work, we utilized the 1M GPT-4 generated instruction-response pairs from Orca and constructed our own dataset along similar lines using *Selective Translated Augmentation* into 50 different languages with the help of GPT-4[2]. We categorize them into three groups: high-resource, mid-resource, and low-resource languages as outlined in Table 7. For high-resource languages, we randomly sample 100k instruction-response pairs from the Orca 1M dataset and generate the responses from GPT-4 with *Selective Trans-*

[2]GPT-4 inference hyper-parameters in Azure OpenAI interface set as: `temperature=0.0`

lated Augmentation as shown in Figure 2. 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 on medium and low resource languages (Jiao et al., 2023; Hendy et al., 2023). Furthermore, as highlighted in (Chang et al., 2023; Lin et al., 2023; Xia et al., 2024), 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. Besides, (Shaham et al., 2024) also demonstrates that a small number of multilingual training samples is sufficient to significantly boost multilingual performance, validating our approach of using fewer samples from medium- and low-resource languages.

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 ambiguous instruction. To mitigate this issue, we used GPT-4 to augment the instructions using *Selective Translated Augmentation*, so that task-specific components of instruction responses are translated into the appropriate language without changing the semantic meaning. Figure 12 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 M-ALPACA dataset: wherein the translation of the task input results in the task being ill-defined based on the instructions. As demonstrated, our proposed *Selective Translated Augmentation* method is able to keep the semantic information of the task intact while translating the instructions. For the exact prompt, please refer to Figure 4 in the Appendix.

3.2 Dataset Filtering and Quality Assessment

After creating the dataset, we filtered out samples where GPT-4 failed to generate a response. The final dataset comprised 1.8 million samples in 51 languages(Table: 15), divided into three subsets: Train, Validation, and Few-shot. Each language's dataset was partitioned to ensure that the Validation and Few-shot sets contained 2,000 and 1,000 samples, respectively, while the Train set included the remaining data. This approach guarantees consistent distributions across languages in the Validation and Few-shot sets, ensuring equitable representation regardless of the training distribution. The final split ratio for Train, Validation, and Few-shot sets was 92:5.3:2.7.

We also conducted a small-scale quality assessment of the generated data for languages such as Bengali, Hindi, German, Turkish, and Tamil. The researchers and engineers in our organization, who are native speakers of these languages, evaluated the data on the basis of coherence, fluency, and information retention. Our findings indicate that the generated dataset is moderate to high quality.

3.3 Sample Diversity in SPHINX

Unlike prior multilingual datasets such as BACTRIAN and M-ALPACA, which translate a fixed set of instruction-response pairs into multiple languages, SPHINX ensures diversity by sampling unique subsets of instruction-response pairs for each language.

For instance, BACTRIAN is constructed from 67k English instruction-response pairs (Alpaca + Dolly) and translated into 52 languages, resulting in identical samples across all languages. In contrast, SPHINX samples from 1M GPT-4-generated instruction-response pairs, ensuring that no two languages share the exact same subset.

Mathematically, the probability that all samples in one language dataset A are also in another language dataset B, when sampled without replacement from a larger dataset D, is given by:

$$P(\text{all A in B}) \approx \left(\frac{m}{N}\right)^n = \left(\frac{100{,}000}{1{,}000{,}000}\right)^{20{,}000}$$

where:

- $N = 1{,}000{,}000$ (Total samples in SPHINX),
- $n = 20{,}000$ (Samples in language A),
- $m = 100{,}000$ (Samples in language B).

Since the exponential term results in an extremely small probability, this confirms that no two languages have identical instruction-response sets in SPHINX.

To further enhance diversity, we apply *Selective Translated Augmentation*, translating 10% of samples for high-resource languages, 5% for mid-resource languages, and 2.5% for low-resource languages. This ensures that translated content varies across languages, preventing uniformity.

Additionally, code-switching naturally emerges from this augmentation process, further increasing linguistic diversity. Compared to AYA, which exhibits moderate variation across task instructions, SPHINX introduces greater sample diversity by leveraging a larger and more heterogeneous seed set (Mukherjee et al., 2023) and selective augmentation strategy. Exploring code-switching phenomena would be an interesting task in these synthetically generated datasets, but currently that is out-of-scope of for this work.

4 *LANGIT*

Following Instruction Tuning strategies of (Longpre et al., 2023) and also taking inspiration from (Min et al., 2022), we devise *Language-Specific N-shot Guided Instruction fine-tuning (LANGIT)* Figure: 1. 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. This added context helps guide the model, enabling it to learn more effectively from the provided examples.

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 using the *LANGIT* strategy detailed in the next section (§6). Additionally, we fine-tuned the models on the SPHINX dataset without using *LANGIT* 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 6, showing the average token lengths as tokenized by the PHI-3-SMALL model.

5 Experiments

5.1 Setup

Base Models: We use MISTRAL-7B[3] 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), M-ALPACA (Wei et al., 2023b) and AYA (Singh et al., 2024b) instruction datasets for comparative evaluation. We also utilize the Azure Translator API[4] (SPHINX-T) to translate the original dataset into all our target languages, demonstrating the effectiveness of our *Selective Translated Augmentation* approach. More details about the datasets used for comparative evaluation are present in Appendix §A.2.

Evaluation: We evaluate[5] our fine-tuned models along with the available base and Instruction fine-tuned model variants of MISTRAL-7B and PHI-3-SMALL (IFT[6]) on 4 discriminative tasks; XCOPA (Ponti et al., 2020)(4-shot), XStoryCloze (Lin et al., 2022)(4-shot), XWinograd (Muennighoff et al., 2023)(0-shot), (Tikhonov and Ryabinin, 2021), Belebele(0-shot) (Bandarkar et al., 2023), and 2 generative tasks; XQuAD (3-

[3]We specifically use the v1.0 base model from https://huggingface.co/mistralai/Mistral-7B-v0.1

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

[5]Evaluation prompts and other details in Appendix §A.1 and §A.3.

[6]We take the MISTRAL-7B instruction-tuned variant from https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1 and PHI-3-SMALL variant from https://huggingface.co/microsoft/Phi-3-small-8k-instruct.

shot) (Artetxe et al., 2020) and Translation (4-shot) (Bojar et al., 2014, 2016; Kocmi et al., 2023) using the language model evaluation harness (Gao et al., 2023). The number of few-shot selection are inspired from these works (Ahuja et al., 2023, 2024; Asai et al., 2024)

Apart from generative tasks such as XQuAD, and machine translation, we also evaluate our instruction-tuned models on open-ended generation prompts. For this, we use an LLM-based evaluation approach to simulate win rates. We use the open-source test set from the Aya Dataset (Singh et al., 2024a), which includes 250 prompts per language across six languages. We use GPT-4o as the LLM evaluator to pick the preferred model generation on this test set, and we subsequently compute win rates (%) based on these preferences. To avoid a potential bias, we randomize the order of the models during the evaluation. The prompt for the evaluator is described in the Appendix: §A.1.

6 Results

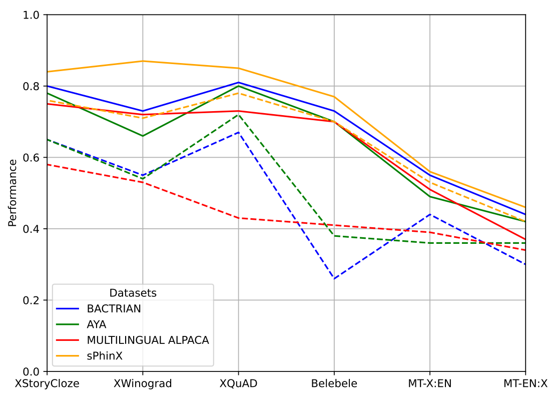

Figure 3: Performance of MISTRAL-7B and PHI-3-SMALL when instruction-tuned on 8B tokens across various datasets on different benchmarks. The solid lines represent the Phi fine-tuned models and the dashed lines represent the Mistral fine-tuned models.

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 average improvement of **39.8%** and **11.2%** respectively on both the models. (Refer to SPHINX-0S in Table 1 for overall results and to Appendix §A.5 for language-wise results). Additionally, as observed in Figure 3, 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 an equal number of tokens.

7 Ablations

7.1 Improvements from *LANGIT*

To demonstrate the effectiveness of our *LANGIT* strategy, we also instruction-tuned the models on SPHINX with 0 shots, referring to this as SPHINX-0s. As shown in Table 1 (with detailed results in Appendix §A.5), models fine-tuned on SPHINX especially MISTRAL-7B exhibit 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 *LANGIT* strategy further boosts the performance by an average of **15%** and **3.2%** respectively as compared to the vanilla fine-tuned model (SPHINX -0s) across multilingual benchmarks.

Furthermore, employing the *LANGIT* strategy leads to additional performance improvements indicating that *LANGIT* can effectively enhance the multilingual capabilities of LLMs. From the detailed results in Appendix §A.5, 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 *LANGIT* fine-tuning strategy (Appendix §A.5).

7.2 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 *LANGIT* strategy outperform the directly translated dataset baselines by an average of **11.7%** and **6.3%** for both MISTRAL-7B and PHI-3-SMALL respectively when compared to the SPHINX-T baselines across multilingual benchmarks. Consequently, even with fewer samples as compared to the other datasets (keeping the number of the tokens the same), mod-

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	<u>**0.79**</u>	<u>**0.56**</u>	0.45
sPHINX	<u>**0.72**</u>	<u>**0.84**</u>	<u>**0.87**</u>	<u>**0.87**</u>	<u>**0.79**</u>	<u>**0.56**</u>	<u>**0.46**</u>

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.

els trained on sPHINX achieve better performance, thereby demonstrating the per-sample efficiency of sPHINX.

7.3 Simulated Preference Evaluation

As shown in Table 2, our win-rate experiments reveal that the GPT-4o evaluator predominantly favored outputs generated by the Mistral base model trained on the sPHINX dataset using the *LANGIT* strategy over other models. For the Phi baselines, we observed a higher percentage of TIEs for all the languages except English, where the evaluator rated both outputs equally, rather than favoring a specific model. This performance gap between the Mistral and Phi models likely arises from the age of their respective base models. Since the Mistral base model is older, it benefits more from additional training on our dataset, whereas the more recently released Phi models are already competitive enough on these benchmarks resulting in preferring both the outputs equally.

7.4 Regression Analysis on Standard LLM Benchmarks

It is well studied that training in multiple languages causes regression in performance in English due to catastrophic forgetting (Chang et al., 2023). We test this phenomenon for our trained models by checking the performance of the PHI-3-SMALL model fine-tuned with sPHINX on English in the multilingual benchmarks we evaluate ((Appendix §A.5) and on popular English-only benchmarks (Table 3).

We find that the PHI-3-SMALL fine-tuned on sPHINX 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 3). 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.

8 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 performance improvement by an average of **39.8%** and **11.2%**

Model	ar	en	po	te	tu	zh
MISTRAL-7B						
IFT	**75** / 9 / 16	59 / **36** / 5	57 / 27 / **16**	62 / 15 / **23**	**65** / 13 / 22	**70** / 26 / 4
M-ALPACA	**85** / 2 / 13	**74** / 21 / 5	52 / **33** / 15	**69** / 8 / 23	**70** / 4 / 25	**75** / 15 / 10
AYA	**78** / 11 / 11	**85** / 11 / 4	55 / 30 / **15**	56 / 18 / **25**	62 / **19** / 19	**78** / 14 / 8
BACTRIAN	**82** / 4 / 13	**85** / 12 / 3	57 / **31** / 12	56 / 18 / **26**	62 / **20** / 18	**74** / 14 / 12
SPHINX-T	61 / **23** / 16	63 / **27** / 10	56 / 24 / **20**	56 / 24 / **20**	62 / **19** / 19	66 / **23** / 11
SPHINX-0S	65 / **19** / 16	**74** / 20 / 5	55 / **31** / 14	51 / 22 / **27**	52 / **36** / 12	68 / **20** / 11
PHI-3-SMALL						
IFT	**46** / 38 / 17	55 / **43** / 2	50 / **44** / 6	39 / 29 / **36**	30 / 27 / **43**	50 / **44** / 6
M-ALPACA	**46** / 5 / 49	59 / **29** / 12	44 / 21 / **35**	33 / 6 / **60**	31 / 10 / **59**	30 / 6 / **64**
AYA	40 / 18 / **42**	**80** / 11 / 9	**56** / 16 / 28	37 / 18 / **46**	25 / 21 / **54**	36 / 22 / **41**
BACTRIAN	32 / 13 / **55**	**68** / 17 / 15	51 / 14 / **36**	24 / 14 / **62**	32 / 15 / **53**	26 / 11 / **63**
SPHINX-T	35 / 14 / **51**	**75** / 12 / 13	**61** / 12 / 27	23 / 14 / **63**	36 / 18 / **47**	37 / 12 / **51**
SPHINX-0S	23 / 11 / **66**	61 / 17 / **22**	32 / 18 / **50**	14 / 9 / **77**	15 / 10 / **74**	14 / 7 / **79**

Table 2: Win rates (%) according to GPT-4o: The first value represents the percentage of outputs where the evaluator preferred the SPHINX and *LANGIT* trained model. The second value indicates the percentage of outputs preferred from the target model. The third value reflects cases where the evaluator rated both outputs equally (TIE).

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

for MISTRAL-7B and PHI-3-SMALL respectively when compared to their corresponding base models across multilingual benchmarks. Moreover, SPHINX exhibits greater sample efficiency and diversity compared to other multilingual instruction tuning datasets. We also proposed *LANGIT*, a strategy that enhances model performance by incorporating N few-shot examples, boosting results by **15%** and **3.2%** for MISTRAL-7B and PHI-3-SMALL, respectively, over vanilla fine-tuning with SPHINX. Compared to the SPHINX-T translation baseline, *LANGIT* yielded gains of **11.7%** and **6.3%**. Models fine-tuned on SPHINX also showed improved performance in unseen languages without degrading English performance. We also observed that the GPT-4o win-rate evaluations favored MISTRAL-7B with *LANGIT*, while PHI-3-SMALL showed more ties due to its stronger baseline. Finally, we plan on releasing a subset of our augmented dataset built on OpenOrca (OPEN-SPHINX).

9 Future Work

All experiments were conducted using 7B base models with full fine-tuning. It would be interesting to explore our methods with adaptive fine-tuning techniques like LoRA (Hu et al., 2022) or PEFT (Mangrulkar et al., 2022), and on smaller models, where we expect similar gains in multilingual performance. Our *LANGIT* strategy uses N examples from the same language and future work could investigate using N examples from the same script to introduce greater diversity, especially for improving performance in low-resource languages. We also observed code-switched data being generated when we employed this strategy to generate data. It will be interesting to explore this phenomena in a future study.

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 that our methodologies are broadly applicable for fine-tuning smaller datasets using techniques like LoRA and PEFT.

While some languages were not included, we made a conscious effort to cover a diverse set spanning multiple scripts and language families.

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 mitigating 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*.

Divyanshu Aggarwal, Ashutosh Sathe, and Sunayana Sitaram. 2024. Exploring pretraining via active forgetting for improving cross lingual transfer for decoder language models. *arXiv preprint arXiv:2410.16168*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

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 Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

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.

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*.

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.

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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

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.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`.

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.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*.

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*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. 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*.

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.

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.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024a. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

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. 2024b. 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`.

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.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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*.

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 4 is the template for *Selective Translated Augmentation* 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 5, 6, 7,8, 9 and 10 are the prompts used to evaluate XQuAD, XstoryCloze, Xwinograd, XCOPA, Belebele and Translation respectively. Figure 11 denotes the prompt used for simulating win-rate evaluations.

A.2 Baseline Datasets For Comparative Evaluation

- 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.
- M-ALPACA (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 data points.
- AYA (Singh et al., 2024a) 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.

A.3 Evaluation Benchmarks

- **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.

A.4 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 [7]. We use DeepSpeed (Rasley et al., 2020) to do distributed fine-tuning over these GPUs. We use the same hyperparameters (Table 4) to fine-tune both MISTRAL-7B and PHI-3-SMALL models.

A.5 Detailed Results

Tables 8, 9, 10, 11, 12, 13 and 14 show the granular results on our models and dataset.

```
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 4: Prompt for Selective Translation using GPT-4

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

```
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 5: 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 6: 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 7: 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 8: 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 9: Belebele evaluation prompt

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

Figure 10: Translation evaluation prompt

```
System: You are a helpful following assistant whose goal is to
select the preferred (least wrong) output for a given instruction
in {LANGUAGE_NAME}.
User: Which of the following answers is the best one for given
instruction in {LANGUAGE_NAME}. A good answer should follow these
rules: 1) It should be in {[LANGUAGE_NAME}.
2) It should answer the request in the instruction.
3) It should be factually and semantically comprehensible.
4) It should be grammatically correct and fluent.

Instruction: {INSTRUCTION}
Answer (A): {COMPLETION A}
Answer (B): {COMPLETION B}

FIRST provide a one-sentence comparison of the two answers,
explaining which you prefer and why. SECOND, on a new line, state
only ‘Answer (A)’ or ‘Answer (B)’ to indicate your choice. If the
both answers are equally good or bad, state ‘TIE’. Your response
should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <‘Answer (A)’ or ‘Answer (B)’ or ‘TIE’>
```

Figure 11: Preference simulation prompt taken from (Üstün et al., 2024) evaluation suite to evaluate our models on free-form generation using GPT-4o.

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

Table 4: 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 5: Probabilities of selecting number of shots in the *LANG* strategy

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

Table 6: Average Token Length in each dataset

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 12: 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 Translated Augmentation strategy. Again, we observe that the samples generated using Selective Translated Augmentation translate only the required amount of information as controlled via prompting whereas in MULTIALPACA the translations are direct translations where only a part of the instructions have been followed to translate the input queries.

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 7: Language distribution and samples across three tiers

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 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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	<u>**0.39**</u>	0.42	0.39
BACTRIAN	0.14	0.51	0.49	0.44	0.10	<u>0.35</u>	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**	<u>**0.55**</u>	**0.26**	0.36	<u>**0.49**</u>	**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	<u>**0.61**</u>	0.50	<u>**0.27**</u>	**0.36**	<u>**0.49**</u>	0.45
SPHINX	<u>**0.35**</u>	<u>**0.64**</u>	<u>**0.61**</u>	**0.51**	<u>0.26</u>	0.33	<u>**0.49**</u>	<u>**0.46**</u>

Table 14: 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.

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 15: Language Distribution in Sphinx Dataset