Aya Dataset:
An Open-Access Collection for Multilingual Instruction Tuning

Shivalika Singh[♦1] **Freddie Vargus**[♦1] **Daniel D'souza**[♦1] **Börje F. Karlsson**[♦2]
Abinaya Mahendiran[♦1] **Wei-Yin Ko**[♦3]
Herumb Shandilya[1] **Jay Patel**[4] **Deividas Mataciunas**[1] **Laura O'Mahony**[5] **Mike Zhang**[6]
Ramith Hettiarachchi[7] **Joseph Wilson**[8] **Marina Machado**[3] **Luisa Souza Moura**[3]
Dominik Krzemiński[1] **Hakimeh Fadaei**[1] **Irem Ergun**[3] **Ifeoma Okoh**[1] **Aisha Alaagib**[1]
Oshan Ivantha Mudannayake[1] **Zaid Alyafeai**[9] **Vu Minh Chien**[1] **Sebastian Ruder**[3]
Surya Guthikonda[1] **Emad A. Alghamdi**[10] **Sebastian Gehrmann**[11] **Niklas Muennighoff**[1]
Max Bartolo[3] **Julia Kreutzer**[12] **Ahmet Üstün**[12] **Marzieh Fadaee**[12] **Sara Hooker**[12]

[1]Cohere For AI Community [2]Beijing Academy of Artificial Intelligence [3]Cohere
[4]Binghamton University [5]University of Limerick [6]Aalborg University [7]MIT
[8]University of Toronto [9]King Fahd University of Petroleum and Minerals
[10]King Abdulaziz University, ASAS.AI [11]Bloomberg LP [12]Cohere For AI

shivalikasingh95@gmail.com {marzieh, sarahooker}@cohere.com

Abstract

Datasets are foundational to many breakthroughs in modern artificial intelligence (AI). Many recent achievements in the space of natural language processing (NLP) can be attributed to the fine-tuning of pre-trained models on a diverse set of tasks that enables a large language model (LLM) to respond to instructions. Instruction fine-tuning (IFT) requires specifically constructed and annotated datasets. However, existing datasets are almost all in the English language. In this work, our primary goal is to bridge the language gap by building a human-curated instruction-following dataset spanning 65 languages. We worked with fluent speakers of languages from around the world to collect natural instances of instructions and completions. Furthermore, we create the most extensive multilingual collection to date, comprising 513 million instances through templating and augmenting existing datasets across 114 languages. In total, we contribute three key resources: we develop and open-source the **Aya** [1] **Dataset**, the **Aya Collection**, and the **Aya Evaluation Suite**. The **Aya** initiative also serves as a valuable case study in participatory research, involving collaborators from 119 countries. We see this as an important framework for future research collaborations that aim to bridge gaps in resources.

♦ First authors.

[1]The word **Aya** has its origins in the Akan (Twi) language and is translated as "fern" in English (Willis, 1998).

1 Introduction

Datasets are static representations of the world, far from the rich, ever-evolving environment we navigate as humans. Yet, these frozen snapshots in time are the foundation upon which progress in AI has been built. Many recent breakthroughs in language modeling can be attributed to fine-tuning pre-trained models on a diverse set of tasks that enable a large language model (LLM) to follow instructions (McCann et al., 2018; Sanh et al., 2022; Wei et al., 2022a; Muennighoff et al., 2023c; Longpre et al., 2023a). Instruction fine-tuning (IFT) leverages the precept that Natural Language Processing (NLP) tasks can be described via natural language instructions, such as *"What were the reviews like for the Barbie movie?"*.

The factors underlying the construction of the datasets impact how models perform for users around the world. Models perform better on the distribution they are trained to mimic (Kunchukuttan et al., 2021). This often introduces known biases towards languages and dialects not included during training and introduces critical security flaws (Yong et al., 2023a; Nasr et al., 2023; Li et al., 2023b; Lukas et al., 2023; Deng et al., 2023). In this work, our goal is to reduce this linguistic inequality. Efforts that aim to improve multilingual performance have often focused on improving data coverage (Chen et al., 2023b). However, most of the limited effort to date has focused on multilin-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11521–11567

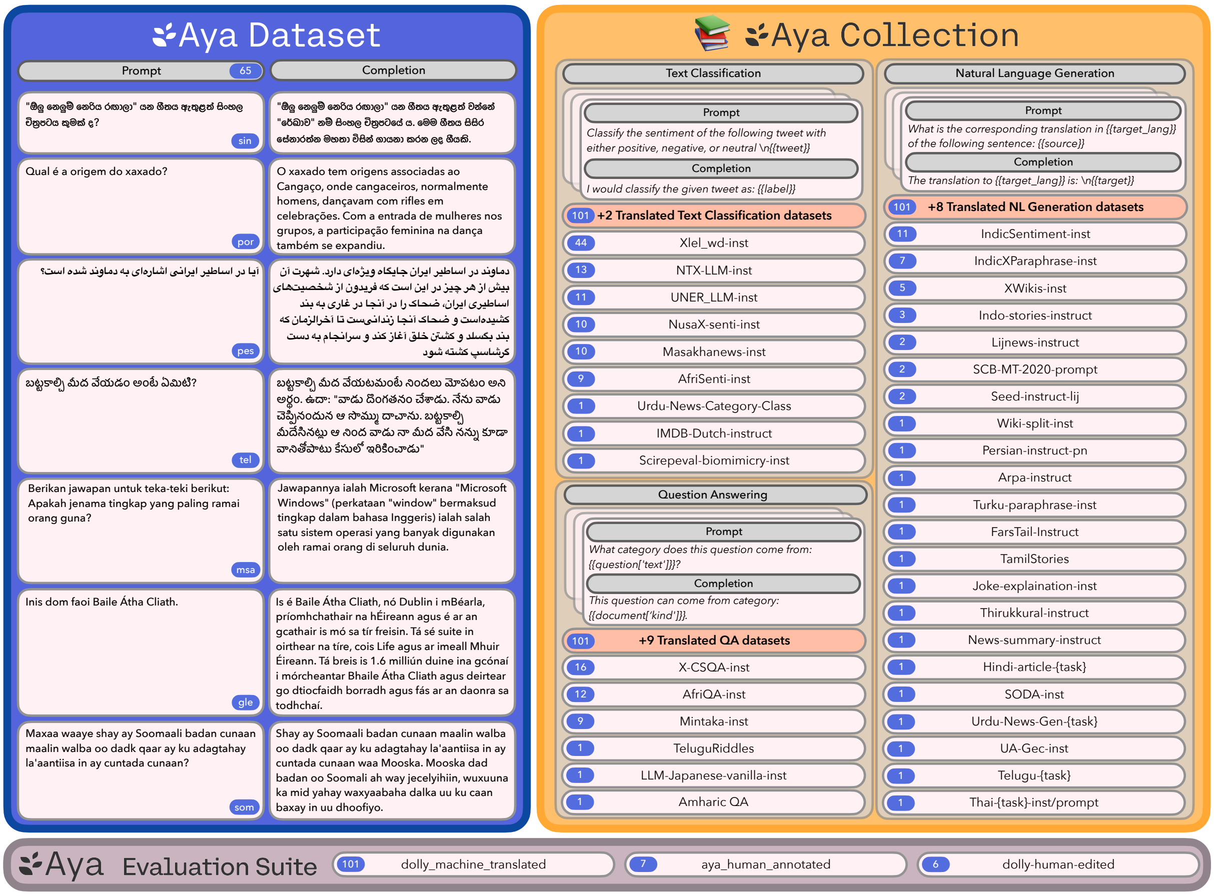

Figure 1: **Aya Dataset, Aya Collection & Aya Evaluation Suite.** On the left, we show examples of contributions in the **Aya** Dataset. These are original human-curated prompt-completion pairs written by fluent speakers of 65 languages. On the right, we have the **Aya** Collection, an aggregation of 44 monolingual and multilingual templated instruction datasets and 19 translated datasets ranging over 114 languages. The bottom block showcases the **Aya** Evaluation Suite for multilingual open-ended generation. We indicate the number of languages in a dataset with the value in the blue ovals in the figure. (Translated datasets have been visually merged due to space constraints).

gual pre-training (Scao et al., 2022a; Wei et al., 2023; Lample and Conneau, 2019; Strømberg-Derczynski et al., 2021) with even less work centered on imparting instruction following abilities.

A key aspect of our work is focused on collecting harder-to-obtain human-curated data from fluent speakers of a language. This curation process has received far less attention due to the lack of access to fluent speakers, especially in low-resource languages (Joshi et al., 2019). We set about to close this gap by conducting a year-long participatory research initiative that involved working with fluent speakers of languages from around the world to collect human-curated instances of instructions and completions. Overall, **Aya** contributes three key resources (See Figure 1):

1. **Aya Dataset**: We create the largest human-annotated multilingual instruction fine-tuning dataset to date, consisting of over 204K instances that cover 65 languages.

2. **Aya Collection**: We collect instruction-style templates from fluent speakers and applied them to a curated list of 44 datasets, including tasks such as Text Classification, Text Generation, Machine Translation, Paraphrasing, and Open-domain Question Answering. Some of these datasets also include equivalent multilingual versions produced through translation. We release 513M instances that cover 114 languages. These contributions are made available as an open-source collection.

3. **Aya Evaluation Suite**: We curate and release a diverse evaluation suite for multilingual open-ended generation. It consists of 250 human-written prompts for each of 7 languages, 200 human-selected and automatically translated prompts for 101 languages

(114 dialects), human-edited prompts for 6 languages, and the English originals. The first set represents culturally-grounded and original prompts, while the translated and post-edited prompts are sourced from English Dolly (Conover et al., 2023) and selected for their cross-cultural relevance.

By fully open-sourcing the **Aya** Dataset, **Aya** Collection and **Aya** Evaluation Suite with a permissive Apache 2.0 License[2] we hope to empower researchers and practitioners to further advance multilingual models and applications. All datasets are accessible for download.[3][4][5]

2 Aya Dataset

The goal of the **Aya** project is to facilitate annotations to a crowd-sourced dataset by individuals fluent in different languages. Inputs from speakers of each language ensure that the dataset is more likely to be organic, articulate, and representative of the speakers' cultures.

The **Aya** project was initiated to provide annotations for the 101 languages available in the mT5 model (Xue et al., 2021). Ultimately, some of these languages did not receive enough contributions to include them in the final dataset. Conversely, we received substantial contributions from languages not initially part of the original list, like Wolof, leading to their inclusion; the final **Aya** Dataset covers 65 languages. Table E.4 provides details of these languages.

2.1 Annotation Tasks

On the **Aya** Annotation Platform, contributors were able to contribute to three different tasks, following the find-fix-verify paradigm (Bernstein et al., 2015): Writing new examples from scratch (**Original Annotations**), editing existing examples to improve the quality and comprehensiveness (**Re-annotations**), and giving feedback on the quality of existing contributions (**Annotation Feedback**). We describe each briefly below:

Original Annotations. This task facilitates the inclusion of human-generated organic content by allowing annotators to submit original prompt-completion pairs in their language. Existing multilingual models have been shown to produce generations influenced by Western culture (Yuan et al., 2021; Naous et al., 2023; Lee et al., 2023) reflecting the underlying representation bias (Mehrabi et al., 2021) of their training datasets. This task aims to encourage annotators to submit fresh samples that are representative of their language, culture, literature, history, and region. The guidelines for contributors are available in Appendix C.2.

Re-Annotations. The purpose of this task is to facilitate the re-annotation or editing of prompt and completion pairs. The decision to add a re-annotation task partly stems from the need to help annotators understand the expected format of instruction-style datasets and to convey the variety of tasks in existing datasets, including question answering (Saad-Falcon et al., 2023; Arefeen et al., 2023), summarization (Stiennon et al., 2020; Wu et al., 2021), paraphrasing (Witteveen and Andrews, 2019; Reimers and Gurevych, 2019), and translation (NLLB-Team et al., 2022; Barrault et al., 2023). Editing examples from existing datasets not only helped familiarize annotators with the expected format but also allowed for human evaluation and rating of existing widely used instruction-style datasets.

In total, we collected datasets from 19 public data sources and translated them into 114 available languages, including dialects using the NLLB 3.3B parameter machine translation model (NLLB-Team et al., 2022). From each collection, we randomly chose 100 examples (per dataset, per language and per split), creating our dataset for annotation, after which we had 1M translated prompt-completion pairs initially populated in the **Aya** UI as re-annotation tasks. These translated pairs served as a starting point for prompts and completions which annotators could improve. We release the raw translations as part of the **Aya** Collection, provide more details about the provenance of the translated datasets, and how they were selected in Section 3.2. In addition to translated examples, there are other available data sources suitable for re-annotation: original **Aya** pairs, pre-existing instruction-style datasets (e.g., xP3), and the transformation of datasets into an instruction-style format, i.e., templated datasets. By re-annotating examples from different sources, we simultaneously enhance the quality of individual examples

[2] https://www.apache.org/licenses/LICENSE-2.0
[3] https://hf.co/datasets/CohereForAI/aya_dataset
[4] https://hf.co/datasets/CohereForAI/aya_collection
[5] https://hf.co/datasets/CohereForAI/aya_evaluation_suite

while obtaining a signal on the overall quality of the dataset in a specific language.

Annotation Feedback. Data quality is critical to ensure that a model can represent a language well. Learning from noisy, low-quality datasets harms the overall model performance and the relatively high cost of encoding these noisy examples is a misuse of capacity (Hsueh et al., 2009; Dodge et al., 2021; Luccioni and Viviano, 2021; Kreutzer et al., 2022). Prior work has shown that improvements to quality through data pruning or selection can have a significant impact on the downstream performance of a model (Longpre et al., 2023b; Marion et al., 2023; Boubdir et al., 2023; Yang et al., 2023). In particular, for instruction-tuning datasets, a small subset of higher-quality instructions can greatly outperform a larger volume of lower-quality instructions (AlShikh et al., 2023; Zhou et al., 2023; Chen et al., 2023a). Given these findings, ensuring high quality contributions is of paramount importance. Ensuring consistent quality is particularly challenging in an open science initiative with a large number of contributors.

2.2 Validating the quality of contributions

We follow a peer-review approach where each annotator acts as a reviewer for the other annotators working on the same language. These reviews form the basis for a quality **Aya** score which is displayed on the leaderboard in the UI. The quality score for an annotator is calculated by averaging the combined average ratings of their examples provided by other annotators who serve as reviewers. All three tasks in the **Aya** UI are connected in a sequential pipeline where submissions from “Original Annotations” are reviewed in the “Re-Annotations” task, and the re-annotations are further reviewed as part of the “Annotation Feedback” task. This systematic approach allows for a robust evaluation and enhancement of the collected data.

2.3 Criteria for Inclusion in Aya Dataset

The **Aya** Dataset includes all original annotations and a subset of all re-annotations. We only release re-annotations if there is a difference between the original and the edited version. To determine this subset, we compute the sum of edit distances **d** (Levenshtein distance (Levenshtein et al., 1966)) between the original and re-annotated prompts and completions on the character level and use an acceptance threshold of ($d \geq 5$). This ensures that we do not release duplicates of existing data.

		Count
Original Annotations		138,844
Re-Annotations	xP3 datasets	2859
	Translated datasets	7757
	Templated datasets	11013
	Original Annotations	43641
Aya Dataset Total		**204,114**

Table 1: **Aya Dataset Statistics.** We show the number of pairs of prompts and completions obtained through various annotation tasks.

Only languages with at least 50 contributions were included in the final release of **Aya** Dataset. This threshold was picked as it represents a balance between achieving a reasonable level of data quality and considering the practical limitations of human resources for some languages. The goal is to include as many languages as possible without lowering the overall quality of the dataset.

2.4 Analysis of the Aya Dataset

The **Aya** Dataset contains a total of 204,114 instances collected via the **Aya** Annotation Platform. Table 1 provides the breakdown of original annotations and re-annotations in the final dataset. The dataset covers 65 languages: 22 high-resource, 12 mid-resource, and 31 low-resource languages (see Appendix C.3 for more details on our language mappings). One objective of this project was to collect fluid original human prompts and completions. Table E.9 provides examples of prompts and completions from the **Aya** Dataset. During the data collection process, annotators were provided with examples and guidelines but were also trusted to explore their own creativity and cultural background to come up with new examples. As a result, it is meaningful to understand differences in aggregate statistics like length across datasets, language type and relationship with perceived quality.

Impact of Re-Annotation. When editing existing instances, we instructed the annotators to prioritize improving both the quality and richness of the prompts and completions. The average length of completions before and after edits are shown in Figure 2. We observe that across all data sources, the average length of completions increased after editing. On average, the length of completions after edits is 25% longer than before edits. We observed the largest increase for **Aya** original annotations

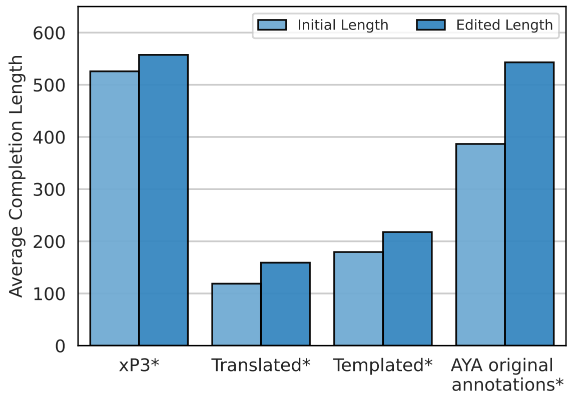

Figure 2: **Average Completion Length Before and After Re-annotation.** Here (*) indicates the subset of all dataset categories (xP3, translated, templated, and **Aya** original annotations) that were included in the **Aya** Dataset after re-annotation. Re-annotation improves average completion length across all datasets.

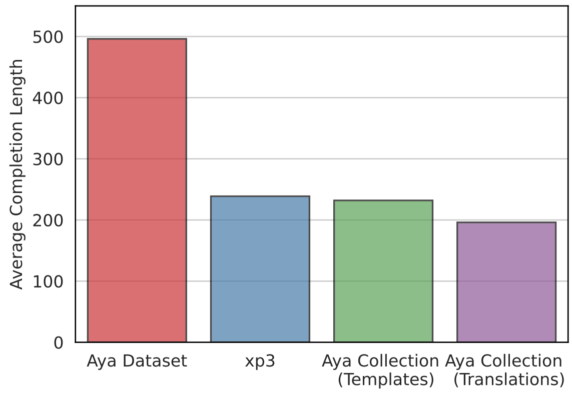

Figure 3: **Comparison of Completion Lengths.** We show the differences in completion lengths between the **Aya** Dataset, the **Aya** Collection, and xP3 (excluding the "code" split).

surfaced in the UI—which were 40% longer on average than the original length.

Length vs. Perceived Data Quality. Although longer completions can be valuable for training models to generate long and natural text, it does not necessarily imply higher quality. Using annotators' feedback in the UI, we further investigate the impact of length on the perceived quality of the samples. We observe in Figure 4 a positive correlation between how long the prompts and completions are and their resulting average approval ratio. Specifically, when we plot combined prompt and completion length against quality, we observe a correlation coefficient of 0.27. This finding emphasizes the importance of using longer prompts and completions and incorporating complete sentences to ensure a positive human experience when engaging with such a model.

Comparison in Completion length Across Datasets. The **Aya** Dataset has considerably longer completions on average when compared with other data collections as shown in Figure 3. This is particularly noteworthy given that the **Aya** Dataset is human-curated. Given the presence of longer completions in the training data for many low-resource languages, we expect that models trained on the **Aya** Dataset will generate longer and more natural responses.

3 Aya Collection

We introduce the **Aya** Collection, a comprehensive, large corpus of datasets that can be used by researchers around the world to train multilingual models. Our goal is only to include datasets with permissive licensing for manipulation and redistribution.[6] The **Aya** Collection consists of three different sources of data: ① **Templated data:** We collaborated with fluent speakers to create templates that allowed for the automatic expansion of existing datasets into various languages. ② **Translated data:** We translated a hand-selected subset of 19 datasets into 101 languages (114 dialects) using the NLLB 3.3B parameter machine translation model (NLLB-Team et al., 2022). The full list of datasets translated is listed in Table E.8. ③ **Aya Dataset:** We release the **Aya** Dataset described in Section 2 as a subset of the overall collection. It is the only dataset in the collection that is human-annotated in its entirety.

Dataset Selection Criteria. The templated and translated datasets in the **Aya** Collection were selected to achieve a mix of different task types. Our criteria prioritized datasets with high-quality natural and complete sentences, suitable for creating pairs of prompts and completions. Datasets that could potentially yield single-word answers were excluded. Finally, to create a high-quality collection, we examined all datasets and excluded those identified as unclean or noisy, primarily attributable to their automatic creation processes.

[6] https://en.wikipedia.org/wiki/Permissive_software_license

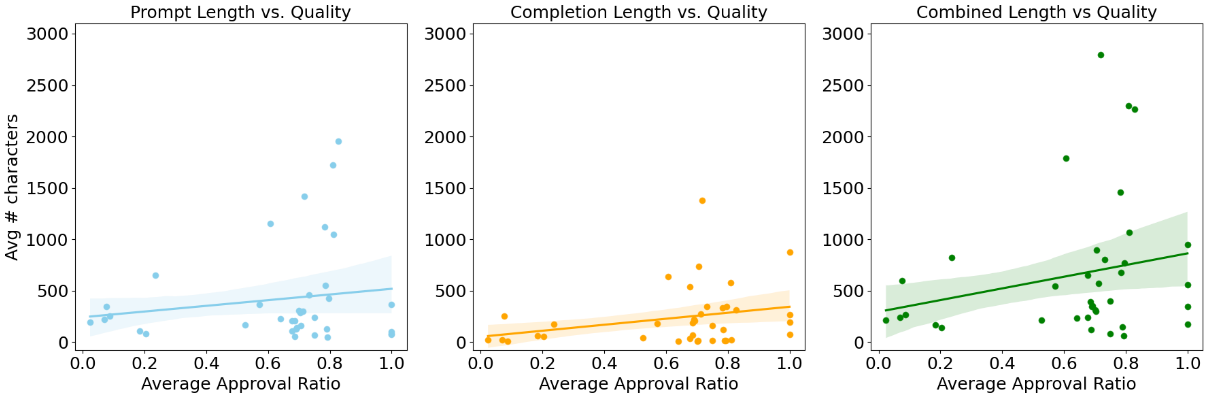

Figure 4: **Relationship between Prompt and Completion Length against Quality.** We show the trend between the average number of characters in the prompt and completion length and the average approval rate of the example.

3.1 Templating Existing Datasets

We explored the automatic expansion of existing datasets in various languages with human-written *prompt templates*, following previous work (Mishra et al., 2022; Bach et al., 2022; Wei et al., 2022a; Wang et al., 2022d). Unlike prior work that still either use English prompts in a multilingual dataset or rely on automatic translation to generate multilingual prompts, to our knowledge, the **Aya** Collection is the first broad effort to involve fluent speakers in creating prompts in their language to expand existing datasets for IFT. We used the PromptSource framework (Bach et al., 2022) to template these datasets. The **Aya** community members submitted instructions and created templates for datasets in the languages they were proficient in. Our process includes: ① Templating datasets with instructions in the same language as the original dataset; ② if the dataset is not in English, we create instructions in English. Our input prompts can be monolingual or code-mixed, depending on whether we apply templates in the same language or in English to the dataset of a particular language. Note that code-mixed input prompts here refer to a *structured* mix of English instructions with non-English monolingual data (Lin et al., 2022), which is different from the typical sociolinguistic definition of code-mixing (or code-switching) of languages in natural conversational utterances (Srivastava and Singh, 2021; Winata et al., 2023a; Yong et al., 2023c; Doğruöz et al., 2023). We used the suggested templates to convert each dataset into an instruction-style format. We release these datasets under the **Aya** Collection. We list the details of all datasets we apply templates to in Table E.7.

3.2 Automatic Translation

Despite the potential drawbacks of having lower quality than naturally found data, training models with translated data can yield significant benefits (Aharoni et al., 2019; Zhang et al., 2018; Tang et al., 2021). We thus experimented with improving coverage of low-resource languages by selectively translating high-quality datasets from various existing collections.

Setup. We chose 19 high-quality IFT datasets from xP3 (Muennighoff et al., 2023c), the Flan Collection (Longpre et al., 2023a), Dolly (Conover et al., 2023), along with additional sources such as SODA (Kim et al., 2022) and Mintaka (Sen et al., 2022). Datasets were prioritized for translation based on the richness of task diversity and length of completions. The complete list of these datasets is given in Table E.8. We process datasets for translation using the No Language Left Behind (NLLB 3.3B; NLLB-Team et al., 2022) machine translation model, which is capable of single-sentence translations between 200 different languages and dialects in various scripts. We open-source all translations as part of the **Aya** Collection.

4 Analysis of Aya Collection

Overview. The **Aya** Collection consists of existing NLP datasets that are templated to include instructions as well as datasets already in instruction format submitted by the **Aya** community. Table E.7 in the Appendix shows the detailed list of datasets. The final **Aya** Collection consists of 44 multilingual and non-English templated and 19 translated datasets, with 513M individual instances. Overall, the collection covers 114 languages[7].

Language Balance. One of the objectives of templating (and translating) existing datasets is to broaden the available resources for languages that have limited digital data. To examine if our final collection adheres to a similar distribution pattern, we use the number of Wikipedia pages in each language as a proxy for the online presence of its fluent speakers. Figure 5 showcases that although the number of instances for languages varies in the **Aya** Collection (templated subset), it does not disadvantage languages with fewer Wikipedia pages. The distribution still ensures a reasonable coverage across all languages. It is imperative to emphasize that our analysis does not involve a direct comparison of absolute values, given the disparate units of measurement involved. Instead, we examine the *patterns* of data scarcity for various languages in our collection versus Wikipedia. Including the translated datasets in the **Aya** Collection further reduces disparities between languages and contributes to creating a more balanced collection as seen in Figure F.8 in the Appendix.

5 Aya Evaluation Suite

Lastly, we release an evaluation suite tailored to multilingual models. This set aims to tackle the scarcity of multilingual data, a challenge that becomes even more apparent when considering evaluation sets. While there are several test sets available for evaluating multilingual models (Conneau et al., 2018; Hu et al., 2020; Ponti et al., 2020; Lin et al., 2022; Leong et al., 2023; Ruder et al., 2023), they focus primarily on discriminative tasks or on regional subgroups of languages. To evaluate multilingual models' generations, the literature includes task-specific evaluation sets such as Translation (Goyal et al., 2021b), Summarization (Hasan et al., 2021) and Question Answering (Clark et al., 2020), as well as combinations thereof (Gehrmann et al., 2022). However, there exists a gap in evaluating *open-ended generation* capabilities of LLMs within a multilingual context. We aim to address this gap by curating a multilingual evaluation set tailored for assessing the open-ended generation capabilities of LLMs, such as brainstorming, planning, and other unstructured, long-form responses.

To strike a balance between language coverage and the quality that comes with human attention, we create an evaluation suite that includes **(1)** human-curated examples in a limited set of languages, **(2)** automatic translations of handpicked examples into a more extensive number of languages, and **(3)** human-post-edited translations into a small number of languages. We consider two primary sources of data: original annotations from **Aya** dataset (comprising new examples culturally curated for different languages) and Dolly prompts (high-quality, human-written examples carefully selected to have a universal reach). The subsets comprising the **Aya** evaluation suite are:

AYA-HUMAN-ANNOTATED Test Set. We partitioned the **Aya** dataset into train and test splits. The test set of the **Aya** Dataset contains 1,750 of the total instances (250 instances from 7 languages), randomly selected from original annotations. To guarantee enough data remains for training, we focused on languages with at least 2000 original annotations. To ensure linguistic diversity, we included languages that were varied in terms of resourcedness, script and language family. For these reasons, the test set consists of English (high-resource, Latin script, Indo-European), Portuguese (high-resource, Latin script, Indo-European), Simplified Chinese (high-resource, Han, Sino-Tibetan), Standard Arabic (high-resource, Arabic script, Afro-Asiatic), Telugu (low-resource, Telugu script, Dravidian), Turkish (high-resource, Latin script, Turkic), and Yoruba (low-resource, Latin script, Atlantic-Congo).

DOLLY-MACHINE-TRANSLATED Test Set. We curate a subset of 200 Dolly prompts (Conover et al., 2023) to serve as an additional translated evaluation set. Our aim with this selection was to exclude any culturally or geographically specific prompts and completions. Hence, two reviewers initially inspected a set of 500 English prompts that were uniformly sampled based on the task categories in Dolly. The reviewers excluded prompts that rely on prompts such as *"Why is NFL football called football when players use their hands mainly?"* that rely on overly specific cultural references. When two reviewers disagreed, a third reviewer was asked to break the tie. This selection aimed to gather a test set that allows us to evaluate the fluency and quality of responses in various languages while avoiding model assessment

[7] We release the **Aya** Dataset as part of the **Aya** Collection, bringing the total number of languages in the collection to 115. However, for the sake of clarity, when referencing the **Aya** Collection statistics in this paper, we exclude the **Aya** Dataset.

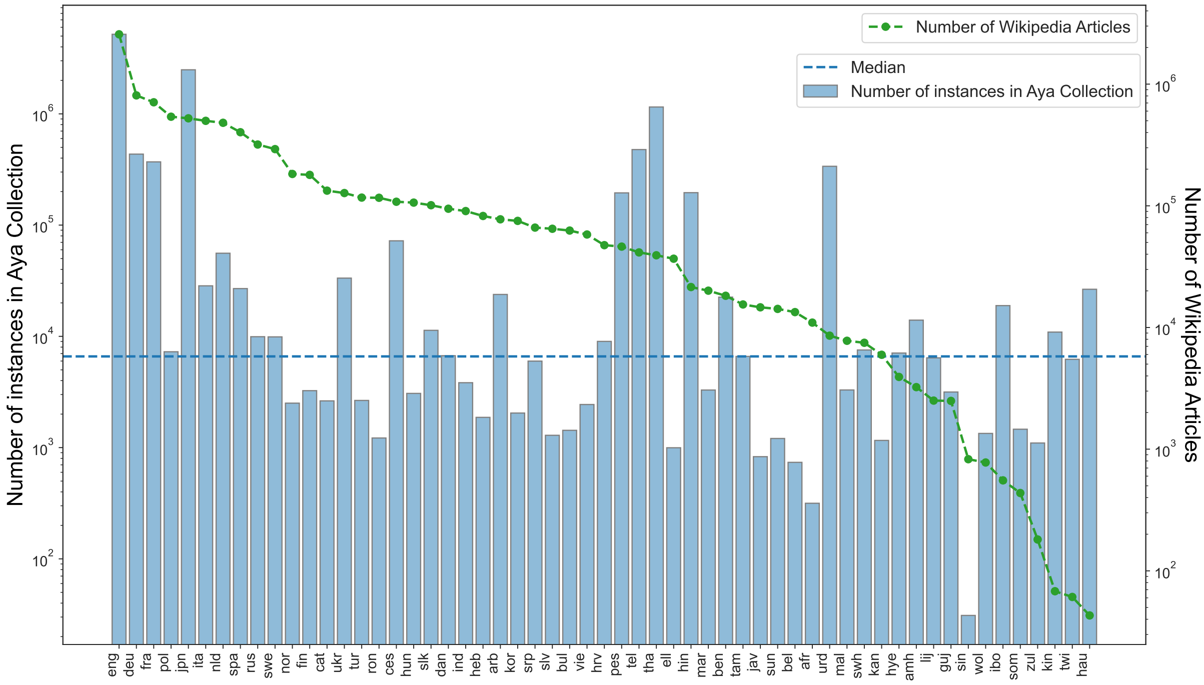

Figure 5: **Number of Prompt–Completion Pairs in Each Language** in the **Aya** Collection (templated). Many languages with limited digital presence, as indicated by a low number of Wikipedia pages, are well-represented in the templated portion of the **Aya** Collection. Note that both axes are in log-scale.

on prompts tied to specific cultural or geographic references that might have language-dependent validity. We automatically translate the prompts with NLLB into 101 languages and their dialects that are captured by NLLB. Including the original English prompts this dataset covers 115 dialects.

DOLLY-HUMAN-EDITED Test Set. The automatic translation process may introduce errors in the prompts that render them nonsensical. To enhance the reliability of testing on these prompts, we therefore enlist professional human annotators to post-edit the examples (e.g. for the example above *"Alburno o Cansado"* (=*"[Fish name] or Tired"*). We post-edit the prompts for a subset of six languages: Arabic, Hindi, Spanish, French, Serbian and Russian. Appendix E.1 describes the post-editing process and effort in more detail. The example above illustrates that some prompts, even when translated correctly, might still not transfer well into other languages—which is the main difference between a translated English-centric set like this and an evaluation set originally written in each target language like AYA-HUMAN-ANNOTATED.

We open-source the **DOLLY-MACHINE-TRANSLATED Test Set** to be an additional resource for researchers, although warn that the expressiveness of a translated evaluation set is limited by the quality of the translation model (and human post-editing) and may adversely impact an estimate of ability in languages where translations are inadequate (Nogara et al., 2023). Ultimately, this is a compromise between having evaluation coverage in a more complete set of languages (101 languages and 114 dialects in total) versus having human-annotated evaluation sets. We recommend pairing evaluation on the automatically DOLLY-MACHINE-TRANSLATED test set with evaluation on the professionally post-edited DOLLY-HUMAN-EDITED for 6 languages, or the AYA-HUMAN-ANNOTATED test set created by proficient speakers in 7 languages. We additionally recommend using human evaluation strategies to assess generated outputs on this evaluation suite. Automatic metrics underperform in creative tasks and non-English outputs, making them unsuited for this application (Gehrmann et al., 2023).

6 Conclusion

Open participatory research continues to be under-resourced and undervalued, particularly when that work focuses on data creation (Sambasivan et al., 2021). **Aya** involved participants from many different countries, different ages, and different levels of familiarity with the field of natural language processing. We see continued opportunities for computational linguists and machine-learning engineers to collaborate with social scientists such as sociolinguists, anthropologists, sociologists, and media studies scholars. As new norms in open science emerge (Krishna, 2020; Bowser et al., 2020), collaborations like these can help ensure that projects in NLP are motivated by an understanding of what language means to the people who use it every day. With **Aya** we hope to change the way data is created for multilingual NLP research. In line with this view, we release the **Aya** Dataset which is the first human-curated open-source, multilingual instruction-style dataset consisting of 204,114 prompt-completion pairs covering 65 languages. This dataset was built with the help of our open-science community of 2,997 collaborators from 119 countries over a period of eight months.

We also release the **Aya** Collection, which consists of 44 instruction-style datasets. These were prepared by transforming existing NLP datasets into prompt-completion pairs that can be leveraged for instruction tuning. Furthermore, we translate several high-quality datasets into 101 languages, thereby expanding coverage, particularly for many low-resource languages. This collection consists of 513M prompt and completion pairs covering 114 languages in total and is the largest multilingual instruction fine-tuning collection today. Additionally, we release **Aya** Evaluation Suite, consisting of human-curated examples in 13 languages and translation of carefully selected prompts in 101 languages.

7 Limitations

Language and Dialect Coverage. The **Aya** Dataset and **Aya** Collection cover 65 and 114 languages respectively—significantly more than existing multilingual datasets. However, this is still only a tiny fraction of the world's linguistic diversity. Of the world's approximately 7,000 languages, only half of them are captured in any sort of written form (Adda et al., 2016). Of this half, only a few hundred are included on the internet in machine readable corpora (Adda et al., 2016). This means that 93% of the world's languages are still not being used to train LLMs. It is also notoriously difficult to determine the dividing line between different languages and different dialects of the same language (Rooy, 2021). Geo-cultural variation within a language often gives rise to new dialects or creole languages over time (Zampieri et al., 2020; Wolfram, 1997; Brown et al., 2020; Lent et al., 2022; Blaschke et al., 2023) and, as such, dialects can serve an important function in establishing and maintaining cultural identity (Falck et al., 2012). Many different dialects that are generally recognized as belonging to a single parent language are not represented in the dataset. For example, in the case of Malay, one of the largest Southeast Asian languages in the dataset, there are no contributions for regional dialects that are widely spoken in certain parts of Malaysia. Contributions by volunteers who wished to self-identify as speaking a particular dialect were tagged as such in the data to allow for limited analysis of the use of regional dialects in annotations. Lastly, socio-linguistic data show that multilingual speakers often 'code-switch' between languages or dialects depending on context (Myers-Scotton, 2017), but in this project, we kept the languages isolated to make them easier to classify and to be used downstream for language-specific applications. The current project setup is not able to serve languages without a written tradition.

Imbalanced Distribution of Contribution. As explored in Appendix C.5, despite the large number of participants, the activity of annotators was skewed, with a 'long tail' of annotators only contributing one or two annotations. Relatively few contributors accounted for the most annotations. Similarly, there is a huge gap between languages with the highest number of contributions and ones with the lowest number of contributions. Consequently, this suggests a potential imbalance in dataset distributions across different languages and a lack of annotator diversity within some languages dominated by one or two prolific contributors.

Cultural or Personal Bias. Some languages in our dataset have limited representation, with only a few annotators responsible for annotating the bulk of their data. This might mean that data for a particular language is dominated by annotations that represent the opinions and perspectives of a particular contributor or a narrow selection of cultural viewpoints. For example, annotations in French might contain many examples about the history of France, its food, songs, and other cultural practices, but not much information about the cultural heritage of French-speaking communities in Québec, Togo, or Senegal (Vigouroux, 2013). This bias is particularly problematic given the skewed distribution of the most active annotators. There is also a potential bias in the availability of particular kinds of content. For example, it is easier to find online text from news sites for many African languages than it is to find text from other domains. Accordingly, these datasets will be skewed towards the grammar and lexicon used in news reports instead of the kind of natural language people use in everyday life (Hovy and Prabhumoye, 2021).

Toxic or Offensive Speech. The **Aya** Annotation Platform did not contain specific flags for toxic, harmful, or offensive speech, so it is possible that malicious users could submit unsafe data. We believe this is of relatively low risk because of the high rate of human-verified annotations and peer-review, making it unlikely that toxic prompts or completions made it into the final dataset. However, there is no guarantee that every entry was audited. While data poisoning has rarely been observed as a viable threat in practice, it has been demonstrated to be of concern for instruction-tuning with very few examples (Xu et al., 2023b; Wan et al., 2023) and for pre-training under realistic conditions (Carlini et al., 2023). During the eight months of crowd-sourced annotating, there were no reported cases of hateful or toxic speech in the existing datasets nor were there any instances of offensive speech reported in the peer-reviewing phase of new annotations.

Ethics Statement

This work was carried out as an open science initiative by volunteer participants. All datasets used in this work have permissive licensing. We publicly release the datasets under Apache 2.0 license.

Acknowledgements

We thank David Adelani, Sasha Luccioni, Rosanne Liu, and Luciana Benotti for helpful feedback on earlier versions of this draft.

Language Ambassadors. We thank the Language Ambassadors for spreading the word about Aya to speakers of their language, recruiting new contributors and supporting them to understand the goals of Aya project (listed alphabetically): Diana Abagyan, Muhammad Abdullahi, Elyanah Aco, Henok Ademtew, Adil, Emad A. Alghamdi, Zaid Alyafeai, Ahmad Anis, Daniel Avila, Michael Bayron, Nathanael Carraz Rakotonirina, Alberto Mario Ceballos Arroyo, Yi Yi Chan Myae Win Shein, Vu Minh Chien, Ionescu Cristian, Ripal Darji, Suchandra Datta, Rokhaya Diagne, Irem Ergün, Hakimeh (Shafagh) Fadaei, Surya Krishna Guthikonda, Aleksandra Hadžić, Shamsuddeen Hassan Muhammad, Ramith Hettiarachchi, Mochamad Wahyu Hidayat, Rin Intachuen, Eldho Ittan George, Ganesh Jagadeesan, Murat Jumashev, Börje F. Karlsson, Abhinav Kashyap, JiWoo Kim, Alkis Koudounas, Kevin Kudakwashe Murera, Falalu Ibrahim Lawan, Wen-Ding Li, Abinaya Mahendiran, Mouhamadane MBOUP, Oleksander Medyuk, Iftitahu Nimah, Solam Nyangiwe, Laura O'Mahony, Ifeoma Okoh, Hui-Lee Ooi, Inigo Parra, Jay Patel, Hanif Rahman, Olanrewaju Samuel, Suman Sapkota, Giacomo Sarchioni, Rashik Shrestha, Bhavdeep Singh Sachdeva, Sean Andrew Thawe, Alperen ÜNLÜ, Joseph Wilson, Emilia Wiśnios, Yang Xu, Zheng-Xin Yong, and Mike Zhang.

Dataset Contributors. We thank the Aya community members who contributed instruction-style datasets in their languages to be included in the **Aya** collection (listed alphabetically): Diana Abagyan, Henok Ademtew, Ahmad Anis, Hakimeh (Shafagh) Fadaei, Hamidreza Ghader, Md. Tahmid Hossain, Eldho Ittan George, Ganesh Jagadeesan, Börje F. Karlsson, Surya Krishna Guthikonda, Abinaya Mahendiran, Desik Mandava, Iftitahu Nimah, Wannaphong Phatthiyaphaibun, Mike Zhang.

Aya Quality Champions. We thank the Aya community members who contributed high-quality data for the **Aya** Dataset (listed in ascending order based on Aya Score): Vu Minh Chien, Hui-Lee Ooi, Gamage Omega Ishendra, Surya Krishna Guthikonda, Hoang Anh Quynh Nhu, Moses Oyeleye, Amarjit Singh Sachdeva, Mike Zhang, Almazbekov Bekmyrza Ruslanovich, Ramla Abdullahi Mohamed, Börje F. Karlsson, Regina Sahani Lourdes De Silva Goonetilleke, Zaid Alyafeai, Yong Zheng Xin, Yavuz Alp Sencer Öztürk, Mohammed Hamdy, Anitha Ranganathan, Ramith Hettiarachchi, Ooi Hui Yin, Caroline Shamiso Chitongo, Bhavdeep Singh Sachdeva, Valentyn Bezshapkin, Yang Xu, Dominik Krzeminski, Iftitahu Nimah, Muna Mohamed Abdinur, Nurbaeva Zhiidegul Talaibekovna, Younes Bensassi Nour, Eldho Ittan George, Caio Dallaqua, Hakimeh (Shafagh) Fadaei, Henok Ademtew, Vijayalakshmi

Varadharajan, Yogesh Haribhau Kulkarni, Laura O'Mahony, Jay Patel, Luísa Souza Moura, Rama Hasiba, Geoh Zie Ee, Gabriela Vilela Heimer, Pratham Prafulbhai Savaliya, Deividas Mataciunas, Ifeoma Okoh, Alberto Mario Ceballos Arroyo, Basiiru Silla, Yiorgos Tsalikidis.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *CMLC 2021-9th Workshop on Challenges in the Management of Large Corpora*, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Tilahun Abedissa, Ricardo Usbeck, and Yaregal Assabie. 2023. AmQA: Amharic Question Answering Dataset. *arXiv preprint arXiv:2303.03290*.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News Topic Classification for African Languages.

Asif Agha. 2006. *Language and Social Relations*. Studies in the Social and Cultural Foundations of Language. Cambridge University Press.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation.

AhmadMustafa. 2023a. Urdu-Instruct-News-Article-Generation. https://huggingface.co/datasets/AhmadMustafa/Urdu-Instruct-News-Article-Generation. Accessed: 2023-11-28.

AhmadMustafa. 2023b. Urdu-Instruct-News-Category-Classification. https://huggingface.co/datasets/AhmadMustafa/Urdu-Instruct-News-Category-Classification. Accessed: 2023-11-28.

AhmadMustafa. 2023c. Urdu-Instruct-News-Headline-Generation. https://huggingface.co/datasets/AhmadMustafa/Urdu-Instruct-News-Headline-Generation. Accessed: 2023-11-28.

AI Tamil Nadu. 2023a. Tamil stories. https://huggingface.co/datasets/aitamilnadu/tamil_stories. Accessed: 2023-12-15.

AI Tamil Nadu. 2023b. Thirukkural Instruct. https://huggingface.co/datasets/aitamilnadu/thirukkural_instruct. Accessed: 2023-11-30.

Christopher Akiki, Giada Pistilli, Margot Mieskes, Matthias Gallé, Thomas Wolf, Suzana Ilić, and Yacine Jernite. 2022. Bigscience: A case study in the social construction of a multilingual large language model.

Waseem AlShikh, Manhal Daaboul, Kirk Goddard, Brock Imel, Kiran Kamble, Parikshith Kulkarni, and Melisa Russak. 2023. Becoming self-instruct: introducing early stopping criteria for minimal instruct tuning. *arXiv preprint arXiv:2307.03692*.

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023. Farstail: A Persian natural language inference dataset. *Soft Computing*.

Lauri Andress, Tristen Hall, Sheila Davis, Judith Levine, Kimberly Cripps, and Dominique Guinn. 2020. Addressing power dynamics in community-engaged research partnerships. *Journal of Patient-Reported Outcomes 4: 24*.

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2023. Leancontext: Cost-efficient domain-specific question answering using llms. *arXiv preprint arXiv:2309.00841*.

Kailash Awati and Simon Buckingham Shum. 2015. Big data metaphors we live by. *Towards Data Science*.

Paul Azunre, Salomey Osei, Salomey Afua Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Adowaa Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac. K. E. Ampomah, Joseph Otoo., Reindorf Nartey Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. Nlp for ghanaian languages. *ArXiv*, abs/2103.15475.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM*, 58(8):85–94.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. ACM.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Meriem Boubdir, Edward Kim, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2023. Which prompts make the difference? data prioritization for efficient human llm evaluation.

Anne Bowser, Caren Cooper, Alex De Sherbinin, Andrea Wiggins, Peter Brenton, Tyng-Ruey Chuang, Elaine Faustman, Mordechai Haklay, and Metis Meloche. 2020. Still in need of norms: the state of the data in citizen science. *Citizen Science: Theory and Practice*, 5(1).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, et al. 2023. Nusacrowd: Open source initiative for indonesian nlp resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum

Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical.

Joe Castaldo. 2023. “AI chatbots fall short in dozens of languages. A non-profit project aims to fix that”. *Globe and Mail*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023a. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023b. Monolingual or multilingual instruction tuning: Which makes a better alpaca.

Sanxing Chen, Yongqiang Chen, and Börje F. Karlsson. 2023c. Dataset and baseline system for multi-lingual extraction and normalization of temporal and numerical expressions. *arXiv preprint arXiv:2303.18103*.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. 2018. Deep learning for classical japanese literature. *ArXiv*, abs/1812.01718.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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.

ConseggioLigure. 2023a. Lij news instruct ita-lij. https://huggingface.co/datasets/ConseggioLigure/lijnews-instruct-ita-lij. Accessed: 2023-11-28.

ConseggioLigure. 2023b. Lij news instruct lij-ita. https://huggingface.co/datasets/ConseggioLigure/lijnews-instruct-lij-ita. Accessed: 2023-11-28.

ConseggioLigure. 2023c. Seed instruct eng-lij. https://huggingface.co/datasets/ConseggioLigure/seed-instruct-eng-lij. Accessed: 2023-11-28.

ConseggioLigure. 2023d. Seed instruct lij-eng. https://huggingface.co/datasets/ConseggioLigure/seed-instruct-lij-eng. Accessed: 2023-11-28.

Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and public participation in ai. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO ’23, New York, NY, USA. Association for Computing Machinery.

Kate Crawford. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven, Connecticut.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *ArXiv*, abs/2310.06474.

desik98. 2023. Telugu Riddles. https://huggingface.co/datasets/desik98/TeluguRiddles. Accessed: 2023-11-30.

Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Shrivastava, Samson Tan, et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the english colossal clean crawled corpus. *CoRR*, abs/2104.08758.

A Seza Doğruöz, Sunayana Sitaram, and Zheng-Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. *arXiv preprint arXiv:2310.20470*.

Oliver Falck, Stephan Heblich, Alfred Lameli, and Jens Südekum. 2012. Dialects, cultural identity, and economic exchange. *Journal of urban economics*, 72(2-3):225–239.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav

Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–6. IEEE.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

ganeshjcs. 2023a. Hindi Article Summarization. https://huggingface.co/datasets/ganeshjcs/hindi-article-summarization. Accessed: 2023-11-28.

ganeshjcs. 2023b. Hindi Headline Article Generation. https://huggingface.co/datasets/ganeshjcs/hindi-headline-article-generation. Accessed: 2023-11-28.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, and Bernd Bohnet. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Eldho Ittan George. 2023a. Aya Indic Sentiment. https://huggingface.co/datasets/el2e10/aya-indicsentiment. Accessed: 2023-11-28.

Eldho Ittan George. 2023b. Aya Paraphrase. https://huggingface.co/datasets/el2e10/aya-paraphrase. Accessed: 2023-11-28.

Charles Goodwin. 2017. *Co-Operative Action*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021a. Larger-Scale Transformers for Multilingual Masked Language Modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021b. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. *arXiv preprint*.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for low-resource text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6504–6513, Marseille, France. European Language Resources Association.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2023. Targen: Targeted data generation with large language models.

Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. 2020. An ethical highlighter for people-centric dataset creation. *arXiv preprint arXiv:2011.13583*.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

hghader1. 2023. FarsTail-Instruct-LLM. https://huggingface.co/datasets/hghader1/FarsTail-Instruct-LLM. Accessed: 2023-11-28.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35.

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

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.

Kuan-Hao Huang, I-Hung Hsu, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction.

Khalid Hussain, Nimra Mughal, Irfan Ali, Saif Hassan, and Sher Muhammad Daudpota. 2021. Urdu News Dataset 1M. Technical report, Mendeley Data, V3.

Mika Hämäläinen. 2021. *Endangered Languages are not Low-Resourced!*, page 1–11. University of Helsinki.

Iftitahu. 2023a. Indonesian Instruct Stories. https://huggingface.co/datasets/Iftitahu/indonesian_instruct_stories. Accessed: 2023-11-28.

Iftitahu. 2023b. Javanese Instruct Stories. https://huggingface.co/datasets/Iftitahu/javanese_instruct_stories. Accessed: 2023-11-28.

Iftitahu. 2023c. Sudanese Instruct Stories. https://huggingface.co/datasets/Iftitahu/sundanese_instruct_stories. Accessed: 2023-11-28.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

jjzha. 2023. IMDB Dutch Instruct. https://huggingface.co/datasets/jjzha/imdb-dutch-instruct. Accessed: 2023-11-28.

Joseph Cheung. 2023. GuanacoDataset (Revision 8cf0d29) .

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 288–298, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Venni V. Krishna. 2020. Open science and its enemies: Challenges for a sustainable science–society social contract. *Journal of Open Innovation: Technology, Market, and Complexity*, 6(3):61.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Anoop Kunchukuttan, Siddharth Jain, and Rahul Kejriwal. 2021. A large-scale evaluation of neural machine transliteration for Indic languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3469–3475, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals.

In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023b. Privacy in large language models: Attacks, defenses and future directions. *ArXiv*, abs/2310.10383.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

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.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. 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. 2023a. 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, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023b. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Yanni Alexander Loukissas. 2019. *All Data Are Local: Thinking Critically in a Data-Driven Society*. MIT Press, Cambridge, Massachusetts.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T Rutherford, and Sarana Nutanong. 2022. A large English–Thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation*, 56(2):477–499.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. *International Conference on Language Resources and Evaluation*.

Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. 2023. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. Fingpt: Large generative models for a small language. *arXiv preprint arXiv:2311.05640*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2023. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. *arXiv preprint arXiv:2311.09122*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. 2023a. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023b. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

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. 2023c. 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.

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Carol Myers-Scotton. 2017. Code-switching. *The handbook of sociolinguistics*, pages 217–237.

Gabriel Nakamura, Bruno Soares, Valério Pillar, José Diniz-Filho, and Leandro Duarte. 2023. Three pathways to better recognize the expertise of global south researchers. *npj Biodiversity*.

Tarek Naous, Michael Joseph Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *ArXiv*, abs/2305.14456.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

Huu Nguyen, Sameer Suri, Ken Tsui, and Christoph Schuhmann. 2023. The open instruction generalist (oig) dataset. https://laion.ai/blog/oig-dataset/.

Jinjie Ni, Fuzhao Xue, Kabir Jain, Mahir Hitesh Shah, Zangwei Zheng, and Yang You. 2023. Instruction in the wild: A user-based instruction dataset. https://github.com/XueFuzhao/InstructionWild.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang.

2022. No language left behind: Scaling human-centered machine translation.

Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2023. Toxic bias: Perspective api misreads german as more toxic.

Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

osyvokon. 2023. UA-GEC instruction tuning . https://huggingface.co/datasets/osyvokon/ua_gec_instruction_tuning. Accessed: 2023-11-28.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context.

Michael Park, Erin Leahey, and Russell J. Funk. 2023. Papers and patents are becoming less disruptive over time. *Nature*, 613:138–144.

Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922*.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of The 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic.

Martin Petty. 2023. Explainer: Why is Myanmar's military holding an election? Accessed on Jan. 17, 2024.

McKevitt C Pinel C, Prainsack B. 2020. Caring for data: Value creation in a data-intensive research laboratory. *Social Studies of Science*, 50(2):175–197.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. 2022. Multilingual event linking to Wikidata. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 37–58, Seattle, USA. Association for Computational Linguistics.

Cornelius Puschmann and Jean Burgess. 2014. Big data, big questions| metaphors of big data. *International Journal of Communication*, 8(0).

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA. Association for Computing Machinery.

PyThaiNLP. 2023a. scb_mt_2020_en2th_prompt. https://huggingface.co/datasets/pythainlp/scb_mt_2020_en2th_prompt. Accessed: 2023-11-29.

PyThaiNLP. 2023b. scb_mt_2020_th2en_prompt. https://huggingface.co/datasets/pythainlp/scb_mt_2020_th2en_prompt. Accessed: 2023-11-29.

PyThaiNLP. 2023c. Thai-Pos-prompt. https://huggingface.co/datasets/pythainlp/Thai-Pos-prompt. Accessed: 2023-11-29.

PyThaiNLP. 2023d. thai-wiktionary-prompt. https://huggingface.co/datasets/pythainlp/thai-wiktionary-prompt. Accessed: 2023-11-29.

PyThaiNLP. 2023e. thai_usembassy_en2th_prompt. https://huggingface.co/datasets/pythainlp/thai_usembassy_en2th_prompt. Accessed: 2023-11-29.

PyThaiNLP. 2023f. thai_usembassy_th2en_prompt. https://huggingface.co/datasets/pythainlp/thai_usembassy_th2en_prompt. Accessed: 2023-11-29.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reuters. 2023. Explainer: What is happening between Armenia and Azerbaijan over Nagorno-Karabakh? Accessed on Jan. 17, 2024.

Raf Van Rooy. 2021. *Language or Dialect? The History of a Conceptual Pair*. Oxford University Press.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. 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*.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A Rossi, and Franck Dernoncourt. 2023. Pdftriage: Question answering over long, structured documents. *arXiv preprint arXiv:2309.08872*.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: Lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '12, New York, NY, USA. Association for Computing Machinery.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

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

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. 2022b. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*.

Nick Seaver. 2021. Care and scale: Decorrelative ethics in algorithmic recommendation. *Cultural Anthropology*, 36(3):509–537.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shafagh. 2023a. Aya Persian Instruction pn Summary. https://huggingface.co/datasets/Shafagh/aya_persian_instruction_pn-summary. Accessed: 2023-11-28.

Shafagh. 2023b. Aya Persian Instruction pn Summary Title. https://huggingface.co/datasets/Shafagh/aya_persian_instruction_pn-summary-title. Accessed: 2023-11-28.

Jack Sidnell and N. J. Enfield. 2012. Language diversity and social action: A third locus of linguistic relativity. *Current Anthropology*, 53(3):302–333.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Z. Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David Ifeoluwa Adelani, Amelia V. Taylor, Jamiil Toure Ali, Kevin Degila, Momboladji Balogoun, Thierno Ibrahima Diop, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. 2021. AI4D - african language program. In *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. *ArXiv*, abs/2211.13308.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Pete Walsh, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 33–43, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Bernard Spolsky. 2018. Language policy in french colonies and after independence. *Current Issues in Language Planning*, 19(3):231–315.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Vivek Srivastava and Mayank Singh. 2021. Challenges and considerations with code-mixed nlp for multilingual societies.

Luke Stark and Anna Lauren Hoffmann. 2019. Data is the new what? popular metaphors & professional ethics in emerging data culture. *Journal of Cultural Analytics*, 4(1).

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

SuryaKrishna02. 2023a. Aya Telugu Food Recipes. https://huggingface.co/datasets/SuryaKrishna02/aya-telugu-food-recipes. Accessed: 2023-11-28.

SuryaKrishna02. 2023b. Aya Telugu Jokes. https://huggingface.co/datasets/SuryaKrishna02/aya-telugu-jokes. Accessed: 2023-11-28.

SuryaKrishna02. 2023c. Aya Telugu News Articles. https://huggingface.co/datasets/SuryaKrishna02/aya-telugu-news-articles. Accessed: 2023-11-28.

SuryaKrishna02. 2023d. Aya Telugu Paraphrase. https://huggingface.co/datasets/SuryaKrishna02/aya-telugu-paraphrase. Accessed: 2023-11-28.

SuryaKrishna02. 2023e. Aya Telugu Poems. https://huggingface.co/datasets/SuryaKrishna02/aya-telugu-poems. Accessed: 2023-11-28.

Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. 2023. From base to conversational: Japanese instruction dataset and tuning large language models. *arXiv preprint arXiv:2309.03412*.

syntaxshill. 2023. Arpa aya. https://huggingface.co/datasets/syntaxshill/arpa-aya. Accessed: 2023-11-28.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

TahmidH. 2023. Annotated News Summary. https://huggingface.co/datasets/TahmidH/annotated_news_summary. Accessed: 2023-11-28.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. *Findings of the Association for Computational Linguistics: ACL-IJCNLP*.

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.

Tellarin.ai. 2023a. LLM Japanese Dataset Vanilla Aya Format. https://huggingface.co/datasets/tellarin-ai/llm-japanese-dataset-vanilla-aya-format. Accessed: 2023-11-28.

Tellarin.ai. 2023b. NTX LLM Instructions. https://huggingface.co/datasets/tellarin-ai/ntx_llm_instructions. Accessed: 2023-11-28.

theblackcat102. 2023. Joke explaination. https://huggingface.co/datasets/theblackcat102/joke_explaination. Accessed: 2023-11-29.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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

TurkuNLP. 2023. Turku Paraphrase Corpus. https://huggingface.co/datasets/TurkuNLP/turku_paraphrase_corpus. Accessed: 2023-11-28.

Universal NER. 2023. UNER LLM Instructions. https://huggingface.co/datasets/universalner/uner_llm_instructions. Accessed: 2023-11-28.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023. Not enough data to pre-train your language model? MT to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836, Toronto, Canada. Association for Computational Linguistics.

Cécile B. Vigouroux. 2013. Francophonie. *Annual Review of Anthropology*, 42(1):379–397.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022c. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022d. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

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

Richard E West, Timothy Newby, Zui Cheng, Alyssa Erickson, and Kyle Clements. 2020. Acknowledging all learning: Alternative, micro, and open credentials. *Handbook of Research in Educational Communications and Technology: Learning Design*, pages 593–613.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*.

W Bruce Willis. 1998. *The Adinkra dictionary: A visual primer on the language of Adinkra*. Pyramid Complex.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023a. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023b. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Peter Wittenburg. 2021. Open Science and Data Science. *Data Intelligence*, 3(1):95–105.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*.

Walt Wolfram. 1997. Issues in dialect obsolescence: An introduction. *American speech*, 72(1):3–11.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2023b. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. 2023. Dataset pruning: Reducing training data by examining generalization influence.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023a. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023b. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Zheng Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, et al. 2023c. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. In *Sixth Workshop on Computational Approaches to Linguistic Code-Switching*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2021. Synthbio: A case study in human-ai collaborative curation of text datasets. *CoRR*, abs/2111.06467.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. 2023a. Chinese open instruction generalist: A preliminary release.

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

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023c. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. *CoRR*, abs/1803.00353.

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

Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. 2024. Astraios: Parameter-efficient instruction tuning code large language models. *arXiv preprint arXiv:2401.00788*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A A Participatory Approach to Research

Recent breakthroughs in NLP have predominantly come from narrow collaborations that involve researchers from a handful of institutions and regions of the world (Nakamura et al., 2023). This reliance on small, specialized collaboration networks has been shown to hinder innovation (Park et al., 2023). Dataset creation as a process has often been undervalued, with minimization of the value of creators' contributions (Andress et al., 2020; Peng et al., 2021; Hanley et al., 2020). Under such conditions, the richness and diversity of the data are often compromised, as it reflects a limited perspective that aligns with the interests of those who wield greater power in these transactions. Data is not, as metaphors such as *'data mining'* (Puschmann and Burgess, 2014) or *'data is the new oil'* (Stark and Hoffmann, 2019; Awati and Shum, 2015) might suggest, a natural resource waiting to be exploited. Whenever we engage with data, we are also engaging with the connections that data has to the people who produce, prepare, and distribute it (Seaver, 2021; Pinel C, 2020; Crawford, 2021). Participatory approaches in AI design and research are one way to address gaps in access to resources needed for research: through collaborative partnerships with language speakers and local communities.

Aya is an example of a participatory research project (Birhane et al., 2022; Corbett et al., 2023; Delgado et al., 2023). Here, the research is the result of a broad cross-institutional, global collaboration. This type of cross-sectional work facilitates the collection of vital linguistic data and community engagement, which is crucial for developing effective language technologies (Joshi et al., 2019; ∀ et al., 2020). We describe below some of the guiding principles we followed throughout the year-long **Aya** project.

Fluid Ownership and Growth. Our open science framework allowed us to challenge the norms of how computer science usually proceeds (Wittenburg, 2021; Sabou et al., 2012). Traditional research approaches often involve rigid hierarchies; typically, research is conducted within academic institutions or corporate labs where roles are clearly defined, and collaboration is mostly synchronous, relying on in-person meetings or real-time communication. In contrast, **Aya** took a decentralized and democratic approach to collaboration, supporting fluid leadership and flexible role adoption. This empowered members to take initiative and lead in areas where they had passion or expertise, regardless of their position in academia, or when they became involved in the project. For example, members became Language Ambassadors at many different points during the year-long project, and mentorship roles evolved naturally with more experienced researchers providing guidance to those more junior (see Appendix G for more details of different roles in the project).

Organizational Structure. The communication channels and organizational structure of **Aya** were designed to facilitate rich collaboration that could evolve with the interests of participating researchers over the year-long project. For example, most communication between independent researchers involved within **Aya** was asynchronous over Discord, which allowed researchers in different time zones to participate in discussions. Monthly meetings were open for anyone to attend and were recorded for asynchronous viewing. We describe the structure of meetings and communication more thoroughly in Appendix G.3.1 and Appendix G.3.2.

Inclusion and Access. The open nature of the **Aya** UI allowed us to bypass the gatekeeping mechanisms of academic science that often marginalize non-English speakers and people without formal academic credentials (West et al., 2020). Expertise in the command of a spoken or written language is clearly distinct from expertise in machine learning. The inclusion of such a wide range of volunteers gave us more representative data in a wide variety of languages and also gave volunteers a glimpse into the often obscure world of machine learning.

Who Participated in Aya . The motivations of contributors were not based on financial remuneration but on ideals of community, identity, and social justice. Participants saw their roles as Language Ambassadors and contributors as crucial to ensuring the inclusion of their languages in the ongoing transition to a digital, information-driven economy. The Language Ambassador for Malagasy, a language driven to the risk of extinction by colonial French rule in Madagascar (Spolsky, 2018), is planning hackathons in 2024 to use the **Aya** Dataset to create voice-to-text apps that will help non-literate speakers of Malagasy participate in the modern economy. In Telugu, a traditional genre of poetry known as Sathakam is an integral part of the educational system. However, chatbots that can translate text into Telugu have little to no understanding of the Sathakam form. The Telugu Language Ambassador told a newspaper in Toronto that "in **Aya** , we made sure to include as many Sathakams as we could find" (Castaldo, 2023).

These motivations are not peripheral to the strength of the final **Aya** Dataset but are key factors in the data's provenance (Loukissas, 2019). These qualitative dimensions remind us that language is, for the people who use it every day, an intimately social phenomenon. Beyond the symbolic notation that connects tokens to referents in the real world, we find a robust network of social relations that are necessary for languages to flourish (Sidnell and Enfield, 2012; Goodwin, 2017; Agha, 2006). The social interactions between contributors, ML researchers, and social scientists in the **Aya** project were crucial to its success. Contributors shared playlists of their favorite songs from their home country, recipes from their childhood, and snapshots of the views from their home offices. They debated subtle nuances of how they wanted their language represented in the dataset and pushed back on some of the assumptions made by project coordinators on what constituted a distinct language as opposed to a regional dialect (see Section 7). More than one contributor sat down with their grandparents to contribute to a language that spanned three generations of use.

The realities of the conditions under which many people work and live were present every day. For example, Zoom meetings were cut short for some volunteers due to power outages in their countries or lack of access to a stable internet connection. Burmese, a language spoken in Myanmar, started out strong in the project with a group of 35 motivated volunteers but saw a sudden pause in contributions as civil war broke out in the country resulting in the withdrawal of the volunteers from the project (Petty, 2023). The Language Ambassador for Armenian also had to drop out of the project because of a conflict in that country (Reuters, 2023). In some countries, postal services only functioned a few days per month because of ongoing warfare, creating challenges for organizers when mailing out **Aya** gifts to thank committed volunteers. Ultimately, organizers were not able to send gifts to thank volunteers who participated from Somalia, Yemen and Palestine. For Somalia and Yemen, both Canada Post, DHL and Fedex were not able to support shipments. For Palestine, the cost of shipment proved to be prohibitively expensive – with an estimated shipping cost of 294 US dollars per t-shirt. These geo-political realities shaped both our contributors' experience as well as the progress of the project.

Including these factors in our post-mortem analysis of the project is crucial to understanding both the motivation of people willing to volunteer for open-science projects, and also to understanding the data itself: its breadth, its provenance, its shortcomings, and its living history.

B Related Work

B.1 Multilingual datasets

Low-resource languages have long been a challenge in NLP, with limited data impacting task performance (Kunchukuttan et al., 2021). To address this, researchers have explored techniques like data augmentation (Sennrich et al., 2016; Dhole et al., 2021), transfer learning (Zoph et al., 2016), repeating (Luukkonen et al., 2023; Muennighoff et al., 2023b), and multilingual models (Dabre et al., 2020; Muennighoff et al., 2023c; Yong et al., 2023b), achieving promising results in areas like machine translation. Here, we focus on efforts that are centered on multilingual dataset creation.

Several works have created large-scale multilingual corpora. These are often unstructured texts, ideal for large-scale unsupervised pretraining (Abadji et al., 2021; Ortiz Su'arez et al., 2019; Scao et al., 2022a,b; Laurençon et al., 2022; Kudugunta et al., 2023; Whitehouse et al., 2023). Another group of multilingual datasets is focused on machine translation (Lucia Specia et al., 2010; Fan et al., 2021). They consist of parallel texts in two or more languages, enabling models to learn the mappings between them. Ideally, machine translation datasets encompass diverse domains and language pairs, from commonly spoken languages to resource-scarce ones, promoting inclusivity and linguistic diversity. One of the most extensive collections of parallel corpora is available at the OPUS project website[8] (Tiedemann, 2012). Large capacity models for language understanding may obtain strong performance on high-resource languages while greatly improving low-resource languages (Goyal et al., 2021a). In (Whitehouse et al., 2023), the effectiveness of LLM-powered data augmentation in cross-lingual commonsense

[8]https://opus.nlpl.eu

reasoning was demonstrated. Improved performance was shown when smaller cross-lingual models were fine-tuned with data generated by LLMs. Some recently released datasets focus on specialized language domains such as law (Niklaus et al., 2023), education (Zhang et al., 2023c), or healthcare (Wang et al., 2023).

These corpora often suffer from inadequate data quality and require extensive cleaning (Abadji et al., 2022; Kreutzer et al., 2022). Task-specific datasets (Ponti et al., 2020; Conneau et al., 2018) are often smaller in scale but offer higher quality data targeted at a specific model capability such as cross-lingual understanding and transfer learning. This type of data is crucial for evaluating and enhancing the performance of models in diverse linguistic contexts. Such datasets are aggregated in multilingual benchmarks (Hu et al., 2020; Ruder et al., 2021; Cahyawijaya et al., 2021). Recently, (Ruder et al., 2023) released XTREME-UP, which covers data in 88 under-represented languages across 9 user-centric technologies.

No Language Left Behind (NLLB-Team et al., 2022) open-sourced bitext, mined bitext, and data generated using back-translation in 200+ languages specifically for text-to-text translation. While Seamless4MT (Barrault et al., 2023) released the metadata of SeamlessAlign, an open multimodal translation dataset, there are relatively fewer works for data creation/curation in low-resource languages. (Cahyawijaya et al., 2023) introduced NusaCrowd, a standardized collection of 137 datasets covering 19 Indonesian local languages in text, speech, and image modalities. Our work differs from previous datasets as we create a large-scale instruction-tuning dataset spanning hundreds of different tasks, yet retain high quality by involving human annotation and rigorous quality control across the entire data creation process.

B.2 Instruction-tuning datasets

Instruction-tuning datasets are collections of human-curated instructions and response pairs, templatized NLP tasks, or synthetic instructions generated by a language model. There are a growing number of NLP meta-datasets such as Natural instructions (Mishra et al., 2022), SuperNatural Instructions(Wang et al., 2022c), Flan 2021 (Wei et al., 2022a), Flan 2022 (Longpre et al., 2023a), Public Pool of Prompts (P3) (Sanh et al., 2022), Unnatural Instructions (Honovich et al., 2023), OPT-IML (Iyer et al., 2022), inter alia (Khashabi et al., 2020; Ye et al., 2021; Min et al., 2021) that collate numerous instruction fine-tuned datasets together. Some work focuses on specific applications such as dialogue (Köpf et al., 2023), structured knowledge grounding (Xie et al., 2022), or chain-of-thought reasoning (Wei et al., 2022b; Kim et al., 2023). Manual efforts include Open Assistant (Köpf et al., 2023) crowd-sourcing volunteers who wrote both instructions and responses, Databricks employees creating 15k examples in Dolly (Conover et al., 2023), and LIMA (Zhou et al., 2023) which is a collection of 1,000 author-curated IFT examples.

Synthetic instruction-tuning datasets comprise instructions sampled from a language model, such as the Self-Instruct dataset (Wang et al., 2022a) generated by GPT-3 (Brown et al., 2020), the Alpaca dataset (Taori et al., 2023) generated by GPT-3.5, and the Guanaco dataset (Joseph Cheung, 2023). Increasingly, the synthetic generation of instruction fine-tuned datasets is more sophisticated. (Xu et al., 2023a) propose a novel Evol-Instruct framework to obtain complex and difficult instructions gradually. (Luo et al., 2023) and (Gunasekar et al., 2023) further expand this idea to promote reasoning, code generation, and algorithmic skills. InstructionWild (Ni et al., 2023) and ShareGPT[9] are collections of user-shared conversations with ChatGPT.

B.3 Multilingual Instruction-Tuning Datasets

Despite ever-larger collections of IFT datasets, prior work has been largely English-centric. Most approaches to extend instruction fine-tuned datasets outside of English have relied on **1)** translating English datasets into other languages (Holmström and Doostmohammadi, 2023; Li et al., 2023a; Winata et al., 2023b), **2)** template-based dataset creation (Yu et al., 2023; Gupta et al., 2023) or **3)** human curation of instruction datasets in non-English languages (Muennighoff et al., 2023c; Li et al., 2023c; Wang et al., 2022b). There have been some notable exceptions with large proportions of non-English data (Joseph Cheung, 2023; Köpf et al., 2023; Lai et al., 2023; Li et al., 2023a; Longpre et al., 2023a; Muennighoff et al., 2023a,c; Zhuo et al., 2024; Nguyen et al., 2023).

Template-Based Datasets. The most relevant effort is recent work by (Muennighoff et al., 2023c) releasing Crosslingual Public Pool of Prompts

[9] https://sharegpt.com/

(xP3). xP3 expands the P3 taxonomy and adds 28 new multilingual datasets. However, their datasets usually use the same template in different languages, thus limiting task diversity. For example, a random batch from their dataset may include the same sample in different languages multiple times. Their xP3 corpus has task instructions exclusively in English. In (Muennighoff et al., 2023c), the experiments with matching the task instruction to the respective language of the sample via machine translation (xP3mt) showed slightly improved performance for non-English task instructions at inference. Our work is distinct in that our human-curated constructed dataset is unique for each of the 65 languages. Such diversity has been emphasized as a key ingredient for instruction tuning (Longpre et al., 2023a). Further, we create non-English task instructions via human annotators, ensuring these are of high-quality, which is another pillar of a good performance (Zhou et al., 2023).

Machine Translated Datasets. The most relevant effort is recent work by (Muennighoff et al., 2023c) releasing Crosslingual Public Pool of Prompts (xP3). xP3 expands the P3 taxonomy and adds 28 new multilingual datasets. However, their datasets usually use the same template in different languages, thus limiting task diversity. For example, a random batch from their dataset may include the same sample in different languages multiple times. Their xP3 corpus has task instructions exclusively in English. In (Muennighoff et al., 2023c), automatically translating the task instruction to the respective language of the sample (xP3mt) showed slightly improved performance. Our work is distinct in that our human-created dataset is unique for each of the 65 languages. Such diversity has been emphasized as a key ingredient for instruction tuning (Longpre et al., 2023a). Further, we create non-English task instructions via human annotators, ensuring these are of high-quality, which is another pillar of good performance (Zhou et al., 2023).

Machine Translated Datasets Machine-translated prompts often lack variability and the cultural nuance inherent in natively written text. However, they are still useful for expanding the language coverage of the training data and can help bridge the resource gap for languages with limited training data (Urbizu et al., 2023; Lin et al., 2022). They can also adapt already-trained instruction-tuned language models to follow instructions in new languages (Yong et al., 2023b). Furthermore, LLMs trained on designed prompts have also been shown to be successful at tasks like EAE (Event Argument Extraction) from multilingual data in a zero-shot setup (Huang et al., 2022). (Zhang et al., 2023a) constructed high-quality Chinese instructions from existing English instruction datasets. They first translated the English instructions into Chinese and then used a human verification process to determine whether these translations were usable; the verified dataset set consists of around 200k Chinese instruction-tuning samples. (Li et al., 2023a) constructed instruction data for 52 popular languages using Google Translate to translate English prompts and completions from Alpaca (Taori et al., 2023) (52K) and Dolly (Conover et al., 2023) (15K) datasets, then used this data to fine-tune LLaMA (Touvron et al., 2023) using LoRA (Hu et al., 2021). (Zhang et al., 2023b) fine-tuned LLaMA on multi-turn interactive translations, improving its multilingual translation abilities.

Human-Curated Multilingual Examples. Most relevant to our work on the **Aya** dataset are other datasets that have been curated by humans, often in English. Databricks collected a 15k instruction dataset databricks-dolly-15k by relying on its employees as annotators (Conover et al., 2023). Annotators were instructed to curate prompt / response pairs in each of eight different instruction categories. (Köpf et al., 2023) released the OpenAssistant corpus with over 10,000 dialogues from more than 13,500 international annotators. While this dataset contains multilingual annotations, this was not an explicit goal of the initiative. In contrast to our corpus which only has 2.05% contributions in English, 42.8% of the OpenAssistant data is in English (Köpf et al., 2023).

B.4 Participatory Research in Machine Learning

Prior participatory research initiatives have centered around regions or specific tasks like translation or character recognition. For example, (Clanuwat et al., 2018) tackles the problem of reading and understanding *Kuzushiji*, a cursive style of Japanese writing no longer in common use. Another example of culturally diverse data collection is (Liu et al., 2021), which recruited native speakers from five languages (Indonesian, Swahili, Tamil, Turkish, and Mandarin Chinese) that are typologically, genealogically, and geographically diverse, to provide images of concepts that are representative of their cultures. Then, they recruited native-speaking professional linguists to write captions for these images. However, this dataset is small (less than 8,000 data points) and thus limited to evaluation only. It is worth noting that these works are solely focused on the image domain, unlike our work, which concentrates on text.

More relevant to our work are participatory data creation initiatives focused on NLP. (Guevara-Rukoz et al., 2020) presents a study focusing on the creation of a crowd-sourced corpus for Latin American Spanish dialects to address the scarcity of resources for these languages. ($\forall$ et al., 2020) focuses on the task of Machine Translation (MT), and curates a dataset in 30 under-represented African languages according to a participatory research framework. Our work is very much in the spirit of these prior efforts, with differences in terms of global rather than regional focus. In contrast to these works, which have a specific regional focus, **Aya** collaborators came from multiple continents covering a diverse range of languages.

Several works have explored the organizational structures required to facilitate the development of research communities around under-represented languages. (Siminyu et al., 2021) details work on the AI4D - African Language Program, which aimed to enhance language resources for African languages. The outcome included creating over nine open-source African language datasets and establishing baseline models, demonstrating the program's significant impact on language technology for African languages. (Azunre et al., 2021) describe the establishment of NLP Ghana, with its collaborative open-source community. (Strassel and Tracey, 2016) discuss the challenges of developing resources for low-resource languages under the LORELEI (Low Resource Languages for Emergent Incidents) program. They focus on the pressing need for digital resources in these languages, particularly in critical situations such as mitigating the effects of natural disasters.

Open science community initiatives like **Aya** yield significant advancements in language modeling. Related efforts (in terms of compute and resources required) can be found in the BigScience Workshop (Akiki et al., 2022), which began in 2021. The BigScience project was initiated to address the limitations in LLM development, emphasizing open science and inclusive collaboration. Leveraging open science principles, it united a global network of researchers working to collaboratively and ethically enhance machine learning. Their work culminated in key developments like the BLOOM model (Scao et al., 2022a) and ROOTS corpus (Laurençon et al., 2022). These achievements underscore the value of community-driven, ethical, and diverse research programs for large-scale language technologies. Similar to Big Science, there have been other recent efforts on open science in language modeling (Groeneveld et al., 2024; Soldaini et al., 2024; Srivastava et al., 2022).

C Aya Dataset: Additional Analysis

C.1 Contributors

We aimed to include individuals from diverse backgrounds—not limited to AI experts—enabling anyone proficient in a language to contribute. During the registration process, we request demographic information from each **Aya** UI user such as country of residence, languages of fluent communication, gender, age range, and familiar dialects. The **Aya** community of contributors includes 2,997 registered users across 134 languages.

Figure C.1 illustrates the demographics of registered **Aya** UI users by age and gender. Regarding the age profiles of users, more than two-thirds were aged between 18 and 35. Approximately 68.1% of users identified as male and 28.5% as female. Overall, 6.6% of users self-reported dialects. Within this group, 75% specified one dialect, 20% specified two dialects, and the remaining 5% specified three or more dialects, with a maximum of six.

C.2 Annotation Guidelines

The annotators were provided with the following evaluation criteria for what a good prompt and completion pair must look like. Re-annotations were then performed if they determined that the prompts or completions needed editing.

1. No grammatical or spelling mistakes in both the prompt and completion text.
2. The prompt provides clear instructions on what the task is.
3. The completion answers the prompt correctly. Both the prompts and completions should be in full sentences and coherent, with reasonable length.

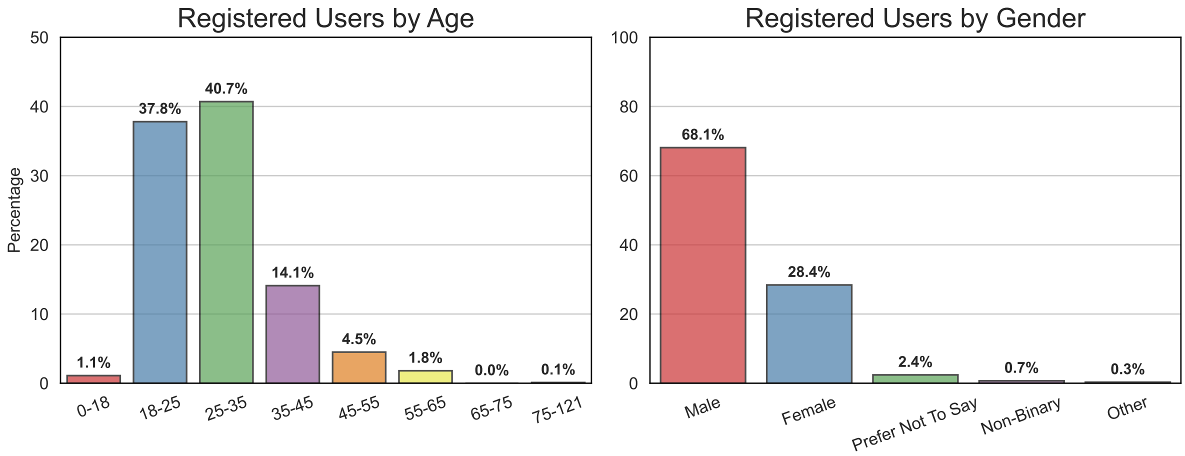

Figure C.1: **Left:** Distribution of registered users on the **Aya** UI by age using specified values. **Right:** Distribution of registered users on the **Aya** UI by gender using specified values

4. For original annotations, the prompts and completions should not be generated from other language models.

Re-Annotations. Before editing, annotators rated the quality of existing prompt and completion pairs by choosing either the thumbs-up or thumbs-down option. If the provided prompt and completion pair were already of good quality according to the criteria above, then annotators rated them with thumbs up and moved ahead without editing. Overall, annotators were encouraged to re-annotate the completions, in particular by adding more details and context to them since many of them were often short one-word answers.

C.3 Language Groupings

In this work we will refer to groups of languages to be "lower-", "mid-" or "higher"-resourced according to their recorded, written, and catalogued NLP resources (Joshi et al., 2020). (Joshi et al., 2020) group languages into 5 distinct clusters based on the amount of data from a combined range of sources (LDC catalog[10], ELRA Map[11], Wikipedia [12]), which we interpret as a proxy for data availability for pretraining and IFT training of LLMs. We group these 5 distinct clusters into a rough taxonomy of **lower-resourced (LR)**, **mid-resourced (MR)** and **higher-resourced (HR)** (See Table C.1). See Table E.4 for full mapping of languages to categories. We note that this grouping is inevitably imperfect; languages and their varieties cannot absolutely nor universally be classified based on this single dimension (Hämäläinen, 2021; Lignos et al., 2022; Bird, 2022). The categorization in our case serves the purpose of aggregation in our analysis of the data distribution.

[10]https://catalog.ldc.upenn.edu/
[11]https://catalog.elra.info/en-us/
[12]https://wikipedia.org/

C.4 Length difference by language

Figure F.6 in the Appendix illustrates the statistics per language. We observe an array of patterns that differ across languages. For instance in Japanese, completions are on average 31% shorter than prompts. On the other end, for Urdu and Yoruba, completions are notably long relative to prompts. On average, completions are 1258% and 2516% longer than the corresponding prompts for Urdu and Yoruba, respectively. The average completion length in Yoruba is 1591% longer than the average prompt length in Japanese. Figure F.6 provides the average length of the combination of prompts and completions per language.

C.5 Annotator Skew

A feature of participatory research projects is the challenge of establishing and maintaining a balanced number of annotations across groups of annotators. In the **Aya** project, the number of annotators per language varied due to numerous factors. As a result, the distribution of annotators is not uniform across languages. Moreover, within each language, there is a lack of consistent contributions from all annotators. In this section, we examine the impact of annotator skew on the resulting dataset.

Group	Category	Languages	Examples
Higher-Resourced	5	7	Arabic, Chinese, English, French, Spanish
	4	18	Hindi, Italian, Portuguese, Russian, Turkish
Mid-Resourced	3	25	Afrikaans, Indonesian, Kazakh, Malay, Latvian
Lower-Resourced	2	13	Hausa, Icelandic, Irish, Lao, Maltese
	1	39	Albanian, Gujarati, Igbo, Luxembourgish
	0	12*	Kurdish, Kyrgyz, Sinhala, Yiddish

Table C.1: Language grouping for the **Aya** Collection. We assign categories to languages based on (Joshi et al., 2020). (*) We assign label 0 to two languages not found in Joshi et al. (2020)'s taxonomy (manipuri and ngaju).

C.5.1 Annotator Skew Across Languages

Annotators were encouraged to contribute to any language in which they could comfortably read and write and were asked to focus most of their efforts on languages other than English. Although a significant number of participants registered for many languages, the engagement level of annotators was not equal, which resulted in considerable differences in the number of contributions across languages. Figure C.2 (top) provides an overview of the percentage of each language present in the final compilation. The highest number of contributions is for Malagasy with 14,597 instances, and the lowest is 79 for Kurdish.

C.5.2 Annotator Skew Within a Language.

The final contributions for each language in the **Aya** Dataset are not evenly distributed among annotators. The median number of annotators per language is 15 (mean is 24.75) with one language having only a single active annotator (Sindhi) and some having over 80 annotators (English and Portuguese). Note that annotators made contributions at varying rates, and there is no direct correlation between the number of annotators and the ultimate count of language contributions. A limited pool of annotators for some languages implies that most instances in that language originate from a smaller group of individuals. Figure C.2 (bottom) illustrates the proportion of instances in a language originating from the most active annotators. We observe a skewed pattern where for 12 languages, the 5 most active annotators contributed all examples. There is an uneven distribution of contributions for many languages because those languages had a smaller number of voluntary annotators throughout the entire project despite rigorous outreach. Additionally, we did not establish a specific quota for annotators to meet; everyone contributed as they desired, resulting in varying levels of activity among annotators.

The most extreme cases are Zulu and Sindhi, where one annotator in each language volunteered for all contributions in Annotation and Reannotation tasks. Thus, in Figure C.2 their top-1 contributor ratio is 1.0 and does not change when moving to top-2 or further. The languages with the least skewed distributions are Malagasy, Tamil, Nepali, Hindi, English and Portuguese. English also had the highest number of unique annotators with 130 individuals out of which 95 annotators contributed to English as their second language for annotation purposes. Given the uneven distribution of annotators per language, it is important to acknowledge that individual annotator quality has a disproportionate influence on some languages.

D Aya Collection: Additional Analysis

D.1 Translation Quality

Figure D.3 shows the translation quality across languages grouped by their resourcefulness. The mean ChrF++ score on FLORES is 48.17 (min: 10.9, max: 69.6) for translations out of English, with a few outliers for HR and LR. We interpret this optimistically as strong enough to sufficiently serve our translation needs. However, upon inspection of translation outputs for fine-tuning data, we encountered significant translation errors with Standard Arabic in Latin script and Minangkabau in Arabic script, so we excluded them from our translated dataset. In total, 19 public datasets were translated into 101 languages (114 dialects).

In addition to releasing the translated datasets used as a basis for re-annotation, we also translated Dolly (Conover et al., 2023). Dolly comprises 15k instructions that Databricks collected by relying on its employees as annotators (Conover et al., 2023). Annotators were instructed to curate prompt and completion pairs in each of eight

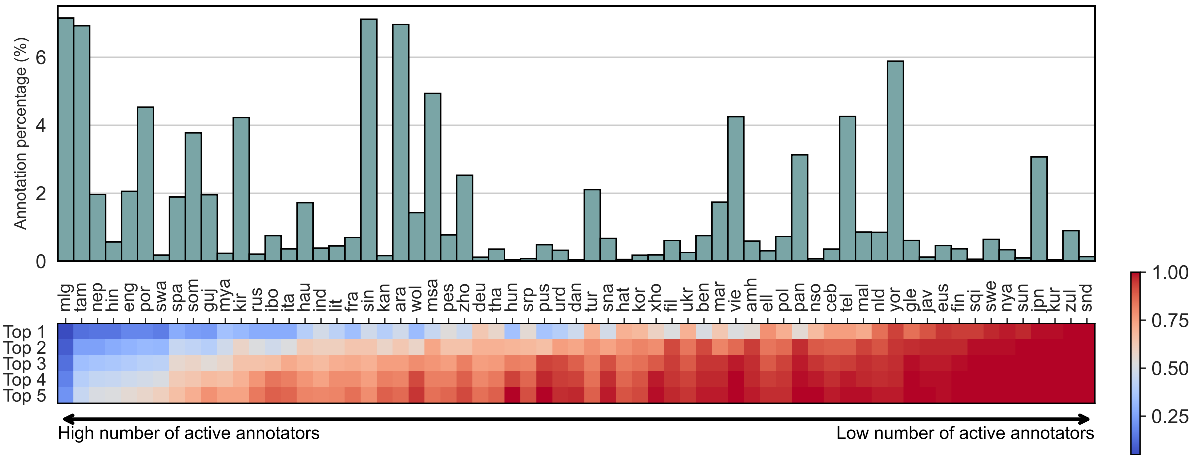

Figure C.2: **Top**: Ratio of all annotations per language with respect to the whole dataset. **Bottom:** Ratio of annotations done by the top-k most active contributors ($k = 1, ..., 5$). Languages annotations follow their respective ISO codes from Table E.4.

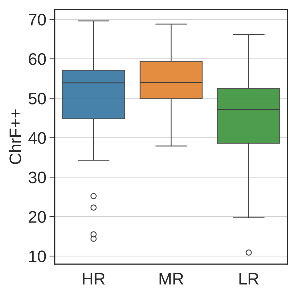

Figure D.3: ChrF++ scores for the NLLB translation model, averaged across resourcefulness buckets.

different instruction categories. In contrast to the mentioned NLP datasets, Dolly was purposefully designed to align language models with human expectations. It stands out as a high-quality, manually curated dataset covering a range of topics including brainstorming, classification, closed question answering, generation, information extraction, open question answering, and summarization. The addition of the translated Dolly datasets is a valuable resource for languages that face a scarcity of conversational instruction fine-tuning data. The list of datasets, along with the number of languages, templates, and other statistics, can be found in Table E.8.

D.2 Tasks Covered Across Templated and Translated Datasets

We aim to include datasets from various tasks in the collection while ensuring that they follow our selection criteria. Table D.2 illustrates our task coverage in the **Aya** Collection, drawing inspiration from xP3 and the Flan Collection. We have a total of three main task types: Question Answering (QA), Natural Language Generation (NLG), and Text Classification (TC). Within these larger umbrella tasks, we define several finer-grained task types based on the datasets, resulting in a total of 11 finer-grained task types. These finer-grained task types are determined by the frequency of datasets in the **Aya** Collection encapsulating that task. For QA, we decided to keep only the main task type, as the intended goal of question-answering tasks is clear: *Answer a proposed question*. The type of the question can be different: open-ended, close-ended, multiple-choice, single response. For NLG, finer-grained task types include Summarization, Translation, Paraphrasing, Dialogue (Generation), and Text Simplification. For TC, we include the following finer-grained task types: Sentiment Analysis, Information Extraction, Named Entity Recognition, Event Linking, Natural Language Inference, and Scientific Document Representation. Finally, we label the task categories of each dataset in the **Aya** Collection in Table E.5 and Table E.6. If we are not able to find a fine-grained task type for the dataset, we keep the main task type.

D.3 Prompt and Completion Lengths

Figure D.4 shows the distribution of length across languages. No discernible pattern is observed when examining lengths for high-resource languages compared to low-resource languages. Low-resource languages appear at both ends of the distribution, occupying both the head and tail. In the **Aya** Collection some low-resource languages (e.g., Somali and Amharic) have longer average com-

Main Task Type	Fine-grained Task Type
Question Answering	—
Natural Language Generation	Summarization Translation Paraphrasing Dialogue Text Simplification
Text Classification	Sentiment Analysis Information Extraction Named Entity Recognition Event Linking Natural Language Inference Document Representation

Table D.2: Task Taxonomy of NLP tasks in the **Aya** Collection.

pletions length than medium or even high-resource languages. The dedication of individual participants in identifying datasets in their own language and templating them has made a significant difference for many languages.

D.4 Quality Assessment of All Different Data Sources

As previously stated, contributors could provide binary feedback on the quality of the prompt-completion pairings. We define the average approval ratio per dataset which serves as a valuable metric for assessing the quality of datasets across various languages and diverse data sources. We compute the average approval ratio as $\mathcal{T}_{+}/\mathcal{T}$, where $\mathcal{T}_{+}$ represents the total number of upvotes (thumbs-up), and $\mathcal{T}$ represents the total number of votes per dataset. An average approval ratio of 1.0 would indicate that every annotation was perceived to be of good quality and all prompts and completions had received a thumbs-up. An average approval ratio of 0.0 would indicate that every annotation was perceived to be of poor quality, and all prompts and completions had received a thumbs-down. We constrained our quality analysis to the 40 datasets in our pool for which we had at least 20 instances of feedback.

Overall, we observe that the majority of datasets had an approval ratio of over 0.5, with all translated data as well as Original Annotations being above this threshold. However, across all the datasets within each group —xP3, Templated, Translated, and **Aya** original annotations— **Aya** original annotations were perceived to be of the highest quality, with an approval ratio of 0.81, compared to the lowest quality dataset, xP3, which only had a ratio of 0.50. This corroborates our intuition that a manual curation process leads to the highest-quality annotations. Figure D.5 provides a summary of the results for each group. Figure F.7 in the Appendix provides approval ratios per dataset in each group.

E Aya Evaluation Suite: Additional Analysis

E.1 Post-Editing the DOLLY-MACHINE-TRANSLATED Test Set

E.1.1 Annotators

Annotator Selection. The primary demographic make-up of the participants in the evaluations was recruited based on their proficiency in the language groups. The proficiency was self-reported, and our requirements were natively proficient or professionally proficient in the specific languages needed for the project. Outside of this, the participants come from diverse social backgrounds comprised of students and individuals with full-time or part-time jobs that do annotation as a "side gig".

Socio-Demographics. The annotator pool is comprised of people from diverse backgrounds, and this spans across socioeconomic backgrounds, careers, levels of education, and self-reported gender and sexual identities. We do not ask any annotators to share or report any of these statistical pieces of information in a formal way; any insights into this are gathered organically and through self-reporting by the annotators.

Quality Considerations. We do not believe that any socio-demographic characteristics have led to any impact on the data that has been annotated. Through every part of the project we have reiterated the importance of this work and the fact that this is helping to support a global-scale research project. We are confident in the trust we have built with the annotators in this project, and they care greatly about the overall outcome and therefore have been diligent in completing the task with a high degree of accuracy. Where possible, we have done our best to have annotators work on this project and be representatives of the communities that the project aims to support.

Compensation. The annotators were paid 30 CAD per hour. No special consideration was made to the hourly rate as that is the standard rate offered to Cohere's annotators who work on highly complex tasks.

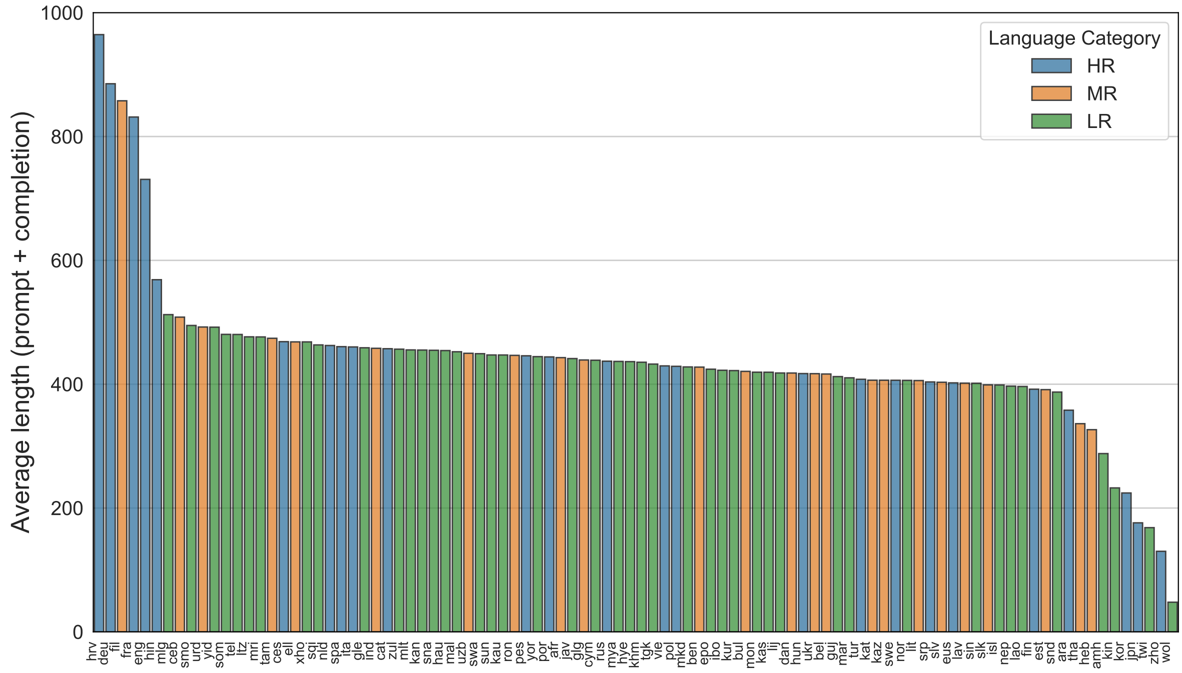

Figure D.4: The average length of prompts and completions for high (HR), medium (MR) and low-resource (LR) languages in **Aya** Collection.

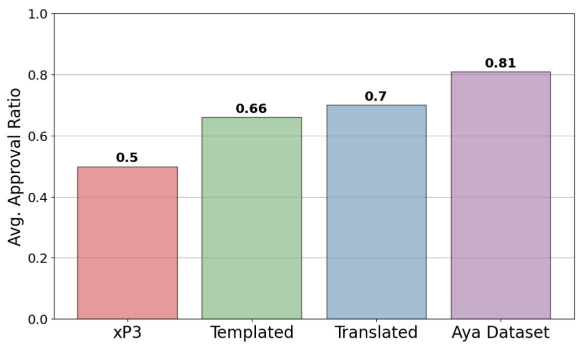

Figure D.5: Average approval ratio per dataset group, for datasets with at least 20 votes.

E.1.2 Annotation Process

Communication. Annotators were briefed by one of the authors in a virtual introduction session, and were able to ask questions and raise issues throughout the annotation task in a Slack channel. They were also encouraged to share frequent error patterns, artifacts, or hard decisions that they encountered throughout the task with the authors and other annotators.

Schedule. There was no fixed time schedule for the annotations and annotators contributed a varying amount of hours and ratings, depending on their availabilities and speed. Each translation was post-edited by one annotator, and there were 3–4 annotators involved in each task. After post-edits were completed, a second annotator (not the original post-editor) assessed the post-edit for quality and proposed new final edits if necessary.

Interface. Post-edits were collected on Google Sheets with an interface built-in Google Apps Script.

E.1.3 Instructions

The instructions given to professional annotators for the DOLLY-MACHINE-TRANSLATED test set post-edits were the following: “As an annotator, you have the task to improve the quality of the prompts for our multilingual model! The prompts are originally machine-translated from English, and sometimes the translation introduces errors in the prompts that make them hard to follow.

We need your help to identify these cases, and to edit these translations so that they...

1. Convey the same instruction/task/request as the English original — not more and not less.
2. Are grammatically correct.
3. Are free from phrases too literally translated from English (we call this “Translationese”).

This is how:
For each pair of English prompt and translated

prompt shown, decide whether the prompt is okay as it is (according to the above criteria), or needs an edit.

- If it needs an edit, edit the prompt until the quality is satisfactory (in the field "Edited Prompt"). Try to keep your edits minimal. Then confirm that the edited prompt fulfills the above three criteria.
- If it's okay as is, just proceed (without editing the "Edited Prompt" field) to confirm that it fulfills the above three criteria.

Annotations were done through an interface built on top of Google Sheets. One annotator edited each prompt, and another verified the edit, if necessary had a discussion and edited the original edit. Three to four editors collaborated on each language.

E.1.4 Post-Editing Effort

For each prompt, we measure the post-editing effort with Human-targeted Translation Error Rate (HTER) (Specia and Farzindar, 2010), an edit-distance metric that compares the original machine translation with the post-edited version in terms of edit operations on units of words. This also gives us an estimate of how severe the errors in the original translations were, and how critically the post-editors assessed the original translations. Analogously, we estimate with a Human-targeted Character F-Score (HChrF) score how much the original translation overlaps with the final post-edited translation. This metric is based on the ChrF score (Popović, 2015) and operates on character-level matches. Computations of HTER and HChrF are based on the sacrebleu implementation (Post, 2018).

Table E.3 reports these statistics for the six languages of the DOLLY-MACHINE-TRANSLATED test set. We find that editors edited at least 41% of prompts in all languages, a surprisingly high number. This indicates that translation errors in the DOLLY-MACHINE-TRANSLATED test set are quite common. For Russian, the post-editing effort was overall largest, with an average of 37.43 HTER, which means that 37.43% of words in the final post-edit had to be edited from the original. This stands in contrast with the post-edits for French, where a similar ratio of original prompts was edited (84.5% compared to 86.5% for Russian), but to a much lesser extent (5.56 HTER).

Language	% of Prompts Edited	HTER	HChrF
Arabic	41.0%	10.78	92.74
French	84.5%	5.56	96.81
Hindi	60.0%	6.16	95.00
Russian	86.5%	37.43	75.92
Serbian	72.5%	9.06	92.79
Spanish	75.5%	9.13	93.25

Table E.3: Post-editing effort measured by the overall percentage of edited dolly test prompts, HTER (Human-targeted Translation Error Rate: the higher, the more effort), and HChrF (Human-Targeted Character F-Score: the lower, the more effort).

ISO Code	Language	Script	Family	Subgrouping	Resources	Included
ace	Achinese	Arabic/Latin	Austronesian	Malayo-Polynesian	Low	♠
afr	Afrikaans	Latin	Indo-European	Germanic	Mid	♠
amh	Amharic	Ge'ez	Afro-Asiatic	Semitic	Low	♦ ♠
ara	Arabic	Arabic	Afro-Asiatic	Semitic	High	♦ ♠
aze	Azerbaijani	Arabic/Latin	Turkic	Common Turkic	Low	♠
ban	Balinese	Latin	Austronesian	Malayo-Polynesian	Low	♠
bbc	Toba Batak	Latin	Austronesian	Malayo-Polynesian	Low	♠
bel	Belarusian	Cyrillic	Indo-European	Balto-Slavic	Mid	♠
bem	Bemba	Latin	Niger-Congo	Atlantic-Congo	Low	♠
ben	Bengali	Bengali	Indo-European	Indo-Aryan	Mid	♦ ♠
bjn	Banjar	Arabic/Latin	Austronesian	Malayo-Polynesian	Low	♠
bul	Bulgarian	Cyrillic	Indo-European	Balto-Slavic	Mid	♠
cat	Catalan	Latin	Indo-European	Italic	High	♠
ceb	Cebuano	Latin	Austronesian	Malayo-Polynesian	Mid	♦ ♠
ces	Czech	Latin	Indo-European	Balto-Slavic	High	♠
cym	Welsh	Latin	Indo-European	Celtic	Low	♠
dan	Danish	Latin	Indo-European	Germanic	Mid	♦ ♠
deu	German	Latin	Indo-European	Germanic	High	♦ ♠
ell	Greek	Greek	Indo-European	Graeco-Phrygian	Mid	♦ ♠
eng	English	Latin	Indo-European	Germanic	High	♦ ♠
epo	Esperanto	Latin	Constructed	Esperantic	Low	♠
est	Estonian	Latin	Uralic	Finnic	Mid	♠
eus	Basque	Latin	Basque	-	High	♦ ♠
fil	Filipino	Latin	Austronesian	Malayo-Polynesian	Mid	♦ ♠
fin	Finnish	Latin	Uralic	Finnic	Mid	♦ ♠
fon	Fon	Latin	Niger-Congo	Atlantic-Congo	Low	♠
fra	French	Latin	Indo-European	Italic	High	♦ ♠
gla	Scottish Gaelic	Latin	Indo-European	Celtic	Low	♠
gle	Irish	Latin	Indo-European	Celtic	Low	♦ ♠
glg	Galician	Latin	Indo-European	Italic	Mid	♠
guj	Gujarati	Gujarati	Indo-European	Indo-Aryan	Low	♦ ♠
hat	Haitian Creole	Latin	Indo-European	Italic	Low	♦ ♠
hau	Hausa	Latin	Afro-Asiatic	Chadic	Low	♦ ♠
heb	Hebrew	Hebrew	Afro-Asiatic	Semitic	Mid	♠
hin	Hindi	Devanagari	Indo-European	Indo-Aryan	High	♦ ♠
hrv	Croatian	Latin	Indo-European	Balto-Slavic.	High	♠
hun	Hungarian	Latin	Uralic	-	High	♦ ♠
hye	Armenian	Armenian	Indo-European	Armenic	Low	♠
ibo	Igbo	Latin	Atlantic-Congo	Benue-Congo	Low	♦ ♠
ind	Indonesian	Latin	Austronesian	Malayo-Polynesian	Mid	♦ ♠
isl	Icelandic	Latin	Indo-European	Germanic	Low	♠
ita	Italian	Latin	Indo-European	Italic	High	♦ ♠
jav	Javanese	Latin	Austronesian	Malayo-Polynesian	Low	♦ ♠
jpn	Japanese	Japanese	Japonic	Japanesic	High	♦ ♠
kan	Kannada	Kannada	Dravidian	South Dravidian	Low	♦ ♠
kas	Kashmiri	Arabic	Indo-European	Indo-Aryan	Low	♠
kat	Georgian	Georgian	Kartvelian	Georgian-Zan	Mid	♠
kau	Kanuri	Arabic/Latin	Saharan	Western Saharan	Low	♠
kaz	Kazakh	Cyrillic	Turkic	Common Turkic	Mid	♠
khm	Khmer	Khmer	Austroasiatic	Khmeric	Low	♠
kin	Kinyarwanda	Latin	Niger-Congo	Atlantic-Congo	Low	♠
kir	Kyrgyz	Cyrillic	Turkic	Common Turkic	Low	♦ ♠
kor	Korean	Hangul	Koreanic	Korean	Mid	♦ ♠
kur	Kurdish	Latin	Indo-European	Iranian	Low	♦ ♠
lao	Lao	Lao	Tai-Kadai	Kam-Tai	Low	♠
lav	Latvian	Latin	Indo-European	Balto-Slavic	Mid	♠
lij	Ligurian	Latin	Indo-European	Italic	Low	♠
lit	Lithuanian	Latin	Indo-European	Balto-Slavic	Mid	♦ ♠
ltz	Luxembourgish	Latin	Indo-European	Germanic	Low	♠
mad	Madurese	Latin	Austronesian	Malayo-Polynesian	Low	♠
mal	Malayalam	Malayalam	Dravidian	South Dravidian	Low	♦ ♠
man	Manipuri	Bengali	Sino-Tibetan	Kuki-Chin-Naga	Low	♠
mar	Marathi	Devanagari	Indo-European	Indo-Aryan	Low	♦ ♠
min	Minangkabau	Latin	Austronesian	Malayo-Polynesian	Low	♠
mkd	Macedonian	Cyrillic	Indo-European	Balto-Slavic	Low	♠
mlg	Malagasy	Latin	Austronesian	Malayo-Polynesian	Low	♦ ♠

mlt	Maltese	Latin	Afro-Asiatic	Semitic	High	♠
mon	Mongolian	Cyrillic	Mongolic-Khitan	Mongolic	Low	♠
mri	Maori	Latin	Austronesian	Malayo-Polynesian	Low	♠
msa	Malay	Latin	Austronesian	Malayo-Polynesian	Mid	♦ ♠
mya	Burmese	Myanmar	Sino-Tibetan	Burmo-Qiangic	Low	♦ ♠
nep	Nepali	Devanagari	Indo-European	Indo-Aryan	Low	♦ ♠
nij	Ngaju	Latin	Austronesian	Malayo-Polynesian	Low	♠
nld	Dutch	Latin	Indo-European	Germanic	High	♦ ♠
nor	Norwegian	Latin	Indo-European	Germanic	Low	♠
nso	Northern Sotho	Latin	Atlantic-Congo	Benue-Congo	Low	♦ ♠
nya	Chichewa	Latin	Atlantic-Congo	Benue-Congo	Low	♦
pan	Punjabi	Gurmukhi	Indo-European	Indo-Aryan	Low	♦ ♠
pes	Persian	Arabic	Indo-European	Iranian	High	♦ ♠
pol	Polish	Latin	Indo-European	Balto-Slavic	High	♦ ♠
por	Portuguese	Latin	Indo-European	Italic	High	♦ ♠
pus	Pashto	Arabic	Indo-European	Iranian	Low	♦ ♠
ron	Romanian	Latin	Indo-European	Italic	Mid	♠
rus	Russian	Cyrillic	Indo-European	Balto-Slavic	High	♦ ♠
sin	Sinhala	Sinhala	Indo-European	Indo-Aryan	Low	♦ ♠
slk	Slovak	Latin	Indo-European	Balto-Slavic	Mid	♠
slv	Slovenian	Latin	Indo-European	Balto-Slavic	Mid	♠
smo	Samoan	Latin	Austronesian	Malayo-Polynesian	Low	♠
sna	Shona	Latin	Indo-European	Indo-Aryan	Low	♦ ♠
snd	Sindhi	Arabic	Indo-European	Indo-Aryan	Low	♦ ♠
som	Somali	Latin	Afro-Asiatic	Cushitic	Low	♦ ♠
sot	Southern Sotho	Latin	Atlantic-Congo	Benue-Congo	Low	♠
spa	Spanish	Latin	Indo-European	Italic	High	♦ ♠
sqi	Albanian	Latin	Indo-European	Albanian	Low	♦ ♠
srp	Serbian	Cyrillic	Indo-European	Balto-Slavic	High	♦ ♠
sun	Sundanese	Latin	Austronesian	Malayo-Polynesian	Low	♦ ♠
swa	Swahili	Latin	Atlantic-Congo	Benue-Congo	Low	♦ ♠
swe	Swedish	Latin	Indo-European	Germanic	High	♦ ♠
tam	Tamil	Tamil	Dravidian	South Dravidian	Mid	♦ ♠
taq	Tamasheq	Latin/Tifinagh	Afro-Asiatic	Berber	Low	♠
tel	Telugu	Telugu	Dravidian	South Dravidian	Low	♦ ♠
tgk	Tajik	Cyrillic	Indo-European	Iranian	Low	♠
tha	Thai	Thai	Tai-Kadai	Kam-Tai	Mid	♦ ♠
tur	Turkish	Latin	Turkic	Common Turkic	High	♦ ♠
twi	Twi	Latin	Niger-Congo	Atlantic-Congo	Low	♠
ukr	Ukrainian	Cyrillic	Indo-European	Balto-Slavic	Mid	♦ ♠
urd	Urdu	Arabic	Indo-European	Indo-Aryan	Mid	♦ ♠
uzb	Uzbek	Latin	Turkic	Common Turkic	Mid	♠
vie	Vietnamese	Latin	Austroasiatic	Vietic	High	♦ ♠
wol	Wolof	Latin	Atlantic-Congo	North-Central Atlantic	Low	♦ ♠
xho	Xhosa	Latin	Atlantic-Congo	Benue-Congo	Low	♦ ♠
yid	Yiddish	Hebrew	Indo-European	Germanic	Low	♠
yor	Yorùbá	Latin	Atlantic-Congo	Benue-Congo	Low	♦ ♠
zho	Chinese	Han	Sino-Tibetan	Sinitic	High	♦ ♠
zul	Zulu	Latin	Atlantic-Congo	Benue-Congo	Low	♦ ♠

Table E.4: 65 languages in the **Aya** Dataset and 114 languages in the **Aya** Collection, each language's corresponding script, family, subgrouping, and if it is classified as "lower-", "mid-" or "higher"-resourced according to the taxonomy classes by (Joshi et al., 2020) (low: [0, 1, 2], mid: [3], high: [4, 5]). The language is either included in the **Aya** Dataset (♦), **Aya** Collection (♠), or both. Note that Ngaju (nij) and Toba Batak (bbc) are not listed in (Joshi et al., 2020).

Main Task Type	Fine-grained Task Type	Dataset
Question Answering		AfriQA-inst (Ogundepo et al., 2023) Amharic QA (Abedissa et al., 2023) LLM-Japanese-Vanilla-inst (Tellarin.ai, 2023a) Mintaka-inst (Sen et al., 2022) X-CSQA-inst (Lin et al., 2021) TeluguRiddles (desik98, 2023)
Natural Language	Summarization	News-summary-instruct (TahmidH, 2023)

Generation		Persian-instruct-pn (Shafagh, 2023a) Hindi-article-summarization (ganeshjcs, 2023a) XWikis-inst (Perez-Beltrachini and Lapata, 2021)
	Translation	IndicSentiment-inst (George, 2023a) Indo-stories-instruct (Iftitahu, 2023a,b,c) Lijnews-instruct (ConseggioLigure, 2023a,b) SCB-MT-2020-prompt (PyThaiNLP, 2023a,b) Thai-USEmbassy-prompt (PyThaiNLP, 2023e,f) SEED-instruct-lij (ConseggioLigure, 2023c,d)
	Paraphrasing	Arpa-instruct (syntaxshill, 2023) IndicXParaphrase-inst (George, 2023b; SuryaKrishna02, 2023d) Turku-paraphrase-inst (TurkuNLP, 2023)
	Text Simplification	Wiki-split-inst (Botha et al., 2018)
	Dialogue	SODA-inst (Kim et al., 2022)
	NL Generation	Telugu-food-recipes (SuryaKrishna02, 2023a) Telugu-jokes (SuryaKrishna02, 2023b) Telugu-news-articles (SuryaKrishna02, 2023c) Telugu-poems (SuryaKrishna02, 2023e) TamilStories (AI Tamil Nadu, 2023a) Joke-explaination-inst (theblackcat102, 2023) Thirukkural-instruct (AI Tamil Nadu, 2023b) Hindi-article-generation (ganeshjcs, 2023b) Thai-Wiktionary-inst (PyThaiNLP, 2023d) UA-Gec-inst (osyvokon, 2023) Urdu-News-Gen-Article (AhmadMustafa, 2023a) Urdu-News-Gen-Headline (AhmadMustafa, 2023c) Thai-POS-inst (PyThaiNLP, 2023c)
Text Classification	Sentiment Analysis	AfriSenti-inst (Muhammad et al., 2023) IMDB-Dutch-instruct (jjzha, 2023) NusaX-senti-inst (Winata et al., 2023b)
	Information Extraction	NTX-LLM-inst (Tellarin.ai, 2023b)
	Named Entity Recognition	UNER-LLM-inst (Universal NER, 2023)
	Natural Language Inference	FarsTail-Instruct (hghader1, 2023)
	Event Linking	Xlel_wd-inst (Pratapa et al., 2022)
	Sci. Doc. Representation	Scirepeval-biomimicry-inst (Singh et al., 2022)
	Text Classification	Urdu-News-Category-Class (AhmadMustafa, 2023b) MasakhaNEWS-inst (Adelani et al., 2023)

Table E.5: Task Taxonomy of Templated Datasets (**Aya** Collection). We classify the templated datasets with a standard task taxonomy of three main tasks: Question Answering, Natural Language Generation, and Text Classification (Table D.2). We then have a fine-grained task taxonomy within each task, such as Summarization, Translation, Paraphrasing, Sentiment Analysis, Information Extraction, and Named Entity Recognition. If there is not a recognized fine-grained task taxonomy for a specific dataset, we put it in the main task type category.

Main Task Type	Fine-grained Task Type	Dataset
Question Answering		Adversarial QA (T) (Bartolo et al., 2020) Flan-Coqa (T) (Wei et al., 2022a; Reddy et al., 2019) Flan-unified-QA (T) (Wei et al., 2022a; Khashabi et al., 2020) HotpotQA (T) (Yang et al., 2018) Mintaka-inst (T) (Sen et al., 2022) MLQA-en (T) (Lewis et al., 2020) NQ-Open (T) (Kwiatkowski et al., 2019) PIQA (T) (Bisk et al., 2020) WIKI QA (T) (Yang et al., 2015)
Natural Language Generation	Summarization	CNN-Daily-Mail (T) (See et al., 2017) (Hermann et al., 2015) Flan-GEM-wiki-lingua (T)(Wei et al., 2022a; Ladhak et al., 2020)
	Text Simplification	Wiki-split-inst (T) (Botha et al., 2018)
	Dialogue	SODA-inst (T) (Kim et al., 2022)
	NL Generation	Joke-explaination-inst (T) (theblackcat102, 2023)

		Flan-CoT-submix (T)(Wei et al., 2022a) Flan-lambada (T) (Wei et al., 2022a; Paperno et al., 2016) Dolly-v2 (T) (Conover et al., 2023)
Text Classification	Event Linking	Xlel_wd-inst (T) (Pratapa et al., 2022)
	Paraphrase Identification	PAWS-Wiki (T) (Zhang et al., 2019)

Table E.6: Task Taxonomy of Translated Datasets (Aya Collection). We classify the translated datasets similar to templated datasets (Table E.5). If there is not a recognized fine-grained task taxonomy for a specific dataset, we put it in the main task type category.

Dataset	#Langs	Template lang	Dataset lang	$\bar{L}_{prompt}$	$\bar{L}_{compl.}$	License	Task
AfriQA-inst (Ogundepo et al., 2023)	12		bem, fon, hau, ibo, kin, swh, twi, wol, yor, zul, eng, fra	46	15	CC BY 4.0	Question Answering
AfriSenti-inst (Muhammad et al., 2023)	9		amh, arq, hau, ibo, kin, ary, por, swh, twi	168	44	CC BY 4.0	Sentiment Analysis
Amharic QA (Abedissa et al., 2023)	1	amh	amh	1114	33	MIT license	Question Answering
News-summary-instruct (TahmidH, 2023)	1	ben	ben	174	67	CC0 1.0	Summarization
Arpa-instruct (syntaxshill, 2023)	1	hye	hye	165	118	Artistic-2.0	Paraphrasing
Telugu-food-recipes (SuryaKrishna02, 2023a)	1	tel	tel	70	870	Apache 2.0	Generation
Telugu-jokes (SuryaKrishna02, 2023b)	1	tel	tel	80	276	Apache 2.0	Generation
Telugu-news-articles (SuryaKrishna02, 2023c)	1	tel	tel	448	426	Apache 2.0	Generation
Telugu-poems (SuryaKrishna02, 2023e)	1	tel	tel	357	198	Apache 2.0	Generation
FarsTail-Instruct (Amirkhani et al., 2023; hghader1, 2023)	1	pes	pes	224	112	Apache 2.0	Natural Language Inference
Hindi-article-summarization (ganeshjcs, 2023a)	1	hin	hin	3813	175	CC BY-SA 4.0	Summarization
Hindi-article-generation (ganeshjcs, 2023b)	1	hin	hin	102	3683	CC BY-SA 4.0	Generation
IMDB-Dutch-instruct (Maas et al., 2011; jjzha, 2023)	1	nld	nld	1470	31	Apache 2.0	Sentiment Analysis
IndicSentiment-inst (Doddapaneni et al., 2023; George, 2023a)	11	eng	ben, guj, hni, kan, mal, mar, pan, tam, tel, urd, eng	174	141	MIT	Translation
IndicXParaphrase-inst (Doddapaneni et al., 2023; George, 2023b; SuryaKrishna02, 2023d)	7	ben, guj, hin, mar, pan, mal, tel	ben, guj, hin, mar, pan, mal, tel	132	93	MIT	Paraphrase Identification
Indo-stories-instruct (Iftitahu, 2023a,b,c)	3	ind, sun, jav	ind, sun, jav	345	322	CC BY 4.0	Translation
Joke-explaination-inst (theblackcat102, 2023)	1		eng	118	548	MIT	Generation
Lijnews-instruct (ConseggioLigure, 2023a,b)	2	ita, lij	it, lij	893	898	CC BY 4.0	Translation
LLM-Japanese-Vanilla-inst (Suzuki et al., 2023; Tellarin.ai, 2023a)	1	jpn	jpn	60	97	CC BY-SA 4.0	Question Answering
MasakhaNEWS-inst (Adelani et al., 2023)	16		amh, eng, fra, hau, ibo, lin, cgg, orm, pcm, run, sna, som, swh, tir, xho, yor	1483	1459	AFL-3.0	Text Classification
Mintaka-inst (Sen et al., 2022)	9	eng	arb, deu, spa, fra, jpn, por, hin, ita, eng	102	49	CC BY 4.0	Question Answering
NTX-LLM-inst (Chen et al., 2023c; Tellarin.ai, 2023b),	13	arb, zho, nld, eng, fra, deu, hin, ita, jpn, kor, por, spa, tur	arb, zho, nld, eng, fra, deu, hin, ita, jpn, kor, por, spa, tur	917	493	CC BY-SA 4.0	Information Extraction
NusaX-senti-inst (Winata et al., 2023b)	12		ace, ban, bjn, bug, eng, ind, jav, mad, min, nij, sun, bbc	219	22	Apache 2.0	Sentiment Analysis
Persian-instruct-pn (Farahani et al., 2021; Shafagh, 2023a,b)	1	pes	pes	1713	128	MIT	Summarization
SCB-MT-2020-prompt (Lowphansirikul et al., 2022; PyThaiNLP, 2023a,b)	2	tha, eng	tha, eng	181	127	CC BY-SA 4.0	Translation
Scirepeval-biomimicry-inst (Singh et al., 2022)	1		eng	996	523	ODC-BY	Scientific Document Representation
Seed-instruct-lij (Maillard et al., 2023; ConseggioLigure, 2023c,d)	2	lij, eng	lij, eng	184	186	CC BY-SA 4.0	Translation
SODA-inst (Kim et al., 2022)	1		eng	412	328	CC BY 4.0	Dialogue
TamilStories (AI Tamil Nadu, 2023a)	1	tam	tam	2266	2172	Apache 2.0	Generation
TeluguRiddles (desik98, 2023)	1	tel	tel	74	44	Apache 2.0	Question Answering
Thai-USEmbassy-prompt (PyThaiNLP, 2023e,f)	2	tha, eng	tha, eng	2131	2077	CC0 1.0	Translation
Thai-POS-inst (PyThaiNLP, 2023c)	1	tha	tha	72	36	CC BY-SA 3.0	Generation
Thai-Wiktionary-inst (PyThaiNLP, 2023d)	1	tha	tha	35	147	CC BY-SA 3.0	Generation
Thirukkural-instruct (AI Tamil Nadu, 2023b)	1	tam	tam	133	542	Apache 2.0	Generation
Turku-paraphrase-inst (Kanerva et al., 2021; TurkuNLP, 2023)	1	fin	fin	108	59	CC BY-SA 4.0	Paraphrase Identification
UA-Gec-inst (Syvokon et al., 2023; osyvokon, 2023)	1	ukr	ukr	192	148	CC BY 4.0	Generation
UNER-LLM-inst (Mayhew et al., 2023; Universal NER, 2023)	11	zho, hrv, dan, eng, deu, por, rus, srp, slk, swe, tgl	zho, hrv, dan, eng, deu, por, rus, srp, slk, swe, tgl	768	109	CC BY-SA 4.0	Named Entity Recognition

Urdu-News-Gen-Article (Hussain et al., 2021; AhmadMustafa, 2023a)	1	urd	urd	109	1313	CC BY 4.0	Generation
Urdu-News-Category-Class (Hussain et al., 2021; AhmadMustafa, 2023b)	1	urd	urd	1407	43	CC BY 4.0	Text Classification
Urdu-News-Gen-Headline (Hussain et al., 2021; AhmadMustafa, 2023c)	1	urd	urd	1314	94	CC BY 4.0	Generation
Wiki-split-inst (Botha et al., 2018)	1		eng	200	166	CC BY 4.0	Text Simplification
X-CSQA-inst (Lin et al., 2021)	16		eng, zho, deu, spa, fra, ita, jpn, nld, pol, por, rus, arb, vie, hin, swa, urd	197	21	MIT	Question Answering
Xlel_wd-inst (Pratapa et al., 2022)	44			379	190	CC BY 4.0	Event Linking
XWikis-inst (Perez-Beltrachini and Lapata, 2021)	4		ces, fra, eng, deu	5662	346	MIT	Summarization

Table E.7: List of datasets in **Aya** Collection (templated datasets).

Dataset	#Langs	$\bar{L}_{prompt}$	$\bar{L}_{compl.}$	License	Task
Adversarial QA (T) (Bartolo et al., 2020)	101	159	721	CC BY-SA 3.0	Question Answering
CNN-Daily-Mail (T) (See et al., 2017) (Hermann et al., 2015)	101	1980	305	Apache 2.0	Summarization
Flan-Coqa (T) (Wei et al., 2022a; Reddy et al., 2019)	101	2143	364	Multiple*	Question Answering
Flan-CoT-submix (T) (Wei et al., 2022a)	101	239	160	Unknown	Generation
Flan-GEM-wiki-lingua (T) (Wei et al., 2022a; Ladhak et al., 2020)	101	1732	572	CC BY-NC-SA 3.0	Summarization
Flan-lambada (T) (Wei et al., 2022a; Paperno et al., 2016)	101	232	7	CC BY 4.0	Generation
Flan-unified-QA (T) (Wei et al., 2022a; Khashabi et al., 2020)	101	281	13	Apache 2.0	Question Answering
HotpotQA (T) (Yang et al., 2018)	101	129	15	CC BY-SA 4.0	Question Answering
Joke-explaination-inst (T) (theblackcat102, 2023)	101	111	545	MIT	Generation
Mintaka-inst (T) (Sen et al., 2022)	101	54	67	CC BY 4.0	Question Answering
MLQA-en (T) (Lewis et al., 2020)	101	819	20	CC BY-SA 3.0	Question Answering
NQ-Open (T) (Kwiatkowski et al., 2019)	101	68	14	CC BY-SA 3.0	Question Answering
PAWS-Wiki (T) (Zhang et al., 2019)	101	308	6	Custom license, attribution	Paraphrase Identification
PIQA (T) (Bisk et al., 2020)	101	304	100	Unknown	Question Answering
SODA-inst (T) (Kim et al., 2022)	101	86	208	CC BY 4.0	Dialogue
WIKI QA (T) (Yang et al., 2015)	101	205	36	MSR DLA*	Question Answering
Wiki-split-inst (T) (Botha et al., 2018)	101	126	220	CC BY-SA 4.0	Text Simplification
Xlel_wd-inst (T) (Pratapa et al., 2022)	101	300	274	CC BY 4.0	Event Linking
Dolly-v2 (T) (Conover et al., 2023)	101	427	357	CC BY-SA 3.0	Generation

Table E.8: List of datasets in **Aya** Collection (translated datasets).

Lang	Prompt	Completion
fra	Qui a écrit le livre La Sagouine?	Antonine Maillet a écrit La Sagouine en 1971. Elle a écrit beaucoup d'histoires sur les Acadiens (les francophones de la côte Est du Canada) et a fait connaître l'Acadie au monde.
ibo	Kedu ka ndị Igbo si eme emume ji ọhụrụ?	Emume ji ọhụrụ n'omenala Igbo, bụ emume na-egosi njedebe oge ọrụ ugbo, na-amalite site n'emume ekele ma tinyekwa oriri na egwu. Ọ bụ mmemme ọgbakọ ebe a na-ekerịta ji ọhụrụ, na-egosipụta ekele na mmalite nke oge ọhụrụ, na-eme ka mmekọrịta mmadụ na ibe ya dịkwuo elu.
por	Qual é a capital do estado da Bahia?	A capital do estado da Bahia, no Brasil, é Salvador. Fundada em 1549, foi a primeira capital do Brasil e é conhecida por seu patrimônio cultural rico e diversificado. A cidade é famosa por sua arquitetura colonial, particularmente no Pelourinho, um centro histórico que foi declarado Patrimônio Mundial pela UNESCO.
msa	Apakah nasi lemak?	Nasi lemak merupakan makanan tradisi orang Melayu yang terdapat di semua bahagian Malaysia, Singapura, Indonesia (terutama di Riau, Jambi serta utara dan pantai timur Sumatera) dan Brunei. Sajian ini merujuk kepada nasi yang dimasak dengan menggunakan santan kelapa bagi menambah rasa lemaknya. Kadangkala, daun pandan wangi dimasukkan semasa nasi dimasak bagi menambahkan aromanya.

Table E.9: Examples of prompt and completion from the **Aya** Dataset.

F Additional Figures

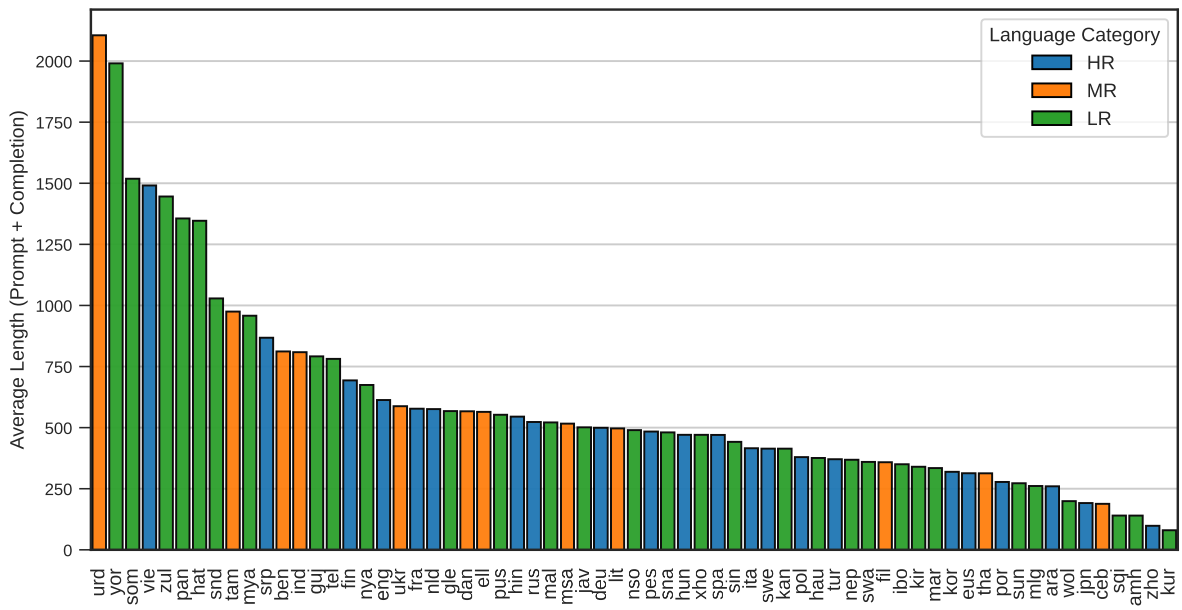

Figure F.6: Average prompt and completion length across different languages in **Aya Dataset**.

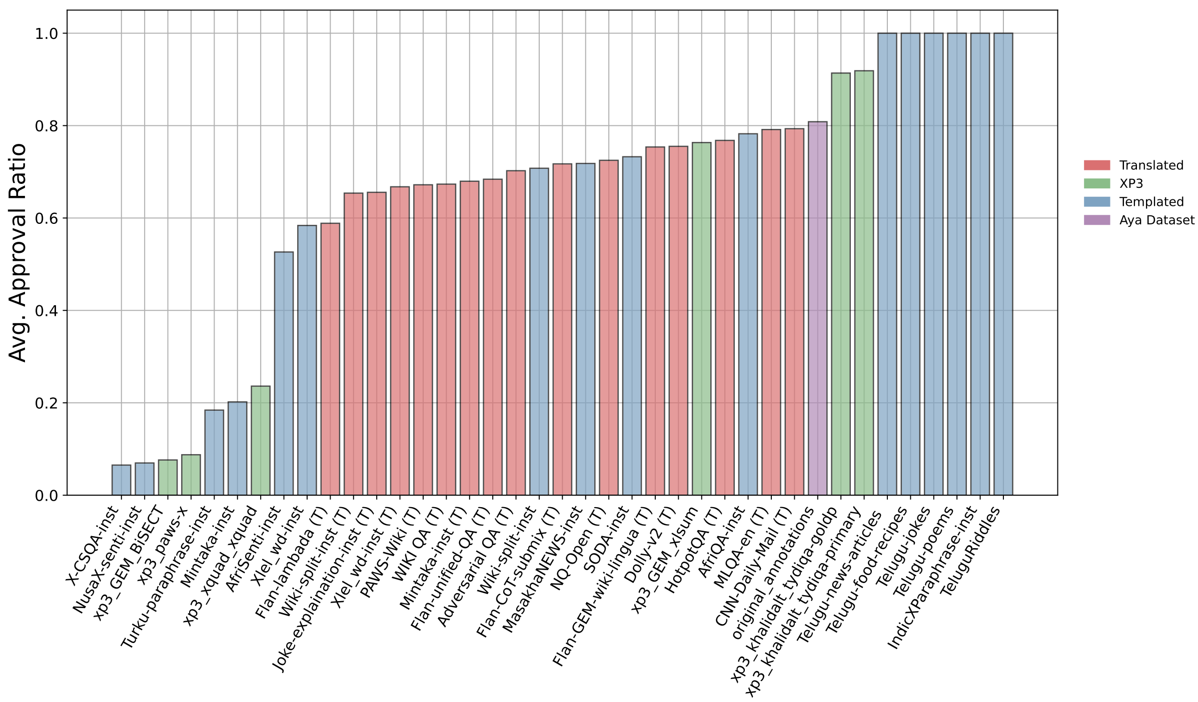

Figure F.7: Average Approval Ratio per dataset, constrained to datasets receiving at least 20 votes.

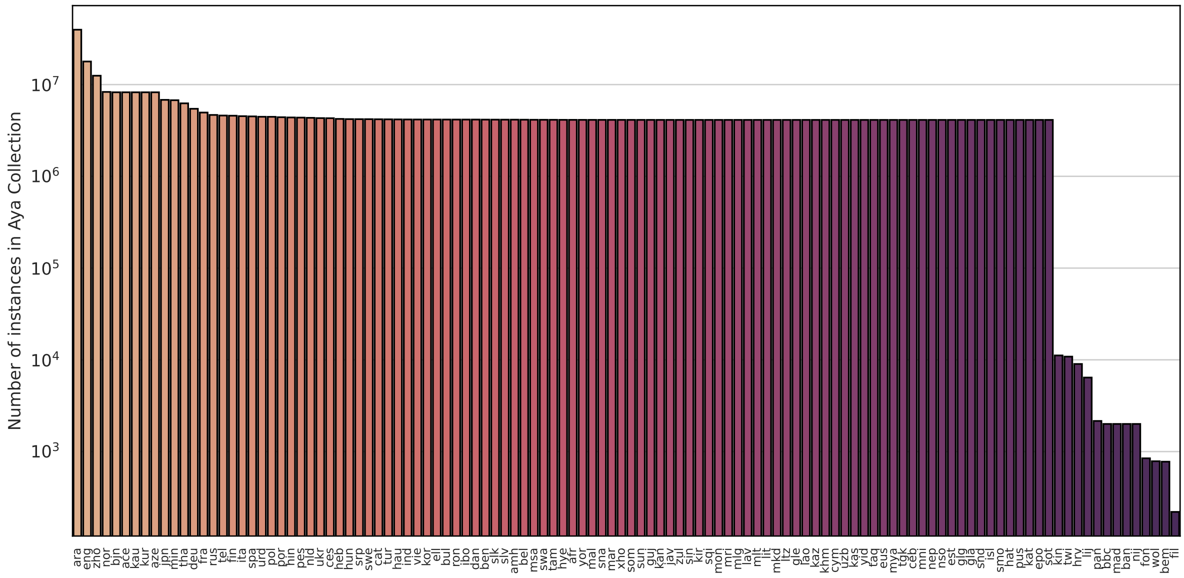

Figure F.8: Number of prompt/completion pairs in each language in the **Aya** Collection. Y-axis is in log-scale.

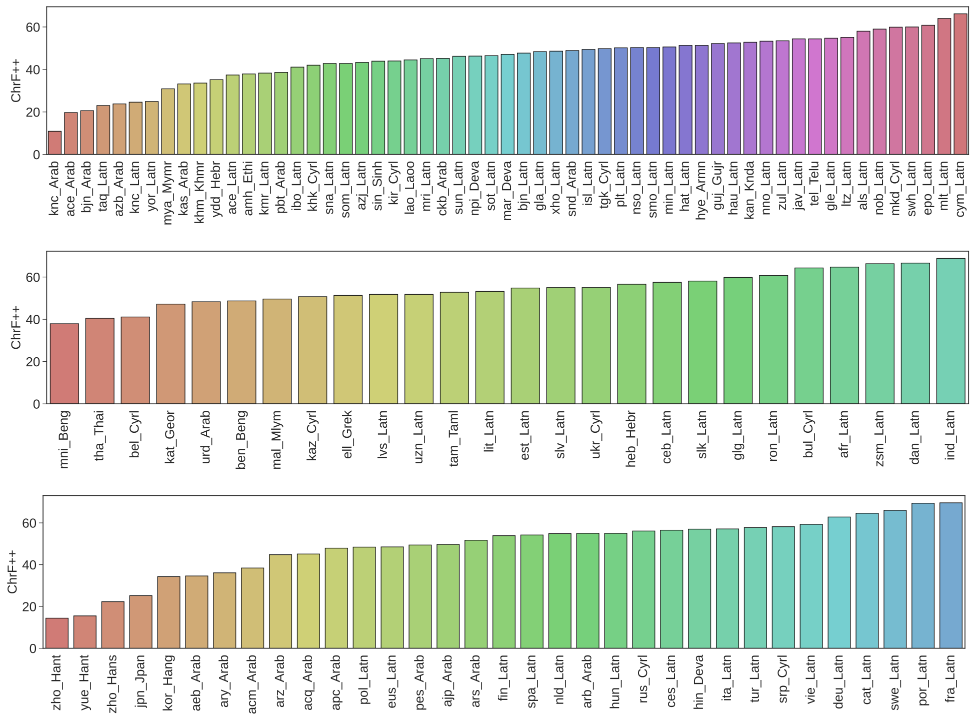

Figure F.9: NLLB Translation Quality: ChrF++ scores on FLORES for translations from English into the **Aya** target languages that are covered by NLLB, grouped by their resourcedness according to (Joshi et al., 2020). **Top**: Low-resource Languages, classes 0, 1, 2; **middle**: Mid-resource Languages, class 3; **bottom**: High-resource Languages, classes 4, 5, following (Joshi et al., 2020)

.

G Language Representation via Community

G.1 Division by Regions

We chose to divide languages into four primary regions: *Africa, Asia, Europe,* and *Latin America*. These four regions were established in order to facilitate the administration of user contributions and were not intended to prescribe boundaries within which certain languages are exclusively spoken.

The language statistics by region are as follows: *Africa* (14 languages), *Asia* (41 languages), *Europe* (42 languages), and *Latin America* (4 languages). Almost all the languages were assigned to a region but there are some exceptions, Maori and Samoan were unassigned to any specific region as they didn't align with the predefined regions. English was left unassigned, serving as a common language across all regions. Additionally, contributions in Spanish and Portuguese were distributed between *Europe* and *Latin America* based on contributors' countries. Similarly, Arabic contributions were shared between *Africa* and *Asia* depending on the contributors' country of origin. Additional dialects of Arabic were included in regions separate from that of their parent language because we had a significant number of speakers from these regions eager to contribute to their respective dialects. Each region had at least one "Regional Lead" responsible for coordinating "Language Ambassadors," and for recruiting fluent speakers for the languages within their area.

G.2 Language Ambassadors

The Language Ambassador's role was pivotal in bridging the gap between the data collected in a language and its speakers. An essential criterion for selection was native fluency in the specific language. The Language Ambassador's expertise in specific languages and familiarity with the cultures of the language speakers was invaluable. They assisted not only in spreading awareness among participants but also in identifying and addressing potential data issues specific to each language, such as languages incorrectly assigned to their region. Their cultural and linguistic insights enabled them to make informed decisions, like choosing suitable data sources for collection in their respective languages. Not every language had a designated Language Ambassador, and some had more than one. In total, we had 84 Language Ambassadors over the course of the initiative. Their combined efforts played a vital role in broadening the contributor base for each language. Support for the Language Ambassadors' progress and trouble-shooting challenges they faced was coordinated asynchronously and through weekly online meetups, discussed in Appendix G.3.1 and Appendix G.3.2.

G.2.1 Regional Leads

There were a total of six Regional Leads: two for Latin America, one for Africa, one for Asia, and two for Europe. The selection for Regional Lead roles was on a voluntary basis, with the only requirement being that they must originate from the regions they intended to lead. The invitation for this role was specifically extended to individuals who were already actively participating in our community projects or engaged in other open science projects. Regional Leads had several key roles throughout the project, such as selecting Language Ambassadors and aiding their efforts in attracting more annotators and maintaining their engagement.

G.3 Communication

G.3.1 Platforms

We established a Discord server for coordination between Regional Leads, Language Ambassadors and annotators. The server provided basic channels for internal communications: introductions, inquiries, and announcements, as well as specific channels for Language Ambassadors, for each region, and for each language, along with any other channels that proved useful for the particular region. For external communications, we used social media platforms (e.g., X, LinkedIn, WhatsApp, Facebook), recognizing that the choice of communication platform varied based on cultural and regional preferences. Using multiple platforms not only facilitated internal organization but also broadened our project's outreach by providing flexible and inclusive means of outreach to diverse communities and audiences.

G.3.2 Meetings

In addition to asynchronous communication through Discord, we conducted meetings to maintain team collaboration and cohesion:

Regional Leads and Language Ambassadors Meeting. A weekly meeting in which Regional Leads and Language Ambassadors shared project updates, exchanged ideas, and addressed questions from Language Ambassadors. It served as an excellent platform for gathering ideas from Language Ambassadors and brainstorming new strategies to engage annotators effectively.

New Contributor Introduction Meeting. Held weekly, this meeting aimed to introduce new contributors to the project's specifics. It included explanations about the motivations behind the project, a walk-through of the **Aya** UI, and a sharing of regional statistics. Additionally, this meeting provided examples of both good and bad annotations and edits to guide new annotators in their work. It concluded with a synchronous challenge for the annotators to submit a few initial annotations in real time to familiarize them with the process and allow them to ask questions if they got stuck.

Regional Leads Meeting. Held bi-weekly, this meeting brought together Regional Leads to assess progress, discuss upcoming steps, and provide advice on how to engage and sustain contributions for their respective regions. Furthermore, this meeting facilitated collaborative troubleshooting efforts and helped make important decisions for the following week.

Technical Update. This meeting was dedicated to sharing technical updates, with a focus on recent UI progress, data, and benchmarking. The purpose of this monthly update was to ensure all team members and annotators were well-informed about the project's current status and upcoming priorities. It was a place for open discussion to hear feedback from everyone interested in the project.

Language Specific Meeting. Held weekly or bi-weekly, these meetings were co-working sessions or datathons led by the language ambassadors with their respective annotators to submit annotations synchronously. It also acted as an onboarding session to welcome new contributors from regions that could not join the New Contributor Introduction Meeting due to conflicting time zones. Demonstrations on using the UI, as well as brainstorming sessions, were conducted to improve the representation of specific languages in the project.

H Data Cards

Following Pushkarna et al. (2022) and the HuggingFace data card template[13], we present the data card for the **Aya** Dataset.

Data Card for the Aya Dataset

The **Aya** Dataset is a multilingual instruction fine-tuning dataset curated by an open-science community. The dataset contains a total of 204,114 annotated prompt-completion pairs.

- Curated by: 2,007 contributors from 110 countries
- Language(s): 65 languages
- License: Apache 2.0
- Repository: https://huggingface.co/datasets/CohereForAI/aya_dataset

Authorship

Publishing Organization:	Industry Type:	Contact Details:
Cohere For AI	Not-for-profit - Tech	https://aya.for.ai/

Example of Data Points

The dataset contains multilingual prompts and completions in the following format:

```
{
  "inputs": "What cultural events or festivals...",
  "targets": "Colombo's cultural calendar is adorned...",
  "language": "English",
  "language_code": "eng",
  "annotation_type": "original-annotations",
  "user_id": "f0ff69570af705b75c5a0851883e..."
}
```

Motivations & Intentions

Curation Rationale: The curation effort employed an open-science approach to create a diverse instruction-style dataset through annotators across the globe that ensures comprehensive representation across all languages. The success of the curation effort, led by volunteers across diverse backgrounds, was significantly influenced by their hope to meaningfully bring NLP advancements to their languages.

Provenance

Methods Used
Crowd-sourced through volunteer annotations, followed by a quality assessment phase in which samples from the dataset were checked.

Methodology Details
Source: Original annotations and edits of open-source NLP datasets
Platform: **Aya** Annotation Platform
Dates of Collection: Jun 2023 - Dec 2023

Dataset Version and Maintenance

Maintenance Status	Version Details	Maintenance Plan
Actively Maintained	Current version: 1.0 Last Update: 12/2023 First Release: 02/2024	Updates will be periodically made available based on volunteer contributions

[13] https://huggingface.co/docs/datasets/v2.15.0/en/dataset_card

Data Card for the Aya Collection

The **Aya** Collection incorporates instruction-style templates from fluent speakers and applies them to a curated list of 44 datasets. It also includes translations of 19 instruction-style datasets into 101 languages. This collection provides 513,579,625 instances of prompts and completions covering a wide range of tasks.

- Curated by: 2007 contributors from 110 countries
- Language(s): 114 languages
- License: Apache 2.0
- Repository: https://huggingface.co/datasets/CohereForAI/aya_collection

Authorship

Publishing Organization: Cohere For AI

Industry Type: Not-for-profit - Tech

Contact Details: https://aya.for.ai

Example of Data Points

The dataset contains multilingual prompts and completions in the following format:

```
{
  "id": 246001,
  "inputs": "The following query in English is taken from..",
  "targets": "The answer is Mount Lucania.",
  "dataset_name": "Mintaka-inst",
  "sub_dataset_name": "-",
  "task_type": "question-answering",
  "template_id": 3,
  "language": "eng",
  "split": "train",
  "script": "Latn"
}
```

Motivations & Intentions

Curation Rationale: Automatic augmentation of existing datasets serves to enhance the available linguistic resources for multiple languages. List of languages were established from mT5 and aligned with annotators' language list and NLLB translation model. The datasets were translated directly from English for all languages.

Provenance

Methods Used
Combination of crowd-sourced templating and automatic translation.

Methodology Details
Source: Existing NLP datasets
Platform: **Aya** Annotation Platform
Dates of Collection: Jun 2023 - Dec 2023

Dataset Version and Maintenance

Maintenance Status
Actively Maintained

Version Details
Current version: 1.0
Last updated: 12/2023
Release date: 02/2024

Maintenance Plan
No updates planned.

Data Card for the Aya Evaluation Suite

The **Aya** Evaluation Suite contains a total of 25,750 open-ended conversation-style prompts covering 114 languages of three subsets:
AYA-HUMAN-ANNOTATED: 250 original human-written prompts in 7 languages each.
DOLLY-MACHINE-TRANSLATED: 200 human-selected prompts from (Conover et al., 2023), automatically translated with the NLLB model (NLLB-Team et al., 2022) from English into 101 languages.
DOLLY-HUMAN-EDITED: 200 DOLLY-MACHINE-TRANSLATED prompts post-edited by fluent speakers for 6 languages.

- Curated by: contributors, professional annotators, and synthetic generation
- Language(s): 101 languages
- License: Apache 2.0
- Repository: https://huggingface.co/datasets/CohereForAI/aya_evaluation_suite

Authorship

Publishing Organization:
Cohere For AI

Industry Type:
Not-for-profit - Tech

Contact Details:
https://aya.for.ai

Example of Data Points

The dataset contains multilingual prompts in the following format. Note that 'source_id' is applicable only for subsets DOLLY-MACHINE-TRANSLATED and DOLLY-HUMAN-EDITED. Furthermore, the 'target' field is not applicable for DOLLY-HUMAN-EDITED.

```
{
  "id": 2,
  "inputs": "How to escape from a helicopter trapped in water ?",
  "targets": "If you are ever trapped inside a helicopter...",
  "language": "eng",
  "script": "Latn",
  "source_id": 6060
}
```

Motivations & Intentions

Curation Rationale: This evaluation suite is tailored for testing the generation quality of multilingual models, with the aim to balance language coverage and human-sourced quality. It covers prompts originally written in each language, as well as English-centric translated and manually curated or edited prompts for a linguistically broad but rich testbed. The list of languages was established from mT5 and aligned with annotators' language list and the NLLB translation model.

Provenance

Methods Used
Combination of original annotations by volunteers, automatic translation, and post-editing of translations by professional annotators.

Methodology Details
Source: Original annotations and translations and post-edits of Dolly
Platform: **Aya** Annotation Platform
Dates of Collection: Jun 2023 - Dec 2023

Dataset Version and Maintenance

Maintenance Status
Actively Maintained

Version Details
Current version: 1.0
Last updated: 02/2024
Release date: 02/2024

Maintenance Plan
No updates planned.