MURI: High-Quality Instruction Tuning Datasets for Low-Resource Languages via Reverse Instructions

Abdullatif Köksal[1, 2, 4], Marion Thaler[1], Ayyoob Imani[1, 2], Ahmet Üstün[3], Anna Korhonen[4], Hinrich Schütze[1, 2]

[1]CIS, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany
[3]Cohere For AI, The Netherlands [4]Language Technology Lab, University of Cambridge, UK
akoksal@cis.lmu.de

Abstract

Instruction tuning enhances large language models (LLMs) by aligning them with human preferences across diverse tasks. Traditional approaches to create instruction tuning datasets face serious challenges for low-resource languages due to their dependence on data annotation. This work introduces a novel method, Multilingual Reverse Instructions (MURI), which generates high-quality instruction tuning datasets for low-resource languages without requiring human annotators or pre-existing multilingual models. Utilizing reverse instructions and a translation pipeline, MURI produces instruction-output pairs from existing human-written texts in low-resource languages. This method ensures cultural relevance and diversity by sourcing texts from different native domains and applying filters to eliminate inappropriate content. Our dataset, MURI-IT, includes more than 2 million instruction-output pairs across 200 languages. Evaluation by native speakers and fine-tuning experiments with mT5 models demonstrate the approach's effectiveness for both NLU and open-ended generation. We publicly release datasets and models at https://github.com/akoksal/muri.

1 Introduction

Instruction tuning refines large language models (LLMs) based on user intentions, enhancing their ability to generalize across tasks and align with human preferences (Ouyang et al., 2022; Sanh et al., 2022; Muennighoff et al., 2023; Wang et al., 2022). While pre-training data can be automatically collected from the web, preparing instruction tuning data is challenging as it requires aligned instruction-output pairs. Three main approaches have been applied to create instruction tuning datasets: human annotation (Ouyang et al., 2022; Köpf et al., 2023; Conover et al., 2023), templatized NLP tasks (Sanh et al., 2022; Wang et al., 2022; Longpre et al., 2023a), and synthetic data generation via LLMs (Wang et al., 2023; Honovich et al., 2023).

For low-resource languages, these approaches face serious limitations. Human annotation is costly, and finding native speakers for low-resource languages is challenging. Templatizing NLP tasks restricts datasets to specific structures and domains, limiting their general applicability, and there is insufficient NLP task-annotated data for low-resource languages (ImaniGooghari et al., 2023). Synthetic data generation is constrained by the languages supported by existing models and suffers from validity (Wang et al., 2023) and creativity (Honovich et al., 2023) issues. Moreover, outputs of both template-based and synthetic data generation methods heavily rely on translation pipelines and are particularly prone to artifacts known as ''translationese'' (Gellerstam, 1986). These artifacts include simplified vocabulary and grammar, unidiomatic word order and expressions, and often neglect linguistic and cultural contexts. Such occurrences of translationese have been shown to negatively impact model training (Yu et al., 2022) by distorting examples and further distancing them from their linguistic and cultural context.

Consequently, no existing model, open source or proprietary, supports low-resource languages with the quality necessary for high-quality instruction tuning dataset generation. This has resulted in a disparity, with English dominating 73% of the most popular datasets (Longpre et al., 2024), leaving low-resource language communities underserved.

In this work, we introduce Multilingual Reverse Instructions (MURI), a novel approach to generate instruction tuning datasets

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1032–1055, 2025. https://doi.org/10.1162/tacl.a.18
Action Editor: Zeljko Agic. Submission batch: 8/2024; Revision batch: 3/2025; Published 8/2025.

for low-resource languages without requiring annotators, task-annotated data, or pre-trained multilingual models. MURI employs the reverse instructions method proposed by Köksal et al. (2024) and combines it with machine translation to develop language-specific target instructions, i_τ, for a target text d_τ. This involves translating d_τ to English d_{eng}, (ii) generating an English instruction i_{eng} for d_{eng} using reverse instructions, and (iii) translating i_{eng} back to the target language, so it can serve as instruction for d_τ. Notably, unlike fully translation-based methods, our approach requires translating only the instructions which enables using authentic outputs in the target language. The instruction-output pair (i_τ, d_τ) can then be used to fine tune a language model to follow instructions. This approach is cost-effective and applicable to any language with available textual data.

By applying MURI to texts from diverse sources, we have created MURI-IT, a dataset containing more than 2 million instruction-output pairs across 200 languages. To our knowledge, this dataset offers the broadest language coverage for multilingual instruction tuning. Our sources include Wikipedia, WikiHow, and various web-crawled pages, providing a rich variety in style, domain, and length. The output documents, sourced directly from data in their original languages, retain cultural and linguistic nuances. Additionally, we use quality filters to ensure the dataset's high standards.

To evaluate MURI-IT, native speakers across 13 languages assess the dataset on five aspects to gauge quality. We also fine-tune several mT5 family models using MURI-IT to execute instruction-based tasks, assessing their performance in natural language understanding and generation. For instance, MURI-101, an mT5-XXL model instruction-tuned with MURI-IT, outperforms prior models like mT0 (Muennighoff et al., 2023) by over 14% in multilingual MMLU. In open-ended generation tasks, it delivers much better outputs than mT0, with win rates of 59% vs. 28%. Additionally, MURI-IT enhances performance when used alongside existing datasets like Aya. We make the fine-tuned models and MURI-IT publicly available.

To summarize, our contributions are:

(i) We introduce Multilingual Reverse Instructions (MURI), a cost-effective method for creating multilingual instruction tuning datasets applicable to hundreds of languages.

(ii) We create and publish MURI-IT, an instruction tuning dataset for 200 languages using MURI. This dataset consists of 2,228,499 instruction-output pairs, with 64% of the data from low-resource languages.

(iii) We evaluate and analyze the dataset with native speakers in 13 languages. We find that the data is highly idiomatic in many languages.

(iv) We fine-tune and release MURI-101, a massively multilingual instruction-following language model using MURI-IT.

2 Related Work

Instruction Tuning Datasets Instruction tuning has emerged as a powerful approach to enhancing the instruction-following capabilities of LLMs, as demonstrated by numerous studies (Ouyang et al., 2022; Sanh et al., 2022; Muennighoff et al., 2023; Wang et al., 2022). The three primary strategies to create instruction tuning datasets are human curation, templatized tasks, and synthetic generation via LLMs.

Human-curated datasets, like Open Assistant (Köpf et al., 2023) and Dolly (Conover et al., 2023), involve extensive human annotation but are difficult to scale and extend to more languages due to high cost. Alternatively, datasets such as Public Pool of Prompts (P3) (Sanh et al., 2022), Super-Natural Instructions (NIv2) (Wang et al., 2022), and FLAN (Longpre et al., 2023a) utilize NLP task reformulation to reduce cost and enhance applicability but still struggle with general-purpose instruction following since their main focus is on NLU tasks.

To address these issues, synthetic datasets have been developed, such as Self-Instruct (Wang et al., 2023), DoG-Instruct (Chen et al., 2023), and Unnatural Instructions (Honovich et al., 2023). These datasets offer greater task diversity but are challenged by issues of validity and creativity. The reverse instructions method (Köksal et al., 2024), employing data augmentation via generative models and pretraining corpora, further exemplifies cost-effective dataset generation. A similar method has been successfully applied in Bactrian-X (Li et al., 2023), demonstrating its effectiveness in multilingual settings by leveraging translation for synthetic data.

Multilingual instruction tuning has shown substantial benefits, especially for low-resource languages (Muennighoff et al., 2023). It not only maintains performance in English but also enhances capabilities in non-English languages with the help of a large scale of English examples (Shaham et al., 2024; Grattafiori et al., 2024). Despite these advancements, large-scale multilingual datasets often remain limited. Efforts to overcome this include pre-training on diverse multilingual data (Chung et al., 2024; Chowdhery et al., 2023) and creating dedicated multilingual instruction fine-tuning sets (Muennighoff et al., 2023), and manual annotation by native speakers (Singh et al., 2024). Extensions to existing datasets often utilize automatic translation (Li et al., 2023; Winata et al., 2023; Holmström and Doostmohammadi, 2023) and template-based generation (Gupta et al., 2024). These methods strive to balance diversity against resource constraints and quality issues inherent in automated translation processes, as seen in the extensive Aya collection (Singh et al., 2024). In summary, while instruction tuning has greatly advanced the capabilities of LLMs in following complex instructions, challenges remain in dataset diversity, validity, and the integration of multilingual content.

Multilingual LLMs In the recent surge of LLMs, English-centric models like the closed-source GPT model family (Radford et al., 2019; Ouyang et al., 2022) and open-sourced ones like Llama and Pythia (Touvron et al., 2023a; Biderman et al., 2023) have gained prominence. Multilingual models, unlike monolingual ones, offer the advantage of facilitating cross-lingual tasks such as translation (Jiao et al., 2023; Xu et al., 2024) and addressing low-resource languages (Artetxe and Schwenk, 2019; Wu and Dredze, 2020). mBERT pioneered the multilingual area by demonstrating that training on multilingual data allows different languages to be represented in a unified semantic space (Devlin et al., 2019). Building on this foundation, subsequent models such as XLM-R, GLoT500, and AfriBERTa extended the capabilities of transformer-based models to hundreds of languages (Conneau et al., 2020; ImaniGooghari et al., 2023; Ogueji et al., 2021).

However, the progression of multilingual encoder-decoder and decoder-only models has been more restrained. Models like Llama and Mistral, for instance, feature datasets predominantly composed of English, with limited data from a select group of high-resource languages (Touvron et al., 2023a; Jiang et al., 2023). In contrast, models like XGLM, BLOOM, and MGPT have been developed from scratch to support extensive language diversity (Workshop et al., 2022; Lin et al., 2022; Shliazhko et al., 2024). Meanwhile, mT5 is trained on 101 languages, an important step in encoder-decoder model training (Xue et al., 2021).

3 The MURI-IT Dataset

We introduce MURI-IT, which includes 2,228,499 instruction-output pairs in 200 languages. The dataset is primarily constructed by applying MURI to the CulturaX (Nguyen et al., 2024) and Wikipedia corpora. The core idea is summarized in Figure 1. Our goal is to utilize existing high-quality human-written multilingual corpora to generate a diverse instruction-following dataset. For a randomly selected text, we aim to generate an instruction for which the high-quality corpus text would serve as a good response. This approach ensures that a model trained with this dataset will not be conditioned on outputs with translation artifacts or culturally irrelevant topics.

MURI-IT also incorporates two additional subsets: 54,578 instances collected from the WikiHow website in 18 languages to augment the dataset, and 455,472 existing NLP task examples in 74 languages to enrich its diversity. This section details the steps used to produce MURI-IT.

3.1 Multilingual Reverse Instructions (MURI)

In this subsection, we describe the process of creating an instruction-following dataset from multilingual corpora using MURI. First, we select documents written in the target low-resource language τ from multilingual corpora (§3.1.1), producing a set $D_\tau = d_\tau^1, d_\tau^2, \ldots, d_\tau^n$. Our goal is to generate instructions for which each of these documents would serve as a suitable answer. Since recent LLMs have limited multilingual capabilities, we combine machine translation models with English LLMs.

We begin by translating each document d_τ^k into English, yielding d_{eng}^k (§3.1.2). Next, we apply reverse instructions using an English LLM to generate instructions that align with d_{eng}^k, producing i_{eng}^k (§3.1.3). Finally, we translate these English instructions back into the original

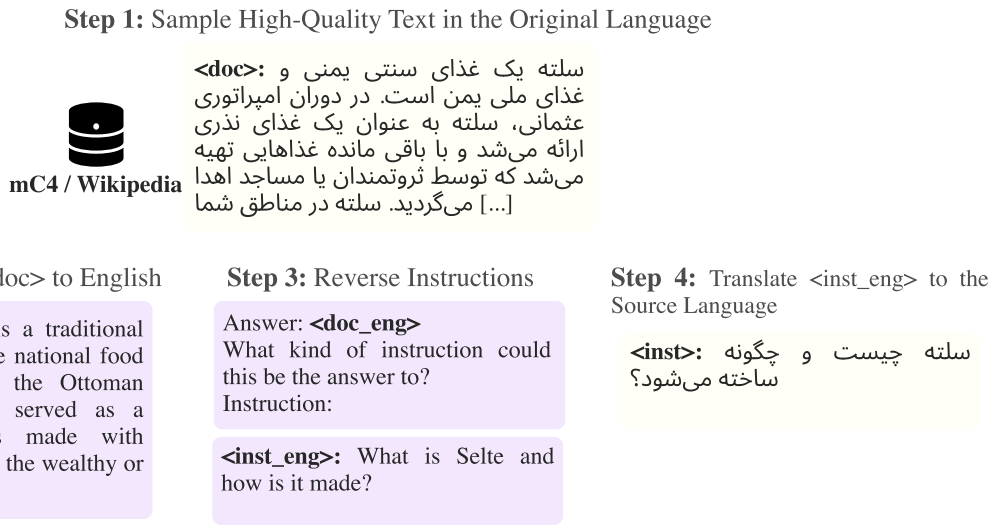


Figure 1: **Multilingual Reverse Instructions** (MURI). Step 1: MURI selects a high-quality human-written example (<doc>) from multilingual corpora. Step 2: Translation into the English document <doc_eng>. Step 3 applies the reverse instructions method to <doc_eng> (i.e., prompting the LLM to generate a matching instruction <inst_eng>). Step 4: <inst_eng> is translated back into the original language (<inst>), resulting in a (<inst>, <doc>) pair where the <doc> output is human-written.

language τ (§3.1.4). This results in the instruction-output pairs $D_{I\tau} = (i^1_\tau, d^1_\tau), (i^2_\tau, d^2_\tau), \ldots, (i^n_\tau, d^n_\tau)$, which form our multilingual instruction tuning dataset.

3.1.1 Data Selection

We randomly sample documents from two multilingual corpora: CulturaX (1,076,575 documents) and Wikipedia (1,554,207 documents). CulturaX covers 167 languages, merging OSCAR (Ortiz Suárez et al., 2020) and mC4 (Xue et al., 2021), enhanced with cleaning, deduplication, language identification, and diversification steps (Nguyen et al., 2024). The Wikipedia corpus spans more than 350 languages and provides high-quality documents.

3.1.2 Document Translation

Each document d^k_τ is first translated into English, yielding d^k_{eng}. We use the MADLAD-400-3B-MT model (Kudugunta et al., 2023) since it supports 400 languages. We use nucleus sampling with top_p $= 1$.

Table 1 summarizes the translation performance of MADLAD-400-3B-MT on Flores-200 (NLLB Team et al., 2022) for the languages we focused on. Translations into English generally achieve high chrF scores across language of 45.5, the others exceed 50, with resource levels 3–5 surpassing 60. The BLEU metric results are reported in Appendix §10.2.

Resource Level	to English	from English
0 (9 langs.)	45.5	28.8
1 (75 langs.)	51.5	40.4
2 (17 langs.)	55.0	41.3
3 (23 langs.)	63.0	52.4
4 (16 langs.)	64.1	58.8
5 (6 langs.)	63.3	53.9

Table 1: Average chrF scores of the MADLAD-400-3B-MT model evaluated on Flores-200, grouped by language resource level as defined in Joshi et al. (2020). Levels range from 0 (low-resource) to 5 (high-resource). Results represent 146 out of the 194 languages included in the reverse instructions subset of MURI-IT.

3.1.3 Reverse Instructions

We next apply an English LLM to generate an instruction i^k_{eng} for each English-translated document d^k_{eng}. We adapt the ‘‘reverse instructions’’ approach of Köksal et al. (2024) to a few-shot setup using open models to produce high-quality instructions. A few-shot example of our prompt is presented in the Appendix (Table 11).

To determine which LLM performs best for reverse instruction generation, we systematically compare multiple models of different sizes in a few-shot in-context learning setup. Each model is evaluated on the Dolly dataset

Model	ROUGE-L	BLEU	WR
Llama 2-7B-Chat	0.400	0.107	0.555
Llama 2-13B-Chat	0.422	0.127	0.632
Llama 2-70B-Chat	0.427	0.129	0.676
Mistral-7B-Instruct v0.1	0.420	0.122	0.615
Mixtral-8×7B	**0.441**	**0.144**	**0.697**

Table 2: Performance of various LLMs on English reverse instructions, evaluated using ROUGE-LSUM, BLEU, and WINRATE (WR) metrics. Mixtral-8×7B achieves the highest scores on all metrics. We report additional ablation studies on sampling strategies, number of few-shot examples, and prompt design in Appendix §10.1.

(Conover et al., 2023), which consists of human-annotated instruction-output pairs. We measure the similarity between the model's generated instructions and the gold standard instructions using ROUGE-LSUM and BLEU metrics. Additionally, we use an LLM-as-a-Judge approach using Llama-3-70B-Instruct to compare the generated instructions against the gold standard. We report the percentage of cases where the model's instructions are preferred over or considered equal to the gold standard, which we refer to as the WINRATE (WR) metric.

Table 2 summarizes the results. We find that Mixtral-8×7B (Jiang et al., 2024) offers the best performance, outperforming various Llama-2 (Touvron et al., 2023b) and Mistral-7B (Jiang et al., 2023) models in generating instructions for a given output. We observe that around 70% of the generated instructions via Mixtral-8×7B are either preferred over or equal to the human-annotated high-quality instructions, which illustrates the quality of synthetically generated instructions in reverse instructions.

Our ablation experiments in Appendix §10.1 show that using four in-context examples, combined with a greedy decoding strategy, provides a higher performance than other settings. Finally, to evaluate whether using different prompts affects the quality and diversity of instructions, we experiment with alternative prompt templates and find that they do not result in significant differences (§10.3).

3.1.4 Instruction Translation

Next, we translate each English instruction i^k_{eng} back into its original language τ, resulting in i^k_τ. We again use MADLAD-400-3B-MT. As shown in Table 1, translating from English into other languages yields lower overall chrF scores compared to translating into English. While resource level 0 languages have a notably low average chrF of 28.8, resource levels 3 through 5 generally perform better, achieving chrF scores exceeding 50.

Interestingly, resource level 5 exhibits lower performance than resource level 4. This decline is attributed to the significantly weaker translation quality for Japanese and Chinese within resource level 5, which achieve chrF scores of 34.3 and 35.7, respectively. Therefore, these lower scores reduce the average performance of resource level 5 as a whole. We observe similar patterns in BLEU scores (see Table 10 in the Appendix).

Despite these limitations, only the instructions (and not the outputs) rely on translating from English. By preserving human-written texts in the target language as outputs, we avoid the more severe artifacts that can arise from fully translated approaches. To ensure language consistency, we apply GlotLID (Kargaran et al., 2023) to detect language mismatches, discarding any instances where instruction and document language IDs do not align.

3.1.5 Content Screening

We observe that some examples contain violent or noisy content because mC4 may include unscreened examples from CommonCrawl (Abadji et al., 2022). We utilize the RoBERTa hate-speech model (Vidgen et al., 2021) to screen the generated instruction-output pairs (i^k_{eng}, d^k_{eng}) in English and eliminate unsuitable examples. To ensure dataset integrity and eliminate redundancy, we employ the MinHashLSHForest method for deduplication.

Manual screening revealed unsuitable instructions lacking necessary context, such as instructions asking for summarization of non-existent prior documents or requesting translations. We excluded these instruction-output pairs from our dataset by filtering out instructions including the words *summarize* or *translate*. Additionally, we observed that web-sourced documents often include extraneous content like footers, headers, or advertisements. We leave the elimination of such extraneous content for future work.

3.2 WikiHow Data

We collected articles from the multilingual WikiHow website using PyWikiHow in 18 languages

(Arabic, Chinese, Czech, Dutch, English, French, German, Hindi, Italian, Japanese, Korean, Malay, Portuguese, Russian, Spanish, Thai, Turkish, and Vietnamese), based on the URLs provided by Wikilingua (Ladhak et al., 2020). Each WikiHow page is composed of the following sections: (i) A *title* that starts with ''How to'', (ii) an *abstract* answer to the question, (iii) a number of *steps*, each comprising a *step-title* and a *step-text* paragraph. We use the *title* of each WikiHow page as the instruction. To introduce variation in the style of the answers, we render the answers to the questions as follows: In 50% of cases, we include the *abstract* in the answer and in the other 50% we don't. Regardless of whether the *abstract* is included or not, in 50% of the cases we only include the *step-titles*, and in the other 50% we include both the *step-titles* and *step-texts*.

3.3 NLP Tasks

To further improve diversity of tasks in MURI-IT, we incorporate several existing multilingual instruction following datasets based on NLP tasks, expanding language coverage and task diversity. These additions, totaling 455,472 samples across 74 languages, complement our primary data sources:

SuperNatural Instructions: We sampled 200 tasks from SuperNatural Instructions (Wang et al., 2022) per translation task type and 500 samples from the remaining set, resulting in 161,986 samples across 55 languages.

xP3 comprises 46 languages across 16 NLP tasks, such as various types of QA and Topic Classification. We adapted 184,000 samples of xP3 (Muennighoff et al., 2023) to our format.

OASST1: From this crowd-sourced chat-style dataset (Köpf et al., 2024) spanning 35 languages, we selected and paired message and output pairs up to the second deepest level, yielding 9,486 examples in 10 languages.

FLAN v2: We incorporated 100,000 samples from FLAN v2 (Longpre et al., 2023a), including 50,000 from its main collection and 50,000 from its Chain-of-Thought subset, all in English following Tulu (Ivison et al., 2023).

These additional datasets were chosen to increase the linguistic and task diversity of MURI-IT, ensuring a more comprehensive and versatile instruction-tuning dataset similar to prior work like Aya (Singh et al., 2024). We have summarized the statistics of our dataset in Table 3.

Source	# Languages	# Examples
Multilingual Reverse Instructions	194	1,718,449
Wikipedia	187	1,031,726
CulturaX	123	686,723
WikiHow	18	54,578
NLP Tasks	74	455,472
SupNatInst-v2	55	161,986
xP3	44	184,000
OpenAssistant	10	9,486
FLAN v2.0	1	100,000
Total	**200**	**2,228,499**

Table 3: Composition of MURI-IT by source, including the number of languages and examples. We split the dataset 90/5/5 into training, validation, and test sets, ensuring similar ratios for sources and languages.

4 Dataset Analysis

This section presents an in-depth analysis of MURI-IT, focusing on three aspects: instruction diversity, linguistic diversity, and data quality assessment. Section 4.1 analyzes the most common instructions generated by our pipeline, accompanied by illustrative examples. Section 4.2 examines the range of languages represented in MURI-IT, highlighting its coverage of low-resource languages, diverse scripts, word orders, and case-marking systems.

Section 4.3 evaluates the dataset's effectiveness in maintaining linguistic and cultural accuracy. We detail the findings from a quality evaluation involving native speakers, assessing alignment, correctness, and overall informational sufficiency.

4.1 Instruction Diversity

We highlight the most common instructions in MURI-IT. Following Wang et al. (2023), we use the Berkeley Neural Parser (Kitaev and Klein, 2018) to identify each verb and its direct object (noun+verb: ''describe history'') or the auxiliary and its dependent verb (auxiliary+verb: ''do make'') for English instructions, before translating them into the source language.

Figure 2 illustrates 17% of the most frequent pairs in the reverse instruction subset of MURI-IT. These include tasks such as writing articles or biographies, describing historical events, listing specific headlines, and telling stories. We also examined how these same verb-object pairs appear across different languages. Notably, the diversity

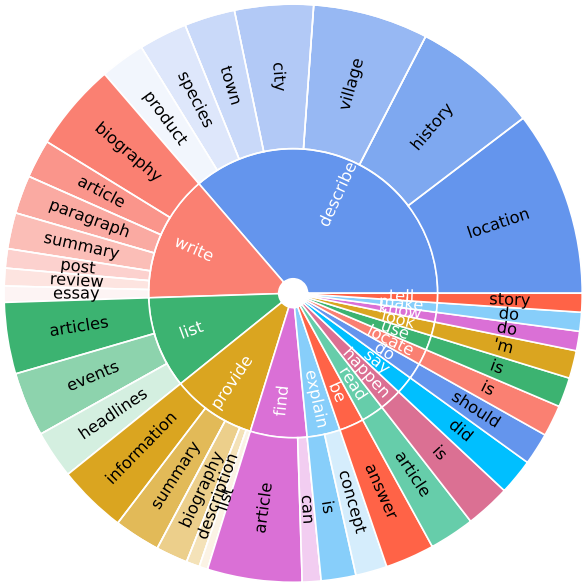


Figure 2: Frequency of the most common noun+verb and auxiliary+verb pairs of generated instructions in a portion of the reverse instructions subset of MURI-IT.

exists even within the same instruction type. For example, for *describe+history*, each language features culturally relevant topics. In Arabic, one finds ‘‘Describe the history of the Al-Azhar Mosque’’; in German, ‘‘Describe the history of the Große Federlhof’’; and in Japanese, ‘‘Describe the history and techniques of Japanese embroidery.’’ We further investigate the impact of prompt selection on instruction diversity in the Appendix (see §10.3) and observe that the model produces similar instructions for different types of prompts.

4.2 Languages and Linguistic Diversity

MURI aims to provide a methodology inclusive of low-resource languages through a culturally respectful approach, utilizing materials in their native languages and avoiding outputs in translationese (Bizzoni et al., 2020; Vanmassenhove et al., 2021). Given that the majority of languages used in NLP systems share typological similarities and geographical origins (Joshi et al., 2020), this often leads to an uneven distribution of resources and tools available to the global community. MURI-IT therefore focuses particularly on languages with limited resources and diverse features.

Joshi et al. (2020) outlined a taxonomy categorizing languages based on their resource levels, ranging from 0 (*left-behinds*) such as Balinese with severely limited resources, to 5 (*winners*) like English or French. Our dataset encompasses a large number of low-resource languages, as shown in Figure 3.a, with over 700,000 examples falling into category 1. Despite this, access to outputs for these low-resource languages remains limited, with 33 languages containing fewer than 1,000 examples each. Nonetheless, MURI-IT proves to be one of the most diverse instruction-tuning datasets to date.

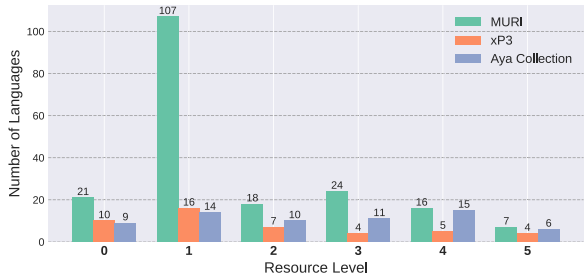
Bagheri Nezhad and Agrawal (2023) emphasize script and word order as important factors in analyzing linguistic diversity. While the majority of languages in MURI-IT employ the Latin script or a combination of Latin, Arabic, and Cyrillic scripts (Figure 3.b), a notable portion (more than one-fifth, categorized as ‘‘Other’’) features low-resource scripts such as Lao or Georgian. As the output texts have not been translated, the use of these scripts is idiomatic, ensuring correct orthography.

To further investigate linguistic diversity, we examined word order and case marking. Focusing on the order of subject, verb, and object, Figure 3.c shows that while European SVO languages predominate (Dryer, 2013) and there are no rare OVS and OSV languages, all frequent patterns are represented. This showcases the ‘‘structural’’ diversity of our dataset.

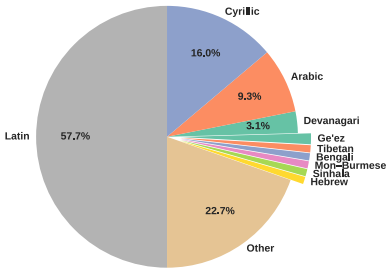
Case-marking patterns align with geographical distribution; e.g., mid-size to large inventories are prevalent in South Asia, Eastern Europe, and east-central Africa (Iggesen, 2013). Figure 3.d illustrates that our dataset encompasses a diverse range of case systems, including complex systems with up to ten cases. This indicates that our dataset has good coverage of both ‘‘analytic’’ and ‘‘synthetic’’ languages. Overall, case marking and word order exemplify the broader coverage of MURI-IT of less common languages compared to previous datasets, contributing to a more comprehensive representation of linguistic diversity in NLP resources.

4.3 Quality Assessment of MURI-IT

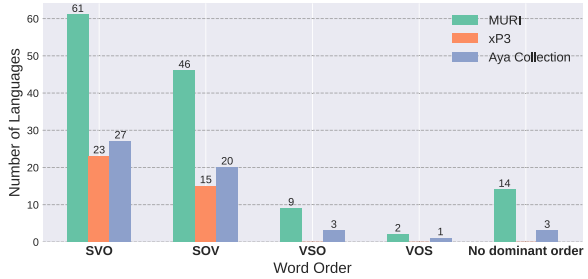
A distinctive feature of MURI-IT is its preservation of cultural and linguistic nuances, often lost in translated datasets. To enhance our linguistic analysis, we conducted a thorough evaluation of a random subset of the dataset, involving native speakers proficient in 13 languages. Each annotator examined 100 randomly selected instruction-output pairs from the reverse instruction subset



(a) Resource levels of languages in MURI-IT, xP3, and Aya datasets (per Joshi et al. (2020)).



(b) Percentage of languages in MURI-IT by script.



(c) Word order in MURI-IT, xP3, and Aya datasets (per Dryer (2013)).

(d) Case marking in MURI-IT, xP3, and Aya datasets (per Iggesen (2013)).

Figure 3: Linguistic diversity of MURI-IT compared to Aya (Singh et al., 2024) and xP3 (Muennighoff et al., 2023) datasets, highlighting differences in (a) resource level (Joshi et al., 2020), (b) script, (c) word order, and (d) case marking (Dryer and Haspelmath, 2013). The classifications by Joshi et al. (2020) and WALS are extensive but not exhaustive, thus not covering all languages in the datasets.

Language	Alignment ↑	Inst. Correctness ↑	Output Correctness ↑	Informational Sufficiency ↑	Prop. of Non-Instructions ↓
Bavarian	0.73	0.74	0.80	0.86	0.60
Chinese	0.66	0.99	0.80	0.85	0.11
Dutch	0.84	0.99	0.77	0.84	0.08
English	0.74	1.00	0.91	0.81	0.04
French	0.82	0.94	0.76	0.77	0.06
German	0.84	0.90	0.77	0.74	0.05
Italian	0.82	0.98	0.90	0.90	0.03
Korean	0.55	0.94	0.91	0.81	0.03
Persian	0.70	0.94	0.92	0.84	0.07
Swedish	0.79	0.98	0.76	0.64	0.02
Turkish	0.70	0.98	0.91	0.92	0.02
Ukrainian	0.69	0.88	0.76	0.69	0.06
Vietnamese	0.85	0.94	0.77	0.74	0.05
Avg.	0.75	0.94	0.83	0.80	0.09

Table 4: Comparison of alignment, proportion of non-instructive phrasing, grammatical correctness, and informational sufficiency across 13 evaluated languages in the reverse instructions subset of MURI-IT. Arrows indicate whether higher (↑) or lower (↓) is better.

of MURI-IT using five predefined evaluation criteria. These criteria assess the quality of both instructions and outputs using (except for Proper Instruction Format) a Likert scale. (i) **Alignment** (range 1–5): Measures the alignment between instruction and output. (ii) **Instruction Correctness**, (iii) **Output Correctness** (range 1–5): Assesses lexical and grammatical accuracy of instruction and output. (iv) **Informational Sufficiency** (range 1–5): Determines whether the instruction can be adequately answered without external context. (v) **Proper Instruction Format** (0: No, 1: Yes): Indicates whether the instruction is appropriately formatted for a language model. All Likert-based criteria (1–5) were then normalized to a 0–1 scale, where a rating of 1 corresponds to 0, and 5 corresponds to 1. This normalization ensures better comparability across different criteria.

Table 4 shows that human assessment is generally good, but with some mixed results.

Issue	Lang.	Instruction & Output
No Orthographic Standard: Standard German words in the instruction. Bavarian has no standard orthography, resulting in translation models mixing different types of Bavarian and Standard German.	Bavarian	I: Beschreibn Sie die Geschichte von Pädaschduaf. A: Pädaschduaf oda Pädasduaf (kuaz: P'dorf), offiziö Perchtoldsdorf, is a Moaktgmoa mit Einwohna (Stand) im Bezirk Mödling in [...]
Code-switching: The instruction is in simplified Mandarin Chinese, while the response is in a Wu dialect using traditional Chinese characters.	Chinese	I: 描述澳大利亚斯蒂尔克里克。 A: Steels Creek勒到澳大利亞維克多利亞州個東北方向45公里，勒勒環墨爾本城市個綠帶浪，接近马龙大公路（The Maroonda Highway），国王湖（Kinglake）搭仔圣安德鲁斯（St Andrews），老辰光風景末交關美，是週末弗出遠門就能白相白相格好地方。[...]
Idiomaticity: The expression *Chronisch-Übersättigung* is incorrect (indicating shortcomings of the translation model).	German	I: Was ist chronische Übersäuerung des Körpers? Was bedeutet die Chronisch-Übersättigung des Organismus? A: Die chronische übersäuerung des Körpers - Die chronische übersäuerung des körpers-Tagebücher Wenn der Körper [...]
Lack of specific product information: The output lacks specific information how to use *TimeTool* for birthday celebration.	Dutch	I: Hoe kan ik mijn verjaardag vieren met TimeTools? A: TimeTools heeft voor elk tijdzegel een uniek werkboek onwikkeld. Veel mensen ervaren dit als een 'feest der herkenning'.
Superfluous Text Markup: The output contains header details, inhibiting readability: *mandy quotes billy and mandy |*	English	I: What is Clickfunnels? A: mandy quotes billy and mandy | What is Clickfunnels? mandy quotes billy and mandy Prices: Is it truly worth it? The Internet [...]

Table 5: Examples of issues identified in the human evaluation of MURI-IT, illustrating various challenges such as orthographic inconsistencies, code-switching, idiomatic inaccuracies, and superfluous text markup. These examples highlight areas where translation and formatting may impact the overall quality and clarity of the dataset.

High-resource languages such as English, German, French and Italian consistently perform well across all criteria. Common problems include lack of specific product information and temporal instructions (e.g., recent news), which are not helpful in instruction tuning. Across all 13 languages, extraneous headers, footers, and metadata are found in some outputs. Thus, the noise contained in the underlying multilingual corpora affects quality and coherence of MURI-IT, as reflected in the slightly lower average output correctness compared to instruction correctness. A relatively minor problem is less idiomatic and culturally appropriate language use. Table 5 provides representative examples of these observed issues, including instances of translation inconsistencies, as seen in the German example.

For lower-performing languages, a major source of error is the lack of standardization. For instance, Bavarian—spoken in Austria, Bavaria, and Alto Adige—lacks a standard. This Aresulted in MADLAD (Kudugunta et al., 2023) translations including Standard German words (Table 5), degrading the quality of generated instructions. Similarly, Chinese instructions and outputs were sometimes mismatched, e.g., inconsistent use of traditional vs. simplified Chinese and of dialects vs. Standard Mandarin.

Overall, we observe a moderately high alignment between instructions and outputs, averaging 0.75. Only 9% of the generated instructions deviate from the typical style of a question or direct instruction. Instruction-output pairs are mostly grammatically and lexically accurate, with higher-performing languages such as English and German aligning particularly well. This directly follows from the superior performance of MADLAD for these languages.

5 Experimental Setup

To evaluate the effectiveness of MURI-IT, we instruction-tune mT5-XXL (Xue et al., 2021). While recent autoregressive models exist with stronger results in English, mT5 remains one of the most comprehensive models supporting numerous languages. We fine-tune using a subset of MURI-IT for the 101 languages supported

	arb	ben	cat	dan	deu	eus	fra	guj	hin	hrv	hun	hye	ind	ita	kan	mal
Okapi	27.7	26.8	30.5	31.8	31.7	27.9	30.7	27.4	26.5	30.0	30.1	27.5	27.5	30.4	26.8	25.8
mT0	31.5	31.6	32.8	33.0	32.7	29.7	32.1	29.5	32.0	31.1	32.3	28.4	33.3	32.4	30.9	28.6
mT0x	31.6	30.2	32.6	32.0	32.5	29.2	32.7	28.5	31.6	31.1	31.7	26.7	32.3	31.3	28.9	26.7
Aya-101	38.2	35.8	39.6	39.7	39.7	36.0	39.7	33.6	38.7	37.5	38.8	30.0	40.0	39.0	34.5	30.4
MURI-101 (ours)	36.5	33.0	38.8	38.4	38.9	34.4	39.0	33.1	35.4	37.0	38.1	29.9	38.9	38.5	32.4	30.9
	mar	**nep**	**nld**	**por**	**ron**	**rus**	**slk**	**spa**	**srp**	**swe**	**tam**	**tel**	**ukr**	**vie**	**zho**	**Avg.**
Okapi	26.1	25.2	31.1	30.1	30.9	30.6	30.2	30.9	30.4	29.3	26.0	25.9	31.6	27.5	28.2	28.8
mT0	31.6	32.4	32.0	32.1	32.4	32.8	32.3	32.1	30.9	31.6	29.4	29.0	31.5	30.9	32.5	31.5
mT0x	29.7	30.1	32.1	32.0	31.8	31.7	31.4	32.2	31.4	32.8	27.7	27.9	32.3	31.1	31.6	30.8
Aya-101	36.0	37.2	40.1	39.0	39.5	39.2	39.4	39.7	38.1	39.7	31.2	32.1	39.9	34.8	38.3	37.3
MURI-101 (ours)	33.0	33.2	38.8	38.1	38.1	37.7	38.0	39.0	36.6	38.5	29.8	31.3	37.0	36.8	36.9	36.0

Table 6: Multilingual MMLU performance of Okapi, mT0, mT0x, Aya-101, and MURI-101 across 31 languages. Scores are accuracy in a few-shot setup. Except for Okapi (25 shots), the number of shots is 5.

by mT5, called MURI-101. Our evaluation encompasses both multilingual Natural Language Understanding (NLU) and open-ended generation (NLG).

5.1 Baselines

We compare our MURI-101 model against four state-of-the-art multilingual instruction-following models.

mT0 (Muennighoff et al., 2023): An mT5-XXL-based model, instruction-tuned using the xP3 dataset, which consists of 16 reformulated NLP tasks, including summarization, QA, and classification for 46 languages.

Okapi (Lai et al., 2023): A series of language-specific instruction-following models based on Bloom-7b (Workshop et al., 2022) and Llama-7b (Touvron et al., 2023a). Each model is independently fine-tuned on translations of English synthetic data, followed by preference optimization for a specific language.

mT0x: An mT5-XXL model instruction-tuned using the extended xP3 dataset, xP3x, covering 101 languages (Üstün et al., 2024), providing a fair comparison regarding the number of languages.

Aya-101 (Üstün et al., 2024): Uses xP3x, translated Aya Collection, a subset of Data-Provenance (Longpre et al., 2024) and translated ShareGPT-Command to instruction-tune mT5-XXL for 101 languages.

5.2 Training Details

Our experiments utilize the TPU Research Cloud (TRC) program, employing a TPU v4-32 with 32 chips and the T5X framework from Google. We set both input and output lengths to 1024 tokens and implement data packing. The effective batch size is 64, achieved through gradient accumulation (batch size of 8 with 8 accumulation steps). Following Üstün et al.'s (2024) findings, we use a fixed learning rate of 3e-4 without a scheduler and trained for 5 epochs. For generation tasks, we apply nucleus sampling with top_p $= 0.8$ and temperature $= 0.9$, as per Holtzman et al. (2020).

5.3 Evaluation Benchmarks

We evaluate the models in both multilingual and monolingual settings for NLU and open-ended generation tasks. Two evaluations use **TranslatedDolly** (Singh et al., 2024), a translated version of Dolly (Conover et al., 2023), a human-annotated English instruction-tuning dataset.

Multilingual Settings. *NLU*: Multilingual MMLU (Lai et al., 2023), created by translating the English MMLU dataset to 31 languages. We evaluate using the lm-evaluation-harness framework (Gao et al., 2024) with a 5-shot setup.
NLG: TranslatedDolly, evaluated on 21 languages using the multilingual Command R+ (Cohere For AI, 2024) model as an LLM judge.

Monolingual Low-Resource Settings. *NLU*: Taxi1500 (Ma et al., 2023) for classification with a 6-shot setup based on a parallel Bible corpus covering 1500 languages.
NLG: TranslatedDolly.

6 Multilingual Model Evaluation

We first evaluate our model MURI-101 on the few-shot multilingual MMLU task. Table 6 shows that MURI-101 clearly outperforms previous models (Okapi, mT0, mT0x), with an average relative improvement of more than **14.3%** (from 31.5 to 36.0). MURI-101 consistently outperforms prior models across all languages, with the exception of Aya-101.

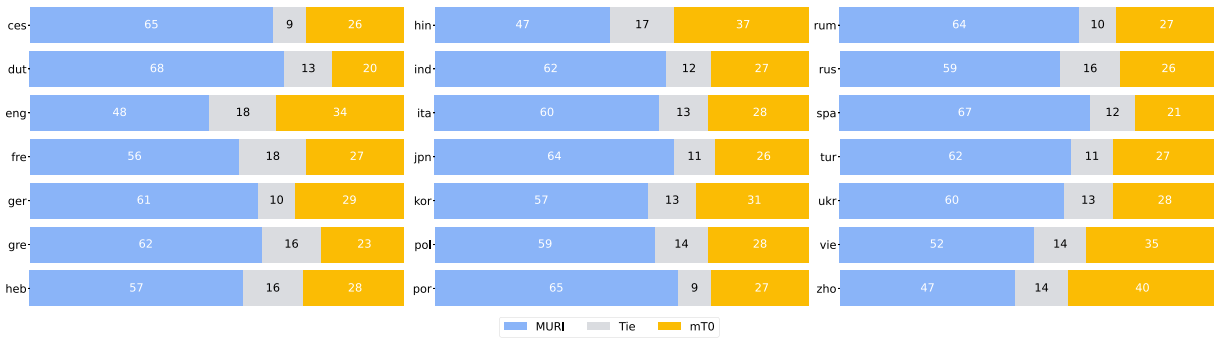


Figure 4: Win rates MURI-101 vs. mT0 outputs across 21 languages on TranslatedDolly, the translated Dolly dataset. Win rates are determined by Command R+ as judge. The average win rates are 59% and 28% for MURI-101 and mT0, respectively.

While Aya-101 shows slightly better performance than MURI-101 in NLU, we note that Aya-101 is the result of a computationally heavy training process involving around 25 million samples. This includes a lot of translated data (47.5% of the training mixture) and data synthetically generated and translated based on the ShareGPT dataset using a proprietary model (22.5% of the training data). Thus, around 60% of their training data relies on translation which may introduce systematic translation artifacts known as translationese (Gellerstam, 1986; Yu et al., 2022) in the model outputs. However, this effect is difficult to evaluate with current metrics. Given these factors, we primarily compare MURI-101 with mT0 in NLG.

For NLG evaluation, we compare MURI-101 with mT0 on TranslatedDolly (Singh et al., 2024) and compare outputs using the multilingual Command R+ model as a judge. From TranslatedDolly, we select the 21 languages that Command R+ supports. Figure 4 shows that MURI-101 consistently outperforms mT0 across all languages. Also across all languages, MURI-101's win rate against mT0 is 59%, with lose and tie rates of 28% and 13%.

The lowest improvement in NLG is for simplified Chinese, with a 47% win rate vs. 40% loss rate. We hypothesize that code-switching within different varieties of Chinese (as discussed in §4.3) contributes to this limited improvement.

7 Monolingual Evaluation in Low-Resource Setting

To evaluate the capabilities of MURI-IT and Aya in low-resource settings, we conduct an additional set of experiments with only monolingual training. We first select ten low-resource languages: Azerbaijani, Kazakh, Lao, Khmer, Welsh, Scottish Gaelic, Belarusian, Bulgarian, Slovenian, and Slovak. While available in Aya, these languages are not part of the human-annotated portion of Aya and only have examples via translation, thus possibly lacking in cultural context and idiomaticity. We test in our experiment how well MURI-IT complements translated content in this setting. Furthermore, the languages were chosen to represent diverse language families: Turkic, Tai-Kadai, Austroasiatic, Celtic, and Slavic.

For this low-resource scenario, we sample at most 15K examples from both Aya and MURI-IT. Then we instruction-tune mT5-XXL for each language and for MURI-IT, Aya and Aya+MURI-IT separately, resulting in MURI_1, Aya_1, and Aya_1+MURI_1 models.

Since many of these languages are not supported by multilingual MMLU and Command R+, we use the few-shot classification task Taxi1500 (Ma et al., 2023) for NLU. For NLG, we use TranslatedDolly; however, we translate model outputs to English (via Google Translate) and calculate win rates with Llama-3-70B-Instruct of translated outputs vs. Dolly's gold English human outputs.

Table 7 presents the NLU results, where accuracy is reported using a 6-shot setup. We observe that MT5 achieves an average of 19.7% accuracy, which is only slightly above random chance (16.7%). AYA_1 obtains 35.1%, demonstrating a substantial gain over MT5. Although AYA_1 already shows strong performance, MURI_1 achieves 36.3%, and the combination (AYA_1+MURI_1) further improves the average accuracy to 37.2%. These results demonstrate that MURI-IT can be competitive with, or even outperform, the Aya dataset at a similar scale, and it also complements Aya in low-resource scenarios.

Language	aze	bel	bul	cym	gla	kaz	khm	lao	slk	slv	Avg.
mT5	20.4	22.4	20.7	18.4	19.3	19.8	16.5	21.3	19.2	18.9	19.7
AYA_1	37.0	32.1	34.4	33.0	28.7	44.7	30.0	32.7	38.1	**40.3**	35.1
MURI_1	37.6	**35.2**	34.4	33.1	34.7	39.2	**33.5**	**36.8**	**42.2**	36.3	36.3
AYA_1+MURI_1	**39.5**	33.7	**38.1**	**35.5**	**35.2**	**46.7**	31.3	33.0	39.1	39.6	**37.2**

Table 7: Monolingual NLU performance on the Taxi1500 classification task across different low-resource languages. Scores are accuracy using a 6-shot setup.

	AYA_1	MURI_1	AYA_1+MURI_1		AYA_1	MURI_1	AYA_1+MURI_1
aze	4%	0%	4%	**kaz**	2%	0%	3%
bel	3%	1%	6%	**khm**	2%	1%	4%
bul	6%	0%	7%	**lao**	3%	2%	1%
cym	2%	0%	4%	**slk**	4%	1%	2%
gla	2%	2%	4%	**slv**	6%	0%	3%

Table 8: Win rate comparison of AYA_1, MURI_1, and AYA_1+MURI_1 vs. gold human outputs across different low-resource languages in NLG. Averages are 3.4%, 0.7%, and 3.8%, respectively.

Table 8 shows that, on average, the win rate of MURI_1 is 0.7%, AYA_1 is 3.4%, and AYA_1+MURI_1 is 3.8%. These low numbers indicate that the models do not produce consistently high-quality outputs in these low-resource languages. While both Aya-101 and MURI-101 demonstrate better NLG performance than prior multilingual instruction-tuning models such as mT0, current models are still limited in their NLG capabilities for low-resource languages and with smaller training datasets.

We hypothesize that the limitations of our base model, mT5, make it hard to achieve large improvements in NLG for low-resource languages. As recent autoregressive models begin to support a larger number of languages, we anticipate that MURI-IT, with its human-written outputs, will be used effectively to improve NLG performance for low-resource languages.

8 Conclusion

This study presents Multilingual Reverse Instructions (MURI), a novel approach for generating high-quality instruction tuning datasets for low-resource languages. Our method addresses limitations of translation-focused multilingual datasets by using human-written texts as outputs, combined with a translation pipeline and LLMs to create contextually appropriate instructions. The resulting dataset, MURI-IT, of more than 2 million pairs across 200 languages greatly expands the resources available for multilingual language models.

Evaluation by native speakers from 13 languages confirmed the dataset's quality and idiomaticity. Our instruction-tuned mT5-XXL model, MURI-101, strongly outperformed previous models on NLU and NLG both multi- and monolingually. Notably, incorporating MURI-IT improved performance for most low-resource languages, effectively complementing existing datasets like Aya.

While challenges remain, particularly in NLG for low-resource languages, MURI-IT represents an important step towards more inclusive and linguistically diverse language models. Future work will focus on refining data quality and leveraging advanced multilingual models to further improve performance across languages.

9 Limitations

Despite the promising results obtained, several limitations must be acknowledged in this study. First, we did not perform clustering—in contrast to Köksal et al. (2024)—due to uncertainties regarding the performance of multilingual encoders. Clustering could potentially enhance content diversity, ensuring a greater variety of linguistic and cultural contexts.

Additionally, the quality of the data can be further improved through more rigorous cleaning such as the removal of headers and footers from documents. Similarly, the MURI methodology, particularly for low-resource languages, would benefit from more standardized source data. Our evaluation, involving native speakers, noted deficits in languages with less standardized orthography or prominent regional dialects. Additional preprocessing could address this issue.

Addressing these limitations in future work will involve integrating advanced clustering

algorithms, enhancing data cleaning protocols, and expanding the dataset to include a wider range of languages.

Acknowledgments

This work was funded by Deutsche Forschungsgemeinschaft (project SCHU 2246/14-1). Anna Korhonen acknowledges the support of UK EPSRC grant EP/T02450X/1. Abdullatif Köksal was additionally supported by the ELSA Mobility Program under grant agreement no. 101070617. We thank our annotators Trung Bui, Tianshi Feng, Anastasia Gutsol, Tobias Linner, Laura Niemann, Haotian Ye, Esmanur Yılmaz, and Raoyuan Zhao for their invaluable contributions.

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.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. https://doi.org/10.1162/tacl_a_00288

Sina Bagheri Nezhad and Ameeta Agrawal. 2023. Exploring the maze of multilingual modeling. *arXiv e-prints*, arXiv:2310.05404v2.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yuri Bizzoni, Tom S. Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.34

Yongrui Chen, Haiyun Jiang, Xinting Huang, Shuming Shi, and Guilin Qi. 2023. DoG-Instruct: Towards premium instruction-tuning data via text-grounded instruction wrapping. *arXiv e-prints*, arXiv:2309.05447v2.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(1).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H.

Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Cohere For AI. 2024. c4ai-command-r-plus-08-2024. https://doi.org/10.57967/hf/3135

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

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. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. [Last accessed: 2025-04-01]

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer. 2013. Order of subject, object and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo. https://doi.org/10.5281/zenodo.7385533

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.12608602

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng

Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, and Tobias Speckbacher. 2024. The Llama 3 herd of models. *arXiv e-prints*, arXiv:2407.21783v3.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2024. TarGEN: Targeted data generation with large language models. In *First Conference on Language Modeling*.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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. https://doi.org/10.18653/v1/2023.acl-long.806

Oliver A. Iggesen. 2013. Number of cases (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.61

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing LM adaptation with Tulu 2. *arXiv e-prints*, arXiv:2311.10702v2.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*, arXiv:2310.06825v1.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary,

Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv e-prints*, arXiv:2401.04088v1.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1001

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. https://doi.org/10.18653/v1/2020.acl-main.560

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.410

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1249

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. 2024. LongForm: Effective instruction tuning with reverse instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7056–7078, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.414

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul E. S., Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. https://doi.org/10.18653/v1/2020.findings-emnlp.360

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 318–327, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-demo.28

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv e-prints*, arXiv:2305.15011v2.

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.

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, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8):975–987. https://doi.org/10.1038/s42256-024-00878-8

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv e-prints*, arXiv:2305.08487v1.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.891

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

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. *arXiv e-prints*, arXiv:2207.04672v3.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.11

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.156

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv e-prints*, arXiv:2203.02155v1. https://doi.org/10.48550/arXiv.2203.02155

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog.

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*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.136

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79. https://doi.org/10.1162/tacl_a_00633

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.620

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *arXiv e-prints*, arXiv:2302.13971v1.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints*, arXiv:2307.09288v2.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.845

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.188

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.132

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.754

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. 2022. 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. https://doi.org/10.18653/v1/2022.emnlp-main.340

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. 2023. 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. https://doi.org/10.18653/v1/2023.eacl-main.57

BigScience Workshop: Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Alshaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, and Tim Dettmers. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv e-prints*, arXiv:2211.05100v4.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.repl4nlp-1.16

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.40

10 Appendices

10.1 Ablation of Reverse Instructions

Reverse instructions aim to generate instructions for a given document/output. We have already determined the most effective LLM in Table 2. We now investigate different setups, including the number of few-shots, sampling strategies, and prompt designs.

Number of Few-shots. We explore different numbers of few-shot examples (`fs` = 1,2,4). Table 9 (*A*) shows that increasing the number of shots generally improves performance. In particular, using four in-context examples consistently achieves the highest ROUGE-LSUM and BLEU scores.

Sampling Strategies. Next, we vary the temperature (T) and nucleus sampling parameter (p) to investigate different decoding settings. As shown in Table 9 (*B*), greedy decoding (`T` = `0`) achieves the best overall scores compared to other settings, such as T = 1.0.

(A) Number of few-shots		
Few-Shot (fs)	**ROUGE-LSum**	**BLEU**
1	0.440	0.129
2	0.414	0.124
4	**0.441**	**0.144**

(B) Sampling Strategies (fs = 4)		
Decoding Settings	**ROUGE-LSum**	**BLEU**
T = 0.7, p = 1.0	0.403	0.126
T = 1.0, p = 0.90	0.384	0.117
T = 1.0, p = 1.0	0.355	0.107
T = 0 (greedy)	**0.441**	**0.144**

(C) Prompt Designs (fs = 4, T = 0)		
Prompt Template	**ROUGE-LSum**	**BLEU**
Prompt #1	**0.441**	0.139
Prompt #2	0.411	0.135
Prompt #3	**0.441**	**0.144**
Prompt #4	**0.441**	**0.144**

Table 9: Ablation studies of reverse instructions on Mixtral-8×7B for (A) different number of few-shots, (B) sampling strategies, and (C) prompt designs.

Prompt Designs. Finally, we experiment with four alternative prompt templates to investigate the impact of the prompt:

- **Prompt #1**: Instruction: X
  Answer: {doc}
  What kind of instruction could this be the answer to? Generated instruction X should not refer to the answer such as ''[VERB] this article'' and it should be clear and self-explanatory when asked to someone.
  X:
- **Prompt #2**: Instruction: X
  Output: {doc}
  Generate a relevant and valid instruction for the given output?
  Instruction:
- **Prompt #3**: X: {doc}
  What is X? X:
- **Prompt #4**: Instruction: X
  Answer: {doc}
  What kind of instruction could this be the answer to?
  Instruction:

As shown in Table 9 (*C*), Prompt #3 and #4 achieve the highest ROUGE-LSUM and BLEU scores, comparable to each other. However, the differences between the various prompt templates are minimal, indicating that prompt design alone does not substantially impact the overall quality or diversity of the generated instructions. This suggests that while certain prompt structures may offer slight performance benefits, the diversity and quality of instructions are largely maintained across different prompts.

10.2 Translation Evaluation

We evaluate the quality of document translations into English and instruction translations from English for the MURI-IT dataset. We report the average BLEU scores of the MADLAD-400-3B-MT model evaluated on Flores-200 in Table 10. We observe patterns in the BLEU scores similar to those of the chrF scores, as described in Sections §3.1.2 and §3.1.3.

10.3 Prompt Selection on Instruction Diversity

We investigate whether relying on a single prompt in reverse instructions would limit the diversity of the generated instructions. We assess the instruction diversity within the Dolly dataset, which consists of English human-annotated

Resource Level	to English	from English
0 (9 langs.)	20.8	5.9
1 (75 langs.)	27.1	13.5
2 (17 langs.)	31.3	15.6
3 (23 langs.)	37.8	25.6
4 (16 langs.)	38.7	31.7
5 (6 langs.)	36.9	27.3

Table 10: Average BLEU scores of the MADLAD-400-3B-MT model evaluated on Flores-200, grouped by language resource level as defined in Joshi et al. (2020). Levels range from 0 (low-resource) to 5 (high-resource). Results represent 146 out of the 194 languages included in the reverse instructions subset of MURI-IT.

instruction-output pairs, to determine if different prompts produce different instructions.

To this end, we follow the methodology of Wang et al. (2023) and parse each instruction using the Berkeley Neural Parser (Kitaev and Klein, 2018). Specifically, we extract either the verb and its direct object (noun+verb, e.g., ‘‘describe+history’’) or the auxiliary verb with its dependent verb (auxiliary+verb, e.g., ‘‘do+make’’) as shown in §4.1. For each prompt template discussed in §10.1, we examine the types of noun+verb or auxiliary+verb pairs generated.

Our analysis reveals that while different prompts may result in varied surface forms of the instructions, the underlying verbs or nouns/auxiliaries remain same. For instance, we find that **84.2%** of the generated instructions using Prompt #1 and Prompt #4 share the same verb or noun/auxiliary. Across all six pairs of the four prompts, an average of **80.3%** of the instructions exhibit identical verb or noun/auxiliary structures. Furthermore, the variations observed in the remaining 20% of the instructions typically retain the same meaning, as demonstrated by examples such as ‘‘I’m planning a ski trip. How do I pick a mountain to visit? *(pick+mountain)*’’ versus ‘‘I’m planning a ski trip. What should I consider when picking a mountain to visit? *(should+consider)*’’, or cases where the neural parser fails to accurately detect the verb or noun/auxiliary. These findings indicate that changing the prompt has a minimal impact on the diversity of the generated instructions.

10.4 Few-shot Examples

Answer: A surfboard is shaped like a narrow plank. Surfers stand on the top of the surfboard and use it to surf on the ocean [...]
> *What kind of instruction could this be the answer to?*
Instruction: What is a surfboard used for?

Answer: Apache Kafka is a distributed system. The main components of Apache Kafka [...]
> *What kind of instruction could this be the answer to?*
Instruction: What are the main components of Apache Kafka?

Answer: Choosing the correct football boot depends on several factors. Of primary importance would be the surface you are playing on. [...]
> *What kind of instruction could this be the answer to?*
Instruction: I’m considering buying football boots. How do I know which one to buy?

Answer: Some common citrus-based beverages include orange juice, grapefruit juice, lemonade, Mountain Dew, mimosas.
> *What kind of instruction could this be the answer to?*
Instruction: List some common citrus-based beverages.

Answer: [DOC]
> *What kind of instruction could this be the answer to?*
Instruction:

Table 11: Few-shot examples used for reverse instructions.

10.5 Annotation Guideline to Evaluate MURI-IT

Task Description

You will be presented with a series of 100 instructions (prompts) and corresponding outputs (answers) based on them. Your task is to evaluate the instruction-output pairs based on several attributes to determine their quality and effectiveness in guiding a Large Language Model toward generating appropriate outputs.

Example Instruction-Output Pair

Instruction: What is a fracture?
Output: A fracture is the (local) separation of a body into two or more pieces under the action of stress.

The following attributes are evaluated for each instruction-output pair:

1. **Alignment:** Determine whether the instruction aligns with the output on a scale of 1 to 5, where:
   - **1:** The instruction and the output are completely misaligned, making it difficult to understand how the output was generated based on the given instruction (e.g., the output does not or not fully answer the instruction).
   - **5:** The instruction and the output are perfectly aligned, providing clear guidance on how to generate the output based on the given instruction.
2. **Instruction Format:** Identify if the instruction is phrased as an instruction or question:
   - Mark as ‘‘Instruction’’ if the given instruction provides a directive for generating the response, e.g., it is phrased as an instruction or question.
   - Mark as ‘‘No Instruction’’ if the given instruction is phrased as a statement, prompting no further answer.
3. **Grammatical and Lexical Correctness and Cohesiveness of the Instruction:** Assess whether the instruction is grammatically and lexically correct on a scale of 1 to 5, where:
   - **1:** The instruction contains numerous grammatical errors and uses inappropriate or unclear language, hindering comprehension and interpretation. The text is not cohesive; parts of the text don’t belong together.
   - **5:** The instruction is grammatically flawless and employs precise and appropriate language, facilitating clear understanding and interpretation.
4. **Grammatical and Lexical Correctness and Cohesiveness of the Output:** Assess whether the output is grammatically and lexically correct on a scale of 1 to 5, where:
   - **1:** The output contains numerous grammatical errors and uses inappropriate or unclear language, hindering comprehension and interpretation. The text is not cohesive; parts of the text don’t belong together.
   - **5:** The output is grammatically flawless and employs precise and appropriate language, facilitating clear understanding and interpretation.
5. **Informational Sufficiency:** Assess whether each instruction provides sufficient information for generating comprehensive outputs and whether it can be reasonably answered based on the provided information on a scale of 1 to 5, where:
   - **1:** The instruction lacks essential information and details, making it impossible to generate a reasonable answer or is ambiguous and not understandable. **Example:** Summarize the article.
   - **5:** The instruction provides ample information, and it is possible to be answered by a Large Language Model. **Example:** What does the word Rigadon mean?

Annotation Instructions

1. Read both the instruction and its output carefully.
2. Evaluate each instruction and its output independently based on the provided attributes.
3. Provide honest and thoughtful assessments for each instruction.
4. The output can contain additional information typical of websites such as links or footers; please account for this in your assessment by penalizing Grammaticality or Alignment accordingly.
5. If you encounter any difficulties or uncertainties, please refer back to these guidelines or reach out for clarification.
6. You can add your own notes in the Notes column if you want to explain your evaluation.