Benchmarking Public Large Language Models in Low-resource Languages

Anonymous EMNLP submission

Abstract

In recent years, Large Language Models have demonstrated impressive performances, particularly in zero-shot and few-shot learning across various languages. However, these models are often evaluated in English or high-resource languages, with limited focus on low-resource languages. This study benchmarks public LLMs which are commonly used in HuggingFace, including XGLM, Falcon, Llama, mT5-base, BLOOM, Mistral, Pegasus-Xsum, and the fine-tuned variants of mT5 and T5, on benchmark datasets in different low-resource languages. We conducted our experiments to evaluate the performance of these models across three natural language processing tasks: machine translation, text summarization, and question answering. Our evaluation results show a significant variability in performance, highlighting both the strengths and limitations of current multilingual large language models when applied to low-resource languages. Specifically, we observed that language models tend to perform better on languages with Latin alphabet, which is the most widely used in alphabetic writing, compared to those with non-Latin scripts, highlighting the need for more balanced training data.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional performance across various NLP tasks. Several studies indicate that providing LLMs with specific task instructions (e.g., summarizing or translating text) significantly enhances their capabilities (Muennighoff et al., 2023). This approach, known as instruction tuning, has been shown to improve performance in both English and multilingual contexts (Shaham et al., 2024; Wu and Dredze, 2019). Despite significant advancements in LLMs, most models remain English-centric, focusing primarily on English tasks (Brown et al., 2020). This limitation makes them less effective in multilingual settings, particularly in low-resource scenarios (Zhang et al., 2020a). Additionally, there is a lack of evaluation studies for public LLMs across different NLP tasks.

Existing evaluation studies have been recently proposed. This is the case of (Chang et al., 2023) who present a comprehensive study of benchmarking LLMs related to NLP tasks, methods and benchmarks, which are commonly used to assess the performance of LLMs in an English setting; To move beyond the English language, (Lai et al., 2023a) evaluate ChatGPT on 7 different tasks , covering 37 diverse languages with high, medium, low and extremely low-resources. However, these evaluation studies highlight the performance of LLMs either on English setting or using non public LLMs.

Low-resource languages (LRLs) are languages with limited linguistic resources and data. They often lack training datasets and necessitate pretraining techniques necessary to develop accurate NLP systems. To learn language patterns, LLMs require massive amounts of training data. However, LRLs still lack training data, making it challenging for LLMs, and generating this data can result in weaker LLM training for LRLs. In addition, LRLs are more complex than high-resource languages due to their unique vocabularies, grammatical structures, and linguistic features, making it difficult for LLMs to generate precise translation. Biases and errors in the training data can further reinforce misconceptions in LRL translations.

In this study, we evaluate popular public LLMs on 5 medium-resource languages (namely: *Hindi, Indonesian, Afrikaans, Bengali and Tamil*), and 11 LRLs (namely: *Amharic, Hausa, Igbo, Nepali, Somali, Swahili, Tigrinya, Telugu, Xhosa, Yoruba, and Zulu*) (Joshi et al., 2020) linked to different language Families. The distribution of our selected language families are displays in Figure 3. To conduct our experiments, we select 9 public LLMs: Llama, BLOOM, mT5 (fine-tuned), mT5-base, XGLM, Pegasus-XSum, T5 (fine-tuned), Fal-

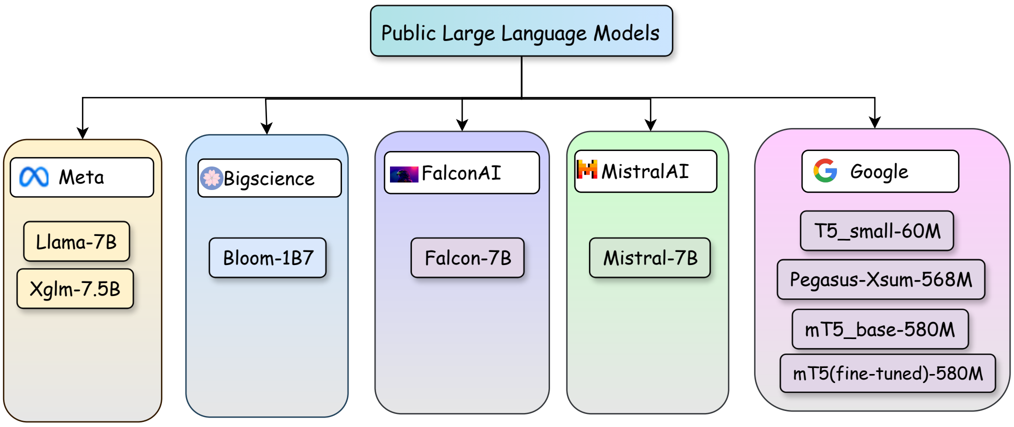

Figure 1: Public Large Language Models used in our evaluation study.

con, Mistral as shown in Figure 1. To underline the performance of these mLLMs particularly on common NLP tasks, we carried out our experiments on 3 NLP tasks: Machine Translation, Text Summarization and Question Answering.

Through our evaluation study, we designed our experiments to mainly answer the research question: *How well do the public LLMs perform in challenging NLP tasks (e.g., text summarization) across low-resource languages*. To answer this question, we included in our experiment different NLP tasks such as machine translation, abstractive text summarization and question answering. Furthermore, we assess the performance of these models on benchmark datasets using different evaluation metrics Our evaluation results demonstrate that LLMs trained on a balanced corpus which covers a diverse set of languages, and models fine-tuned on a few sets of samples tend to perform well while evaluated on languages with Latin scripts compared with languages with non-Latin scripts. We summarize the main contributions of this paper as follows:

- We provide a comprehensive evaluation of 9 LLMs on different NLP tasks across 16 low-resource languages.
- The evaluation results highlight the challenges of benchmarking LLMs on low-resource languages. Specifically, we observed that LLMs tend to perform better on languages with Latin alphabet, which is the most widely used in alphabetic writing, compared to those with non-Latin scripts.

In the following sections, we will discuss LLMs, NLP tasks, benchmark datasets selected for our experiments, the evaluation methodology and the experimental results. We start by discussing related works regarding benchmarking LLMs and multilingual benchmarks. Furthermore, we provide an overview of the different tasks, LLMs and multilingual benchmarks chosen for our evaluation, and describe the evaluation strategy and the experiments conducted in our study.

2 Related Works

In this section, we provide an overview of the related LLMs evaluation studies, specifically for *mLLMs* and *LRLs*.

2.1 Evaluation of Multilingual LLMs

Recent studies focus on creating and evaluating benchmarks (including datasets and frameworks) for LLMs in different domains. For example, in the medical domain, Alonso et al. (2024) proposed MedExpQA, the first multilingual benchmark based on medical exams to assess LLMs in medical question answering, demonstrating its performance in only four high-resource languages. Additionally, (Lai et al., 2023b) introduces Okapi, a benchmark dataset used to evaluate multilingual instruction-tuned LLMs with reinforcement learning from human feedback for three distinct tasks across 26 languages. (Ahuja et al., 2023) introduces MEGA, the first comprehensive benchmarking of generative LLMs, which evaluates models on standard NLP benchmarks, covering 16 NLP datasets across 70 topologically diverse languages. Liu et al. (2024) investigate the importance of translation with LLMs for a variety of scenarios, including multilingual NLP tasks and real-world multilingual user queries to enhance performance in mul-

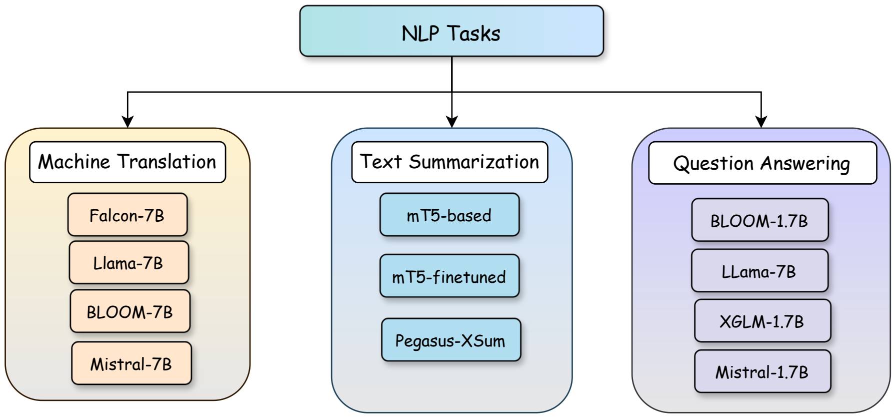

Figure 2: Tasks used in our evaluation study

tilingual NLP tasks with English-centric language models. Srivastava et al. (2023) introduced BIG-bench, which consists of 204 tasks largely related to translation to evaluate the behavior of LLMs. Further, Liang et al. (2023) proposed HELM, a holistic evaluation of 30 language models on 42 scenarios and 7 metrics, by defining a taxonomy of scenarios and metrics that span the space of LLM evaluation. However, these scenarios focused on datasets in high-resource languages, such as standard English or its dialects. Consequently, these LLMs can also exhibit grammatical structure bias, where structures from higher-resource languages influence LRLs.

2.2 Evaluation of Low-resource Languages

Evaluation methodologies have shown great performance on high-resources languages but failed to generalize on LRLs, particularly on languages with non-Latin scripts (Bang et al., 2023). Moreover, the performance of LLMs, such as ChatGPT, GPT-3.5 and BLOOMZ, have been evaluated, and the translation capabilities of these models perform well in high-resource languages but are limited in LRLs. This is because a larger vocabulary is needed to represent tokens in many languages, and a lack of language standardization leading to variations in grammar, vocabulary. and writing systems is observed across languages. To overcome these challenges, NLP communities have been developing benchmarks covering specific language families, such as IndicXTREME (Doddapaneni et al., 2023) for Indian languages, MasakhaNER (Adelani et al., 2021) for African languages, and IndoNLU (Wilie et al., 2020) for Indonesian languages.

Despite the overall progress in benchmarking LLMs, most works focus on evaluating non public LLMs either on high-resource scenarios or on non-English languages in general by highlighting a few NLP tasks. In this study, we benchmark public LLMs for three common NLP tasks—Machine Translation (MT), Text Summarization (TS) and Question Answering (QA)—focusing particularly on LRLs. We use three multilingual benchmark datasets: FLORES-101, XL-SUM, and OKAPI.

3 Tasks

3.1 Machine Translation

Machine Translation (MT) is a task of translating text in one language to another language without human intervention. For LRLs, MT poses significant challenges due to the lack of parallel data. Recent studies have highlighted the remarkable multilingual translation capabilities of very LLMs like GPT-4 for LRLs (Hendy et al., 2023; Garcia et al., 2023), without requiring explicit fine-tuning. On the other hand, medium-size LLMs such as XGLM have demonstrated superior performance compared to supervised state-of-the-art models using only few-shot examples (Lin et al., 2022a).

In our study, we employ publicly available LLMs such as Llama, Falcon, BLOOM, Mistral and XGLM for evaluating translation text from English to various LRLs and vice-versa. We aim to explore their potential application of these public models for improving MT quality, specially for LRLs.

3.2 Text Summarization

Text summarization (TS) is the process of long text into concise summarises that capture the most salient information. There are two types of text

Table 1: We provide a few selected LRLs used for our evaluation experiments including language code, language script, language family and total numbers of speakers.

	Iso-639-3	Language	Script	Language Family	Speakers
Medium	afr	Afrikaans	Latin	Indo-European-Germanic	8M
	ben	Bengali	Bengali	Indo-European-Indo-Aryan	282.9M
	ind	Indonesian	Latin	Austronesian	225M
	hin	Hindi	Devanagari	Indo-European-Indo-Aryan	571M
	tam	Tamil	Tamil	Dravidian	89.4M
Low	amh	Amharic	Ge'ez	Afro-Asiatic	57M
	hau	Hausa	Latin	Afro-Asiatic	77M
	ibo	Igbo	Latin	Atlantic-Congo	31M
	ne	Nepali	Devanagari	Indo-European-Indo-Aryan	32M
	som	Somali	Latin	Afro-Asiatic	22M
	swh	Swahili	Latin	Atlantic-Congo	200M
	tir	Tigrinya	Ge'ez	Afro-Asiatic	7M
	tel	Telugu	Telugu	Dravidian	96M
	xho	Xhosa	Latin	Atlantic-Congo	19M
	yor	Yoruba	Latin	Atlantic-Congo	46M
	zul	Zulu	Latin	Atlantic-Congo	11M

summarization: extractive summarization, which aims to select the most significant phrases from the original text as a final summarize; abstractive summarize that generates concise and human-like sentences based on the original (Liu et al., 2017), (Raposo et al., 2022), (Zhang et al., 2020b). In our study, we focus on the later one, since it is one of the most challenging NLP tasks and requires advanced abilities, such as understanding long texts and generating coherent text. Recently, several fine-tuned LLMs for abstractive summarization have been proposed, however most of them are for monolingual (e.g., English) (Askari et al., 2024), (Zhang et al., 2023). For this task, we evaluate the publicly available models such as mT5-base, mT5 fine-tuned on XLSum dataset, Pegasus-Xsum and the fine-tuned version of T5-small on benchmark datasets in different LRLs. Our goal is to explore the capabilities of these LLMs in generating coherent summaries without prior fine-tuning for these languages.

3.3 Question Answering

Question Answering (QA) systems are designed to interpret and answer queries in natural language. Recently, various QA models and datasets have been developed to accomplish enable machines understand the context of queries and precisely answer them (Rajpurkar et al., 2016) (Yang et al., 2015), (Campese et al., 2023). However, these datasets pose unique challenges such as finding the answer span when the context and the problem are in different language. To address this issue, researchers adopt recent advancements in LLMs to encode input text and use additional layers for classification and solving multilingual QA task (Lewis et al., 2020), (Clouatre et al., 2020), (Yao et al., 2019), (Wang et al., 2020). In our study, we focus on multilingual QA task since it is a crucial step towards cross-lingual machine comprehension in LRLs. We use in our study different public multilingual LLMs, namely BLOOM, Llama, Mistral and XGLM, and evaluate them on different benchmark datasets.

4 Multilingual Large Language Models

Our study benchmarks different LLMs based on two main criteria: i) they are publicly available, and ii) they can be employed in multilingual NLP tasks. The models included in our evaluation are BLOOM (Workshop et al., 2023), XGLM (Lin et al., 2022b), Falcon (Almazrouei et al., 2023), Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023), and fine-tuned variants of mT5, mT5-base (Xue et al., 2021), T5 (Raffel et al., 2023), and Pegasus-XSum.

For **MT**, we employ the 7B of Falcon, a decoder-only model trained on 11 natural languages; the 7B of Llama, an encoder-decoder model trained on 20 natural languages; the 7B of BLOOM, a decoder-only model trained on 46 natural languages; the 7B of Mistral-v0.2, a decoder-only model trained on 6 natural languages; and the 1.7B of XGLM, a decoder-only model trained on 31 natural languages. For **TS**, we consider the fine-tuned variants of mT5 trained on 45 natural languages of

XL-Sum[1] dataset; mT5-base, a pretrained encoder-decoder model covering 101 natural languages; Pegasus-XSum, an encoder-decoder model fine-tuned on XSum (Narayan et al., 2018) dataset and also evaluated on low-resource summarization; the fine-tuned variants of T5 small, an encoder-decoder model which used a text-to-text approach. For **QA**, we use the 1B7 version of BLOOM which covers 48 natural languages; the 7B of Llama trained on 20 natural languages; the 1.7B of XGLM trained on 31 natural languages and the 7B of Mistral-v0.2 trained on 6 natural languages.

By comparing similar-sizes LLMs on different benchmark training data, we highlight their relative strengths and weaknesses in handling multilingual context, particularly for LRLs.

5 Benchmark Datasets

In this section, we present the benchmark datasets used in our evaluation study across NLP tasks: Machine Translation, Text Summarization, and Question Answering.

Benchmark dataset for Machine Translation: We employ FLORES 101 [2] (Goyal et al., 2021) dataset, which contains $3k$ sentences extracted from English Wikipedia articles, converging various topics. The dataset is also designed for many-to-many evaluation, allowing for comprehensive evaluation of multilingual MT models across many source languages and many target languages, specially with low resources.

Benchmark dataset for Text Summarization We use XL-SUM [3], a comprehensive dataset tailored for abstractive summarization, consisting of $1M$ professionally annotated article-summary pairs from BBC news articles (Hasan et al., 2021). The data was extracted using a set of carefully designed heuristics and covers 44 languages ranging in resource level from low to high.

Benchmark dataset for Question Answering To assess the performance of LLMs in multilingual questions answering, we empoly three benchmark datasets, namely: AI2 Reasoning Challenge (ARC)[4] (Clark et al., 2018), Hellaswag[5] (Zellers et al., 2019) and MMLU (Hendrycks et al., 2021) from Okapi[6] framework. These datasets have been translated from the original AI2 Reasoning Challenge (ARC), Hellaswag, and MMLU datasets in English into 26 languages, including LRLs, using ChatGPT.

6 Evaluation Methodology

Two significant techniques can be used for prompting LLMs for a given NLP task. First, prompting LLMs with in-context (Brown et al., 2020), which is a straightforward approach for leveraging LLMs in solving a given NLP task with few-shot examples given in the context without the need for training of fine-tuning. The second technique is instruction tuning (Mishra et al., 2022; Ouyang et al., 2022), which is a novel approach to guide LLMs, following instructions and solve new-tasks based on textual instructions provided in prompt. In our study, we use both techniques as follow:

- **Evaluating Machine Translation**: Following (Zhu et al., 2023), we adopt their learning strategy to evaluate the performance of LLMs in translation text across different languages. (*details can be found in section 7.1*). As an evaluation metric for all languages, we employ SacreBLEU [7] (Post, 2018), a variant of the BLEU score. SacreBLEU works with plain text and generates official WMT scores in comparison to the original BLEU score. Moreover, it facilitates the download and management of test sets throughout assessments.

- **Evaluating Text Summarization**: We fine-tuned the mT5 model as a baseline in the same manner as (Hasan et al., 2021) and then perform our experiments on abstractive summarization in two settings: (i) multilingual, and (ii) low-resource. We employ multilingual ROUGE Scoring [8], a metric used to evaluate TS as an evaluation metric for all languages. The results indicate that there are four ROUGE type scores, namely: ROUGE-1 (unigram based scoring), ROUGE-2 (bigram based scoring), ROUGE-L (longest common sub-sequence based scoring), and ROUGE-Lsum (splits text using "$\backslash n$"). We report the first three types of scores in Table 2.

[1] https://github.com/csebuetnlp/xl-sum
[2] https://huggingface.co/datasets/gsarti/flores_101
[3] https://github.com/csebuetnlp/xl-sum
[4] https://allenai.org/data/arc
[5] https://allenai.org/data/hellaswag
[6] https://github.com/nlp-uoregon/Okapi
[7] https://github.com/mjpost/sacrebleu
[8] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

- **Evaluating Question Answering**: We employ the Okapi[9] framework, an evaluation framework designed for instruction-tuned LLMs. We use the accuracy metric as our evaluation metric since it enables evaluation of all languages.

7 Experiments

In this section, we provide the details of our experiment setup for each NLP task (MT, TS and QA).

7.1 Multilingual LLMs Evaluation on MT

Selected languages: Among 101 languages in the FLORES-101 dataset, we selected a set of LRLs with Latin and non-Latin scripts that are relevant for MT tasks. These include *Amharic, Afrikaans, Indonesian, Hausa, Hindi, Igbo, Somali, Swahili, Tamil, Xhosa, Yoruba, and Zulu*. Our selection criteria were based on linguistic diversity and the number of native speakers. Notably, some LRLs are spoken across multiple continents as shown in Table 1. For example, Hindi is mainly spoken in India but also has speakers in the United States and Canada. We also prioritize languages like *Xhosa, Zulu, Somali,* and *Tamil*, which have limited online resources, to evaluate the MT models' capabilities with less data.

Learning Strategy: With 12 translation pairs, we report the performance of each LLM in MT using the following direction: $X2E$, which means translation from a source language to English and vice-versa $E2X$. We used the*OpenICL* [10] framework (Wu et al., 2023) as a foundation for all implementations.

Results Table 3 shows the performance of five LLMs on the FLORES-101 benchmark for MT. The models evaluated are: Falcon-7B (Almazrouei et al., 2023), Llama-7B (Touvron et al., 2023), BLOOM-7B (Workshop et al., 2023), Mistral-7B (Jiang et al., 2023), and XGLM-7.5B (Lin et al., 2022b), all of which have comparable parameter sizes. The evaluation results show that XGLM outperforms the other models in translating LRLs, specially Afrikaans, Indonesian, Hindi, Swahili and Tamil, with the most notable performance in Indonesian translations.

[9] https://github.com/nlp-uoregon/Okapi
[10] https://github.com/Shark-NLP/OpenICL

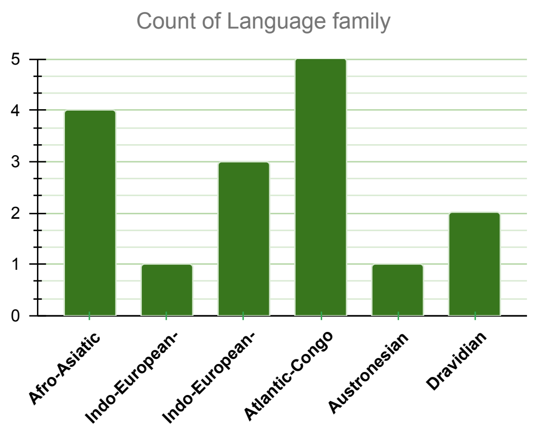

Figure 3: Distribution of language family used in our experiments.

Performance analysis: The superior performance of XGLM can be attributed to its training on a balanced multilingual corpus compared to the other models that are trained primarily on English datasets. This training approach allows XGLM to excel in few- and zero-shot learning across various tasks. Specifically, XGLM has demonstrated its effectiveness by surpassing the official supervised baseline in 45 directions and outperforming models like GPT-3 in 171 out of 182 translation directions with just 32 training examples on the FLORES-101 MT dataset.

7.2 Multilingual LLMs Evaluation on TS

Selected Languages For this abstractive text summarization, we chose 11 LRLs for our evaluation, namely: *Amharic, Bengali, Indonesian, Hausa, Hindi, Igbo, Somali, Swahili, Tamil, Tigrinya, and Yoruba*. These languages were chosen based on different criteria: their presence on at least two continents, the number of speakers, their language families, and the availability of resources.

Results Table 2 shows the performance of different LLMs on summarizing text in the selected languages. In particular, we evaluate four multilingual LLMs on the XL-SUM dataset: i) mT5-multilingual-XLSum (a fine-tuned version of mT5) (Xue et al., 2021), ii) mT5-base (Xue et al., 2021), iii) Pegasus-XSum (Zhang et al., 2020b), and iv) T5 (fine-tuned) (Raffel et al., 2023). The evaluation results indicate that the mT5-multilingual-XLSum model consistently outperforms the other models across all languages. Specifically, mT5 variants show high performance on Hausa, while Pegasus-

Table 2: Summarization performance of LLMs over LRLs. Best results are in bold.

	mT5-multilingual-XLSum			mt5-base			Pegasus-xsum			T5-small (fine-tuned)		
Language	**R-1**	**R-2**	**R-3**	**R-1**	**R-2**	**R-3**	**R-1**	**R-2**	**R-3**	**R-1**	**R-2**	**R-3**
amh	20.65	8.01	18.69	3.58	0.82	3.45	0.13	0.0	0.13	0.13	0.0	0.13
ben	29.76	12.42	25.57	6.12	1.60	5.97	0.0	0.0	0.0	0.0	0.0	0.0
ind	37.01	17.15	30.88	8.41	2.38	7.74	15.81	4.24	12.67	20.41	5.32	15.03
hau	**39.96**	**18.23**	**32.20**	**9.81**	1.98	**8.64**	18.93	4.11	13.83	**25.34**	**6.01**	**17.61**
hin	38.74	17.21	21.24	9.35	1.73	7.91	0.31	0.07	0.31	0.06	0.01	0.06
ibo	32.29	10.90	25.35	7.69	1.46	7.18	**21.86**	3.91	**16.17**	24.78	4.81	17.31
som	31.74	11.80	24.37	7.98	1.37	7.17	20.81	4.47	15.23	20.66	4.13	14.58
swh	37.72	18.00	31.17	9.49	2.64	8.65	16.80	3.77	12.78	22.57	5.57	16.34
tam	24.45	11.27	22.25	3.48	0.97	3.36	0.55	0.0	0.54	0.26	0.13	0.26
tin	25.98	8.99	22.05	5.95	1.08	5.56	0.52	0.0	0.51	0.11	0.09	0.11
yor	32.01	12.69	25.93	5.80	1.10	5.50	21.28	4.23	14.91	22.55	4.48	15.52
AVG	**31.84**	**13.33**	**25.42**	7.06	1.55	6.46	10.63	2.25	6.76	12.44	2.77	8.81

Table 3: Translation performance of LLMs over LRLs. Best results are in bold.

	Falcon-7B		Llama-7B		BLOOM-7B		Mistral-7B		XGLM-7.5B	
Language	**BLEU**$_{X2E}$	**BLEU**$_{E2X}$	**BLEU**$_{X2E}$	**BLEU**$_{E2X}$	**BLEU**$_{X2E}$	**BLEU**$_{E2X}$	**BLEU**$_{X2E}$	**BLEU**$_{E2X}$	**BLEU**$_{X2E}$	**BLEU**$_{E2X}$
amh	0.37	0.02	0.54	0.02	0.27	0.10	0.87	0.01	0.21	0.01
afr	11.18	5.93	15.79	9.37	6.85	3.45	16.00	9.78	**16.73**	3.91
ind	11.03	4.90	13.44	5.71	10.22	8.88	15.87	10.14	**34.32**	**30.38**
hau	1.80	0.96	1.90	1.20	1.57	0.67	2.97	2.04	2.60	0.45
hin	0.47	0.33	7.48	3.59	7.11	7.91	11.85	5.81	**24.19**	**18.40**
ibo	1.93	1.29	1.82	1.10	1.20	0.99	2.46	1.72	1.72	0.26
som	1.74	0.75	2.16	0.83	1.49	0.59	3.44	2.04	2.44	0.31
swh	2.47	1.04	2.65	1.27	7.03	4.14	5.32	2.56	**30.01**	**19.07**
tam	0.40	0.00	0.82	0.15	4.15	5.06	2.69	0.89	**14.86**	**9.00**
xho	2.05	1.45	2.28	1.16	1.39	0.62	3.72	2.19	1.86	0.94
yor	1.78	1.12	1.79	1.39	1.82	1.88	2.81	1.82	1.97	0.69
zul	1.61	1.10	1.87	0.94	0.98	0.45	2.92	1.87	1.49	0.74
AVG	3.06	1.57	4.37	2.22	3.10	2.89	5.91	3.40	**11.03**	**7.01**

Table 4: QA performance of LLMs over medium (ind, hin, tam, ben) and very LRLs (ne and tel). Best results for each task are in bold.

	BLOOM-1B (Acc)			Llama-7B (ACC)			XGLM-1.7B			Mistral-7B-v0.2		
Language	**ARC**	**Hellaswag**	**MMLU**	**ARC**	**Hellaswag**	**MMLU**	**ARC**	**Hellaswag**	**MMLU**	**ARC**	**Hellaswag**	**MMLU**
ind	23.76	33.49	25.14	19.23	29.77	27.93	20.85	31.70	24.60	**32.65**	**38.11**	**40.97**
hin	20.89	29.11	23.60	21.15	27.08	25.52	20.46	28.43	23.67	**22.17**	**29.29**	**30.41**
tam	**23.29**	25.76	24.07	20.67	25.53	24.67	22.15	25.38	23.51	21.45	**25.95**	**27.40**
ben	20.62	26.88	24.99	19.08	25.97	25.09	19.76	26.68	24.01	**21.13**	**27.43**	**29.06**
ne	20.02	26.59	23.91	21.81	26.46	24.54	**22.50**	25.23	23.63	22.16	**27.18**	**28.47**
tel	19.04	25.99	24.13	**20.26**	25.57	24.64	17.54	25.69	23.89	19.65	**26.04**	**26.66**

XSum demonstrates superior results on Igbo compared to other languages.

Performance analysis: The mT5-multilingual-XL-SUM is based on the mT5 checkpoint, which is fine-tuned on 45 languages, including our selected languages. Hausa and Igbo had approximately $6k$ and $4k$ training samples, respectively, which is a good indication that models fine-tuned on such a small training data can still generalize and produce competitive results to multilingual models.

7.3 Multilingual LLMs Evaluation on QA

Selected Languages For this task, we asses LLMs on six LRLs from the Okapi framework. These languages are *Indonesian, Hindi, Nepali, Bengali, Tamil and Telugu*. These languages were chosen due to their limited resources, which are considered as LRLs.

Results We evaluated the multilingual capabilities of four publicly available models: BLOOM-7B, Llama-7B, XGLM-1.7B and Mistral-7B-v0.2 on different QA benchmark datasets. We explore the evaluation results for each dataset as follow: ARC, Hellaswag, and MMLU

- **ARC**: as shown in Table 4, Mistral tends to perform well on Indonesian, Hindi and Bengali compared to the other models. BLOOM outperforms other models especially on Tamil, XGLM on Nepali, and Llama on Telugu.
- **Hellaswag**: on this dataset, Mistral outperforms other LLMs on all languages, and Indonesian presents the highest score across languages.
- **MMLU**: Mistral also outperforms the other models on all languages. The highest score is shown particularly on Indonesian.

Performance analysis: Indonesian is known as the official and national language of Indonesia and is spoken by over $225M$ people. Indonesian is also among the most widely spoken languages in the world. This language is classified as a medium-resource language (Joshi et al., 2020) with a high score while evaluating the baselines models in the framework, thus enabling researchers to develop more open-access tools and resources in that language. Notably, among the selected LRLs, Bengali and Tamil present a quite significant score compared to other languages. This observation aligns with previous findings that the Mistral-7B model outperforms the best open 13B model (Llama 2) across all evaluated benchmarks and the best released 34B model (Llama 1) in reasoning, mathematics, and code generation tasks.

8 Conclusion

In this work, we present a comprehensive study to evaluate a set of 9 public LLMs, commonly used in Hugging Face, on three different NLP tasks—machine translation, text summarization, and question answering—particularly focusing on low-resource languages. We evaluated the performance of these LLMs on 16 low-resource languages based on available resources. Our findings highlight the challenges and limitations of evaluating LLMs on low-resource system due to the scarcity of training data and different writing scripts. To address this limitation and advance the state-of-the-art in low-resource languages, future research efforts could explore the development of specialized models tailored for low-resource languages by incorporating native speakers as human-in-the-loop feedback mechanisms during model training.

9 Limitations

While our study provides valuable insights into the performance of multilingual large language models on low-resource languages, we acknowledge two main limitations as follows: i) our evaluation focused on a subset of publicly available LLMs and multilingual benchmark datasets. Due to the vast number of models and resources available, we selected the popular and publicly LLMs from the Hugging face Hub, and ii) the multilingual LLMs evaluated in this study were not specifically optimized or customized for the selected low-resource languages. To achieve significant results on low-resource languages, we believe further research should focus on developing tailored LLMs by incorporating native speakers as human-in-the-loop feedback mechanisms during model training.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering.

Hadi Askari, Anshuman Chhabra, Muhao Chen, and Prasant Mohapatra. 2024. Assessing llms for zero-shot abstractive summarization through the lens of relevance paraphrasing.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

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.

Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. 2023. Quadro: Dataset and models for question-answer database retrieval.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

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

Louis Clouatre, Philippe Trempe, Amal Zouaq, and Sarath Chandar. 2020. Mlmlm: Link prediction with mean likelihood masked language model.

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.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation.

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

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

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

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.

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.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023a. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

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

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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. 2022a. 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.

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. 2022b. Few-shot learning with multilingual language models.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative adversarial network for abstractive text summarization.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Gonçalo Raposo, Afonso Raposo, and Ana Sofia Carmo. 2022. Document-level abstractive summarization.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy

Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

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. 2023. Llama: Open and efficient foundation language models.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2020. Kepler: A unified model for knowledge embedding and pre-trained language representation.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding.

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. Al-shaibani, 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, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. 2023. Openicl: An open-source framework for in-context learning.

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.

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.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020b. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

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