Beyond Latin Scripts: Performance Gaps in Public Large Language Models for Low-resource Languages

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated exceptional performance across a wide range of natural language processing tasks. However, their capabilities in linguistically diverse, low-resource contexts remain underexplored—particularly for languages that do not use Latin scripts. This study evaluates nine publicly accessible LLMs across 14 low-resource languages (LRLs), encompassing both Latin and non-Latin scripts (e.g., Ge'ez, Devanagari, Cyrillic), focusing on three key tasks: machine translation, text summarization, and question answering. Our analysis reveals significant performance disparities: languages with Latin scripts (e.g., Somali, Swahili, Yoruba) perform better compared to those with non-Latin scripts (e.g., Pashto, Nepali, Sinhala, Amharic), particularly in text summarization, with ROUGE scores differing by up to 39% across languages. These disparities are strongly correlated with the type of tokenizer used: the majority of tokenizer models in this study are not effective when dealing with languages outside their primary training distribution or those with distinct linguistic features (e.g., non-Latin scripts, complex morphology). This highlights a critical need for language-specific tokenizers—or multilingual tokenizers explicitly designed to accommodate a broader range of linguistic characteristics—for optimal LLM performance on linguistically diverse LRLs.

1 Introduction

Large language models (LLMs) have demonstrated state-of-the-art performance across various natural language processing (NLP) tasks. Several studies indicate that providing LLMs with specific task instructions, such as summarizing or translating text, significantly enhances their capabilities (Muennighoff et al., 2023). This method, known as instruction tuning, has been shown to improve LLMs' performance in both English and multilingual contexts (Shaham et al., 2024, Wu et al., 2023). Despite significant advancements, most LLMs remain English-centric, focusing primarily on English tasks (Brown et al., 2020a). Major gaps remain in evaluating public LLMs across diverse languages and tasks, particularly for under-resourced, script-diverse contexts (Zhang et al., 2020).

Recent studies highlight growing efforts to benchmarking LLMs across languages and tasks. For instance, Chang et al. (2023) presented a comprehensive study on benchmarking LLMs related to NLP tasks, methods, and benchmarks, which are commonly used to assess performance in English settings. To extend beyond English, Lai et al. (2023a) evaluated ChatGPT on seven different tasks, covering 37 diverse languages with high, medium, low, and extremely low resources. However, these evaluation studies analyzed the performance of LLMs either in English settings or using non-public LLMs, leaving a major gap in understanding public LLM capabilities for low-resource, script-diverse languages (Liu et al., 2024; Brown et al., 2020b; Liang et al., 2023).

Low-resource languages (LRLs) are defined by limited linguistic resources and data, posing challenges for LLMs in learning robust language patterns (Magueresse et al., 2020). Joshi et al. (2021) categorize languages in six classes based on the availability of labeled and unlabeled data: (0) *The Left-Behinds*, (1) *The Scraping-Bys*, (2) *The Hopefuls*, (3) *The Rising Stars*, (4) *The Underdogs*, and (5) *The Winners*. In a simplified form, class 0 languages have neither labeled nor unlabeled data; class 1-4 languages have unlabeled data, but their labeled data quantity varies from virtually nonexistent to high and, class 5 languages have both high volumes of labeled and unlabeled data.

While prior work identifies cross-lingual exemplars and unintentional bilingualism as drivers of LLM translation, our study uniquely highlights script type as a systemic bias, demonstrating its correlation with pretraining data scarcity by evaluating

public[1] LLMs on 14 LRLs (classes 0 to 2) spanning diverse scripts, including *Amharic, Telugu, Burmese, Nepali, Kannada, Pashto, Tajik, Swahili, Yoruba, Somali, Sinhala, Marathi, Punjabi, Kyrgyz*. These languages cover diverse linguistic families and resource levels, enabling analysis of script and data disparities.

We designed our experiments to answer the research question: *How robust are public LLMs across NLP tasks in LRLs and script diversity settings?* To answer this question, we benchmark nine public LLMs on three high-impact NLP tasks: Machine Translation, Text Summarization and Question Answering.

Our analysis demonstrates two major factors shaping LLM performance: *the type of tokenizer* and *the script type*. Models fine-tuned with minimal data (e.g., mT5) excel on languages well-represented in their training. However, some tokenizers still struggle to handle languages that fall outside their primary training distribution or exhibit distinct linguistic characteristics, such as non-Latin scripts or complex morphology, and models evaluated on Latin-script languages (e.g., Swahili) consistently outperform those tested on non-Latin-script languages (e.g., Nepali), with performance gaps exceeding 39% in the text summarization task.

We summarize the main contributions of this paper as follows:

- We provide a comprehensive evaluation of nine public LLMs on different NLP tasks across 14 languages ranging from class 0 to 2.
- We conduct tokenization errors analysis to understand the model capability to generalize on the selected languages and report the results analysis.
- The evaluation results highlight the challenges of benchmarking LLMs on LRLs in each task.
- We provide our benchmark source code[2].

The remainder of this paper is organized as follows. We provide a review of recent work on benchmarking LLMs in Section 2. In Section 3, we detail our methodology and task definitions, while Section 4 presents the experimental setup. Section 5 presents our results. Section 6 presents our tokenization analysis. We conclude in Section 7 and provide limitations in Section 8.

[1]Public LLMs are openly accessible via APIs or repositories like Hugging Face and GitHub.

[2]https://anonymous.4open.science/r/Benchmarking-LLM-B12C

2 Related Work

2.1 Evaluation of Multilingual LLMs

Evaluating multilingual LLMs is a challenging task due to the lack of comprehensive and language-agnostic benchmarks. Recent studies have focused on creating and evaluating benchmarks (including datasets and frameworks) for LLMs in different domains. For example, in the medical domain, Alonso et al. (2024) introduced MedExpQA, the first multilingual benchmark based on medical exams to assess LLM performance in four high-resource languages. Additionally, Liang et al. (2020) presented XGLUE, a cross-lingual evaluation benchmark with 11 tasks across 19 languages, where training data is available only in English. Hu et al. (2020) introduced XTREME, which covers 40 languages and includes 9 tasks to evaluate cross-lingual transfer in multilingual encoders. Moreover, Lai et al. (2023b) developed Okapi, a benchmark for evaluating multilingual instruction-tuned LLMs with reinforcement learning from human feedback for 43 distinct tasks across 26 languages. Ahuja et al. (2023) introduced MEGA, the first comprehensive benchmark for generative LLMs, covering 16 NLP datasets across 70 topologically diverse languages. Further, Liang et al. (2023) proposed HELM, a holistic evaluation for 30 language models on 42 scenarios and 7 metrics. However, these scenarios primarily focus on high-resource languages like English or its dialects, leading to potential grammatical structure bias, where syntactic patterns from higher-resource languages influence those of LRLs.

2.2 Evaluation on Low-resource Languages

Evaluation methodologies excel in high-resource languages but often fail to generalize to LRLs, particularly those with non-Latin scripts (Bang et al., 2023a). Models 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 (Bang et al., 2023b; Chowdhery et al., 2023; Muennighoff et al., 2023) . This is because a larger vocabulary is needed to represent tokens in many languages, and a lack of language standardization leads to variations in grammar, vocabulary and writing systems is observed across languages. To address these chal-

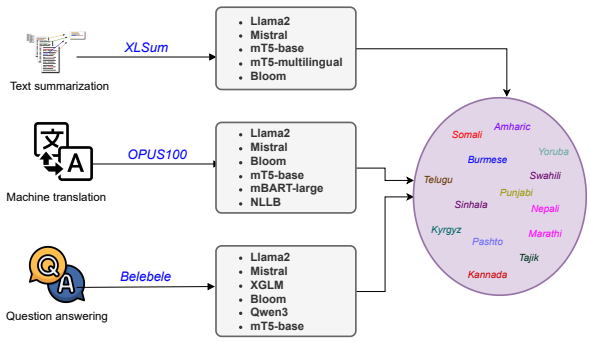

Figure 1: An overview of the tasks, datasets, public LLMs, and languages used in our evaluation study.

lenges, NLP communities have developed 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 progress in benchmarking LLMs, most studies include only a few samples of low-resource language and non-Latin scripts in their pre-training corpora, and focusing primarily on non-public LLMs in high-resource scenarios.

In contrast, our study addresses theses gaps by systematically evaluating nine public multilingual LLMs across a diverse set of low-resource (class 0 to 2) and non-Latin script languages on three NLP benchmarks.

3 Multilingual Large Language Models

Our study benchmarks different LLMs based on two criteria: i) they are publicly available, and ii) they can be employed in multilingual NLP tasks.

An overview of the tasks, datasets, LLMs, and languages considered in our study is given in Figure 1. More details are given in Appendix A. We include the following LLMs in our study: LlaMA 2[3] (Touvron et al., 2023), BLOOM[4] (Workshop et al., 2023), Mistral[5] (Jiang et al., 2023), XGLM[6] (Lin et al., 2022), mT5[7] (fine-tuned) (Xue et al., 2021), mT5-base[8] (Xue et al., 2021), mBART-large-50-many-to-many-mmt[9] (Tang et al., 2020), NLLB[10] (Team et al., 2022) and Qwen[11] (Yang et al., 2025). LLaMA2 and Mistral have the smallest token vocabulary (32K), followed by Qwen (152K), BLOOM (250K), and XGLM (256K), which has the largest vocabulary among them.

We examine three tasks: Machine Translation, Text Summarization, and Question Answering. For each task, we evaluate LLMs of the same size, using three multilingual benchmark datasets related to each task: OPUS100, XL-Sum, and Belebele.

Machine Translation: is the task of translating text from one language to another without human intervention. For LRLs, machine translation poses significant challenges due to the lack of parallel data. Recent studies have highlighted the remarkable multilingual translation capabilities of LLMs such as GPT-4 for LRLs, even without explicit fine-tuning (Hendy et al., 2023; Garcia et al., 2023). In this task, we specifically evaluate LLMs trained on a wide range of languages to assess the effectiveness of their pre-training approaches, which involve predicting subsequent text based on the provided context in an autoregressive manner—particularly for LRLs. Models evaluated include NLLB, mBART-large, mT5-base, Mistral, BLOOM, and LLaMA 2, which we use to translate text from English into various LRLs.

Text Summarization: is the process of condensing long texts into concise summaries that capture the most salient information. In our study, we focus on abstractive summarization, one of the most challenging NLP tasks, as it requires advanced capabilities such as understanding lengthy passages and generating coherent summaries. Although several fine-tuned LLMs for abstractive summarization have been proposed recently, most are designed for monolingual settings (e.g., English) Askari et al., 2024; Zhang et al., 2024. In our work, we consider publicly available LLMs—such as mT5-base, mT5 fine-tuned on the XLSum dataset[12], LLaMA 2, Mistral, and BLOOM—across different LRLs. Our goal is to assess the ability of these models to generate coherent summaries without prior fine-tuning for these languages.

Question Answering: is a system that interprets and responds to natural language queries, leveraging advanced models and datasets to enhance

[3] https://huggingface.co/huggyllama/llama-7b
[4] https://huggingface.co/bigscience/bloom-7b1
[5] https://huggingface.co/mistralai/Mistral-7B-v0.1
[6] https://huggingface.co/facebook/xglm-7.5B
[7] https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum
[8] https://huggingface.co/google/mt5-base
[9] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
[10] https://huggingface.co/facebook/nllb-200-3.3B
[11] https://huggingface.co/Qwen/Qwen3-8B
[12] https://github.com/csebuetnlp/xl-sum

contextual understanding and accuracy Rajpurkar et al., 2016; Yang et al., 2015; Campese et al., 2023. We focus on the multilingual question-answering task, as it represents a crucial step toward cross-lingual machine comprehension in LRLs.

For this task, we evaluate LLMs such as LLaMA 2, Mistral, XGLM, BLOOM, mT5-base, and Qwen3. By comparing models of similar sizes trained on different benchmark datasets, we identify their relative strengths and weaknesses in handling multilingual contexts—particularly for LRLs across each task.

4 Evaluation Methodology

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

Machine Translation: For this task, we employ both instruction tuning and in-context prompting by evaluating LLMs that are either explicitly trained on translation data using $src \rightarrow tgt$ pairs or designed as sequence-to-sequence translation models. In instruction tuning, no explicit prompts are used—the translation is handled directly through the model's input/output format. In contrast, in-context prompting involves constructing a textual prompt; the model generates translations based on its understanding of the instruction embedded in the prompt, without any additional fine-tuning. We use OPUS100 (Zhang et al., 2020) as benchmark dataset and ChrF++ (Popović, 2017) as metric.

Text Summarization: We also use both instruction tuning and in-context prompting for this task. In the first case, we evaluate supervised models trained on summarization datasets, using standard input/output formats. In the second case, we construct manual prompts that rely on the model's general language understanding to infer the summarization task from the prompt without fine-tuning. We use XL-Sum (Hasan et al., 2021) as the benchmark dataset and ROUGE (Lin, 2004) as metric.

Question Answering: For this task, we use zero-shot in-context prompting, where the model is not fine-tuned for the specific QA task but is instead prompted in a zero-shot format using the structure: ⟨context⟩, ⟨question⟩, ⟨answer⟩. Each multiple-choice answer is scored independently using a text classification model, such as BART-MNLI, trained for natural language inference (NLI)—which judges entailment between a premise and a hypothesis. We use Belebele (Bandarkar et al., 2024) as the benchmark dataset and the F1 score as the evaluation metric.

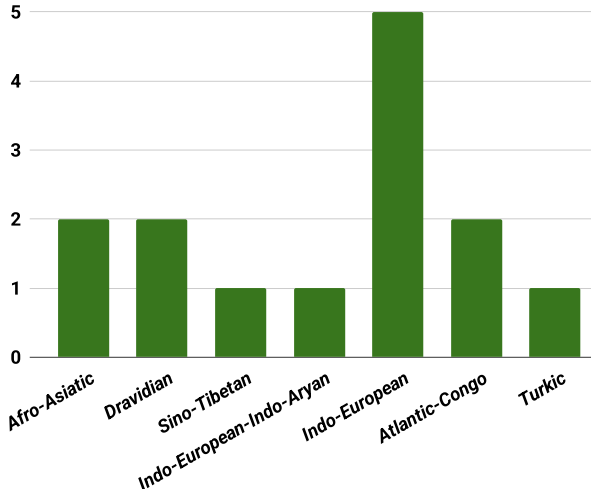

Figure 2: The 14 languages in our experiments categorized to language families.

Low-resources Languages: By considering diverse linguistic distributions and language scripts—spanning a variety of language families—we selected 14 LRLs, ranging from class 0 to 2, to assess the capabilities of LLMs to generalize, even to unseen languages (see Table 4). Each selected language belongs to at least one distinct language family, as shown in Figure 2, and the overall language distribution is presented in Figure 3.

5 Evaluation Results

In the following evaluation results, NL and L denote languages with non-Latin and Latin scripts, respectively. Dashes (i.e., $-$) in the results mark unsupported languages in the dataset, and results in bold indicate the highest scores.

5.1 Evaluation on Machine Translation

Learning Strategy: With 14 translation pairs, we report the performance of each LLM in machine translation from English to the target languages.

Results: Table 1 shows translation performance (ChrF++) across LRLs. With an average between

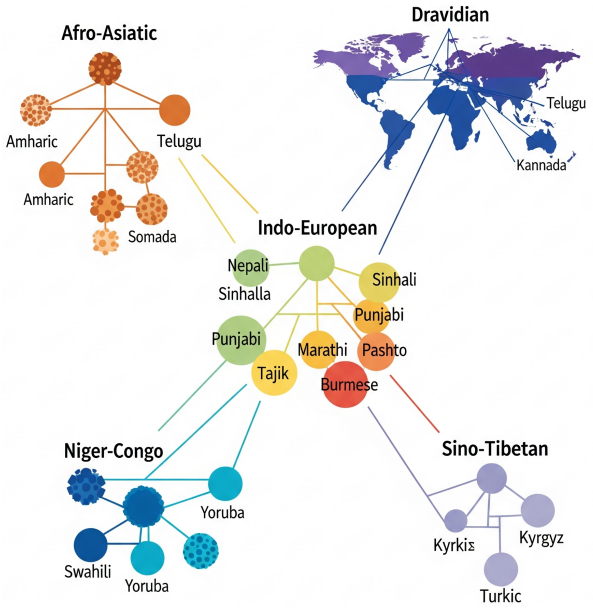

Figure 3: Language family distribution.

4.91 for NLLB and 13.62 for Llama2, the performance across all models for all languages is poor. For example, NLLB performs better on Telugu; mBART-large performs better on Yoruba, Sinhala, Marathi, and Burmese; Mistral performs best on Pashto, Tajik, and Kyrgyz; while LLaMA 2 excels in Kannada, Punjabi, Burmese, and Nepali. We also observe poor performance from NLLB, mT5-base, and BLOOM. Additionally, the Llama2 model achieves the highest average score compared to the others. Overall, the results reveal that translation performance is significantly worse for languages with non-Latin scripts.

Language	NLLB	mBART	mT5	Mistral	Bloom	Llama2
Somali$_{L}$	-	-	-	-	-	-
Swahili$_{L}$	-	-	-	-	-	-
Yoruba$_{L}$	10.07	**44.22**	5.10	13.06	12.13	13.45
Pashto$_{NL}$	2.80	1.82	10.42	**15.28**	12.23	13.85
Kannada$_{NL}$	1.83	11.36	12.50	15.40	9.53	**17.32**
Sinhala$_{NL}$	1.91	**34.99**	6.92	7.80	4.84	12.60
Marathi$_{NL}$	4.76	**12.71**	5.10	10.31	6.51	8.93
Punjabi$_{NL}$	2.90	1.73	5.47	11.17	10.28	**15.68**
Tajik$_{NL}$	9.58	2.60	10.36	**21.10**	20.36	21.05
Kyrgyz$_{NL}$	4.23	5.44	9.49	**15.80**	9.31	13.04
Telugu$_{NL}$	**11.67**	3.47	8.62	9.47	9.08	11.11
Amharic$_{NL}$	-	-	-	-	-	-
Burmese$_{NL}$	2.80	1.69	5.21	8.93	5.87	**16.00**
Nepali$_{NL}$	1.46	1.12	6.56	5.84	6.82	**6.86**
AVG	4.91	11.01	7.79	12.19	9.72	**13.62**
Median	2.90	3.47	6.92	11.17	9.31	**13.45**

Table 1: Translation performance (average ChrF++ score) of LLMs across languages on Opus100 dataset.

Performance analysis: no single model dominates across all languages; performance largely depends on the overlap between the model's training data and the target language. Llama2's strong average performance is likely due to its robust architecture, which includes dense attention and efficient decoding mechanisms. A critical insight from our analysis is the impact of script disparity—specifically, Latin vs. non-Latin scripts. Non-Latin languages often suffer due to several factors:

- Tokenizer bias: BPE or SentencePiece vocabularies are typically dominated by Latin-script tokens, leading to inefficient tokenization of other scripts.
- Data scarcity: LRLs written in non-Latin scripts generally have less high-quality parallel data available.
- Model bias: Models may favor outputs in high-resource, often Latin-script languages, resulting in language interference or degraded fluency in other languages.

Our findings illustrate that multilingual model performance is uneven, shaped by language scripts, data representation, and model architecture. Even the strongest models frequently underperform in non-Latin, low-resource settings—underscoring the urgent need for more balanced training corpora and improved tokenization strategies.

5.2 Evaluation on Text Summarization

Results: Table 2 shows the performance (ROUGE score) of different LLMs in summarizing text across the selected languages. The results indicate that the mT5-multilingual-XLSum model consistently outperforms the others in almost all languages and achieves the highest average score. However, overall, all models perform worse on languages with non-Latin scripts—such as Nepali, Amharic, Telugu, Sinhala, and Pashto—compared to languages with Latin scripts like Somali, Swahili, and Yoruba, highlighting a bias toward Latin-script languages in this task.

Performance analysis: The mT5-multilingual-XL-Sum model is based on the mT5 checkpoint, fine-tuned on the XLSum dataset, which includes high-quality news summaries in over 45 languages, many of which are low-resource and use Latin scripts. This specialization makes it well-suited for summarization, unlike general-purpose models such as LLaMA2, Mistral, or Bloom, which are primarily trained for open-ended text generation.

Language	Llama-2		Mistral		mT5-multi		Bloom		mT5-base	
	R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
Somali$_{L}$	7.42	5.19	7.03	4.82	**34.42**	**25.84**	7.18	5.28	12.01	9.68
Swahili$_{L}$	6.56	4.87	6.50	4.77	**39.02**	**31.75**	6.20	5.04	13.32	11.21
Yoruba$_{L}$	8.66	6.28	8.68	6.25	**39.32**	**29.99**	7.77	6.16	16.88	13.57
Pashto$_{NL}$	0.01	0.01	**0.02**	**0.02**	0.00	0.00	0.00	0.00	0.00	0.00
Kannada$_{NL}$	-	-	-	-	-	-	-	-	-	-
Sinhala$_{NL}$	0.57	0.57	0.81	0.81	**2.33**	**2.33**	0.77	0.77	0.03	0.03
Marathi$_{NL}$	0.85	0.85	0.85	0.85	**4.16**	**4.16**	1.20	1.15	0.77	0.77
Punjabi$_{NL}$	1.39	1.39	1.43	1.43	**5.00**	**5.00**	1.06	1.04	0.83	0.64
Tajik$_{NL}$	-	-	-	-	-	-	-	-	-	-
Kyrgyz$_{NL}$	1.69	1.67	1.58	1.56	**8.63**	**8.63**	2.34	2.25	1.21	1.21
Telugu$_{NL}$	1.80	1.80	1.97	1.97	**3.06**	**3.06**	1.06	1.06	0.26	0.26
Amharic$_{NL}$	0.71	0.71	0.68	0.68	**3.00**	**3.00**	0.96	0.82	0.53	0.51
Burmese$_{NL}$	3.89	3.89	4.05	4.05	**7.27**	**7.27**	4.21	4.21	0.30	0.30
Nepali$_{NL}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AVG	2.80	2.27	2.80	2.27	**12.18**	10.09	2.73	2.32	3.85	3.18
Median	1.54	1.53	1.51	1.50	**4.58**	**4.58**	1.13	1.11	0.65	0.58

Table 2: Text summarization performance (ROUGE) of LLMs across languages on the XL-Sum dataset.

Moreover, mT5-multilingual-XL-Sum is explicitly multilingual and designed to support LRLs by leveraging a shared vocabulary and cross-lingual transfer, helping it generalize better to low-resource Latin-script languages compared to models like LLaMA2 or Bloom, which may have limited exposure to such languages during pretraining. Additionally, this model benefits from better tokenizer compatibility with Latin scripts, as Latin-script languages tend to be tokenized more efficiently by subword tokenizers like SentencePiece—especially when the model has encountered similar orthographies during training.

It is worth mentioning that LLaMA 2 and Mistral seem to perform similar to Bloom with a much smaller token vocabulary.

5.3 Evaluation on Question Answering

Results: Table 3 shows QA performance (F1 scores in %) across different language scripts on the Belebele dataset. The averaged F1 scores, (i.e., mean) over all languages across all models range between 32 and 36, with median values between 33 and 36.

Bloom and Qwen have the highest median (36.22 and 36.10) and averaged F1 score (35.59 and 34.54) values, indicating that these models may offer the best performance overall in this evaluation. XGLM has the lowest averaged F1 score and median, potentially the least performance system.

LLaMA2 and mT5 appear to have balanced data distributions, showing the smallest difference between the averaged F1 scores and median scores, with the median slightly below the average. In contrast, the other models contain low-value outliers that pull the averaged F1 score below the median.

For all models, the F1 scores for Burmese fall below the model-specific median, suggesting it's among the most challenging languages tested.

A closer examination shows that LLaMA2 outperforms other models on Somali, Pashto, Punjabi, and Telugu; Mistral performs best on Amharic; Bloom excels on Swahili, Sinhala, Tajik, Kyrgyz, and Nepali; Qwen leads on Kannada, Marathi, and Burmese; and mT5-base performs best on Yoruba.

Some models seem to archive more concise performance on Latin-script languages compared to non-Latin, where performance decreases—for instance, LLaMA2 on Kannada, Kyrgyz, and Burmese; Mistral on Telugu and Tajik; Bloom on Somali and Burmese; Qwen on Tajik and Kyrgyz; and mT5-base on Telugu, Amharic, and Nepali.

Language	Llama2	Mistral	Bloom	Qwen	mT5	Xglm
Somali$_{L}$	**41.43**	35.37	26.53	37.57	32.84	29.03
Swahili$_{L}$	38.99	37.14	**48.84**	35.93	38.53	34.38
Yoruba$_{L}$	33.44	37.93	38.61	30.24	**44.03**	24.06
Pashto$_{NL}$	**41.28**	38.40	37.60	38.98	33.06	40.66
Kannada$_{NL}$	24.63	33.93	31.94	**42.97**	33.33	27.17
Sinhala$_{NL}$	35.51	28.22	**38.94**	38.23	34.06	26.59
Marathi$_{NL}$	28.93	30.70	31.75	**39.00**	31.87	33.07
Punjabi$_{NL}$	**41.20**	36.99	34.86	36.64	38.07	28.57
Tajik$_{NL}$	31.30	29.50	**37.58**	23.55	34.53	33.43
Kyrgyz$_{NL}$	26.62	36.19	37.60	27.59	30.27	**38.20**
Telugu$_{NL}$	**38.31**	21.58	32.37	36.26	25.51	34.60
Amharic$_{NL}$	29.24	**36.42**	33.53	31.24	28.06	23.97
Burmese$_{NL}$	27.02	31.00	28.83	**33.04**	29.20	32.12
Nepali$_{NL}$	33.52	36.52	39.34	32.37	28.65	**43.97**
AVG	33.67	33.56	**35.59**	34.54	33.00	32.13
Median	33.48	35.78	**36.22**	36.10	32.95	32.60

Table 3: QA performance (F1 score) of LLMs across languages on the Belebele dataset.

Performance analysis: Each model appears to have language-specific strengths, reflecting differ-

ences in pretraining data, architecture, or cross-lingual generalization ability. Additionally, the variety of languages (Afro-Asiatic, Indo-Aryan, Dravidian, etc.) indicates that these models exhibit partial generalization, likely influenced by the distribution of their training data. Script sensitivity—Latin versus non-Latin scripts—is also apparent, likely due to sparse pretraining data in non-Latin scripts (e.g., Kannada, Telugu, Burmese, Amharic) and tokenizer inefficiencies, as many LLMs use subword tokenizers trained predominantly on Latin-based corpora. This highlights that non-Latin scripts continue to pose challenges for LLMs, mainly in QA.

6 Tokenization Analysis

We conducte further experiments on the CulturaX dataset to assess how well a tokenizer understands diverse LRLs.

Our analysis considers the following metrics:

OOV Rate: Indicates the proportion of words not found in the vocabulary; higher values suggest lower coverage.

Vocab Coverage: Higher values indicate better vocabulary coverage.

Sub-word Fragmentation: Higher values indicate that more words are segmented into multiple subword units, suggesting less efficient tokenization.

Tokens/Word: Lower values mean fewer tokens per word on average, suggesting more concise word representations.

6.1 Overall Results

Appendix B contains full details and results for all LRLs, with Figures 5 to 8 illustrating the metrics across models and languages.

Our tokenization analysis reveals consistently very high OOV rates and very low vocabulary coverages across all models and languages. XGLM achieved the best average score for both, with the lowest OOV rate (0.84) and the highest vocabulary coverage (0.15), both for Yoruba, a Latin-script language. These rates indicate limited generalization and semantic understanding of the evaluated models on LRLs.

Across all models, the highest OOV rates and lowest vocabulary coverages, i.e., indicating the most challenging languages, were observed for Amharic (Ge'ez script), Pashto (Arabic script), Tajik (Cyrillic script), and Kyrgyz (Cyrillic script).

Regarding sub-word fragmentation, LLaMA2 and Mistral exhibited the highest values (over 0.68) across all languages. This may be due to the small vocabulary size of both models compared to the other models. Notably, BLOOM and Qwen showed more efficient tokenization compared to other models for Nepali and Marathi. BLOOM further reached below 0.1 for Telugu, Kannada, Marathi, and Punjabi.

XGLM achieved the best token-per-word ratios (i.e., the lowest values) across the majority of languages, while LLaMA2 consistently showed the worst performance in this regard.

6.2 LLaMA2 Tokenizer Results

The radar chart in Figure 4 provides valuable insights how LLaMA2 might perform across LRLs, such as Amharic, Kannada, Nepali, Pashto, Tajik, Swahili, and Punjabi, based on the analyzed linguistic characteristics:

Amharic: Amharic's extremely high average tokens per word (10.00) and highest OOV rate (0.99), even with the LLaMA2 tokenizer, strongly suggest that the tokenizer is highly inefficient for

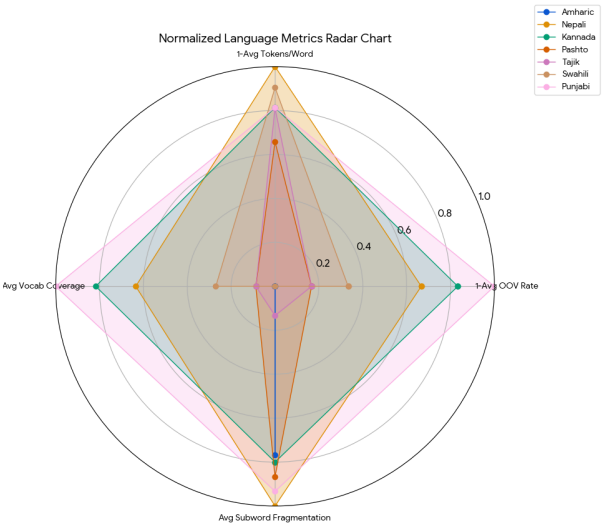

Figure 4: Token analysis with respect to the average out-of-vocabulary (OOV) rate, average tokens per word, average vocabulary coverage, and average sub-word fragmentation rate per language was performed using the LLaMA2 tokenizer. Inverted values mean that a value closer to 0 on the chart axis corresponds to a higher actual OOV rate, while a value closer to 1 indicates a lower actual OOV rate. The same inversion applies to the average tokens per word metric.

Amharic. This is likely due to its unique Ethiopic script and complex morphology, which do not align well with the BPE patterns learned by LLaMA2. The tokenizer breaks words into many small, often character-level tokens that do not form meaningful subword units for the LLM.

Pashto and Tajik: Similar to Amharic, their high OOV rates (0.98) and low vocabulary coverage with the LLaMA2 tokenizer indicate inefficiency. While their average tokens per word is lower than Amharic's, it remains notably higher than in languages like Nepali or Punjabi. This suggests that despite being trained on diverse data, the LLaMA2 tokenizer struggles with these Persian-derived languages, possibly due to script variations or morphological features not well represented in the tokenizer's BPE training.

Swahili: Swahili's lower subword fragmentation rate (0.68) compared to other languages, combined with a high OOV rate (0.97), suggests that the LLaMA2 tokenizer is not effectively segmenting Swahili words into useful subword units that are well-covered in its vocabulary. This may be due to its Bantu agglutinative morphology, which forms word structures that LLaMA2's BPE does not optimally capture.

Nepali and Punjabi: These languages generally show more favorable characteristics with the LLaMA2 tokenizer. They exhibit high subword fragmentation rates (0.98 for Nepali, 0.96 for Punjabi), which are accompanied by lower OOV rates and better vocabulary coverage. This indicates that for these languages, the tokenizer's BPE method effectively breaks down words into meaningful and well-covered subword units.

Kannada: Good vocabulary coverage and high subword fragmentation suggest that the LLaMA2 tokenizer handles Kannada relatively well in breaking down words into known subwords. However, its average tokens per word is still moderate, indicating some degree of fragmentation.

The analysis clearly shows that the effectiveness of the LLaMA2 tokenizer varies significantly across languages, particularly for those outside its primary training distribution or with distinct linguistic features (e.g., non-Latin scripts or complex morphology). While it performs reasonably well for languages like Nepali and Punjabi—benefiting from effective subword fragmentation—it presents substantial challenges for languages such as Amharic, Pashto, and Tajik. In these cases, inefficient tokenization negatively affects the model's context understanding, increases computational cost, and reduces overall output quality. These findings highlight the critical need for language-specific tokenizers or multilingual tokenizers explicitly designed to handle a wider range of linguistic features, especially for low-resource and linguistically diverse languages, to ensure optimal LLM performance.

7 Conclusion

In this work, we present a comprehensive study evaluating nine publicly available LLMs, commonly used via Hugging Face, on three core NLP tasks—machine translation, text summarization, and question answering—with a particular focus on LRLs. We assessed the performance of these LLMs across 14 languages ranging from class 0 to class 2. In addition, we conducted a tokenization analysis on LRLs. Our findings highlight the challenges and limitations of evaluating LLMs on LRLs, primarily due to the scarcity of training data and the diversity of writing scripts. To address these limitations and advance the state of the art, future research should explore the development of specialized models tailored to LRLs, including those that use non-Latin scripts.

8 Limitations

While our study provides valuable insights into the performance of multilingual LLMs on LRLs, We acknowledge two main limitations: (i) our evaluation focused on a subset of publicly available LLMs and multilingual benchmark datasets. Given the vast number of models and resources available, we selected widely used and openly accessible LLMs from the Hugging Face Hub; and (ii) the multilingual LLMs and tokenizers analyzed in this study were not specifically optimized or customized for the selected LRLs. We believe future research should explore the development of LLMs tailored to LRLs, 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 Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen 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. *Preprint*, arXiv:2103.11811.

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.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Preprint*, arXiv:2404.05590.

Hadi Askari, Anshuman Chhabra, Muhao Chen, and Prasant Mohapatra. 2024. Assessing llms for zero-shot abstractive summarization through the lens of relevance paraphrasing. *Preprint*, arXiv:2406.03993.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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. 2023a. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

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. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

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. 2020a. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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. 2020b. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. 2023. Quadro: Dataset and models for question-answer database retrieval. *Preprint*, arXiv:2304.01003.

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. *Preprint*, arXiv:2307.03109.

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. *J. Mach. Learn. Res.*, 24(1).

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. *Preprint*, arXiv:2212.05409.

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. *Preprint*, arXiv:2302.01398.

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

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. *Preprint*, arXiv:2302.09210.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

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. *Preprint*, arXiv:2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world. *Preprint*, arXiv:2004.09095.

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. *Preprint*, arXiv:2307.16039.

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. *Preprint*, arXiv:2211.09110.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *Preprint*, arXiv:2004.01401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

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. *Preprint*, arXiv:2403.10258.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. *Preprint*, arXiv:2104.08773.

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.

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. *Preprint*, arXiv:2203.02155.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *Preprint*, arXiv:2401.01854.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

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. *Preprint*, arXiv:2207.04672.

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. *Preprint*, arXiv:2302.13971.

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. *Preprint*, arXiv:2009.05387.

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. *Preprint*, arXiv:2211.05100.

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. *Preprint*, arXiv:2303.02913.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. 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.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A General Appendix

ISO	Language	Script	Family	Speakers	C
am	Amharic	Ge'ez	Afro-Asiatic	57M	2
te	Telugu	Telugu	Dravidian	96M	1
my	Burmese	Burmese	Sino-Tibetan	42.9M	1
ne	Nepali	Devanagari	Indo-European-Indo-Aryan	32M	1
kn	Kannada	Kannada	Dravidian	44M	1
ps	Pashto	Arabic	Indo-European	55M	1
tg	Tajik	Cyrillic	Indo-European	10.5M	1
sw	Swahili	Latin	Atlantic-Congo	200M	2
yo	Yoruba	Latin	Atlantic-Congo	46M	2
so	Somali	Latin	Afro-Asiatic	22M	1
si	Sinhala	Sinhala	Indo-European	16M	0
mr	Marathi	Devanagari	Indo-European	83M	2
pa	Punjabi	Gurmukhi	Indo-European	150M	2
ky	Kyrgyz	Cyrillic	Turkic	4.12M	2

Table 4: We provide a few selected LRLs used in our evaluation including the ISO-639-3 language code (ISO), language script (Script), language family, total numbers of speakers, and language class (C).

Language	ISO	Script	Family	Mono Data	C
Amharic	amh	Ge'ez	Afro-Asiatic	3.02M	2
Arabic	ara	Arabic	Afro-Asiatic	126M	5
Azerbaijani	azj	Latin	Turkic	41.4M	1
Bengali	ben	Bengali	Indo-European	57.9M	3
Burmese	mya	Myanmar	Sino-Tibetan	2.66M	1
Chinese (Simplified)	zho	Han	Sino-Tibetan	209M	4
Chinese (Traditional)	zho	Han	Sino-Tibetan	85.2M	4
English	en	Latin	Indo-European	-	5
French	fra	Latin	Indo-European	428M	5
Gujarati	guj	Gujarati	Indo-European	9.41M	1
Hausa	Hau	Latin	Afro-Asiatic	5.87M	2
Hindi	Hi	Devanagari	Indo-European	104M	4
Igbo	ibo	Latin	Atlantic-Congo	693K	1
Indonesian	ind	Latin	Austronesian	1.05B	3
Japanese	jpn	Han, Hiragana, Katakana	Japonic	282.9M	5
Kirundi	rn	-	-	-	1
Korean	kor	Hangul	Koreanic	390M	4
Kyrgyz	kir	Cyrillic	Turkic	2.02M	-
Marathi	mar	Devanagari	Indo-European	14.4M	2
Nepali	npi	Devanagari	Indo-European	17.9M	1
Oromo	orm	Latin	Afro-Asiatic	752K	1
Pashto	pus	Perso-Arabic	Indo-European	12M	1
Persian	fas	Perso-Arabic	Indo-European	611M	4
Pidgin	n/a	-	-	-	0
Portuguese	por	Latin	Indo-European	340M	4
Punjabi	pan	Gurmukhi	Indo-European	5.02M	2
Scottish (Cyrillic)	gd	-	-	-	1
Serbian (Latin)	srp	Cyrillic	Indo-European	225M	4
Sinhala	si	-	-	-	0
Somali	som	Latin	Afro-Asiatic	14.1M	1
Spanish	spa	Latin	Indo-European	379M	5
Swahili	swh	Latin	Atlantic-Congo	35.8M	2
Tamil	tam	Tamil	Dravidian	68.2M	3
Telugu	tel	Telugu-Kannada	Dravidian	282.9M	1
Thai	tha	Thai	Kra-Dai	319M	3
Tigrinya	tir	Ge'ez	Afro-Asiatic	-	2
Turkish	tur	Latin	Turkic	128M	4
Ukrainian	ukr	Cyrillic	Indo-European	357M	3
Urdu	urd	Perso-Arabic	Indo-European	28M	3
Uzbek	uzb	Latin	Turkic	7.54M	3
Vietnamese	vie	Latin	Austro-Asiatic	992M	4
Welsh	cym	Latin	Indo-European	12.7M	1
Yoruba	yor	Latin	Atlantic-Congo	1.59M	2

Table 5: List of languages included in the XLSum dataset, along with the corresponding ISO 639-3 , Script, Language family and Class.

A.1 Pre-training Model Details

We report the pre-training details of each model in Table 6.

A.2 NLP tasks

- **Machine translation:** For this task, we evaluate the 3.3B version of NLLB-200 designed for single sentence translation among 200 languages; mBART-large-50-many-to-many-mmt fine-tuned for multilingual machine translation on 53 natural languages; mT5-base, a model covering 101 natural languages; the 7B version of Llama2, trained on 20 natural languages; the 7B version of BLoom, trained on 46 natural languages and the 7B version of Mistral, trained on 6 natural languages.

- **Text Summarization:** For this task, we use the fine-tuned variants of mT5 trained on 45 natural languages of XL-Sum; mT5-base; the 7B version of Bloom; the 7B version of Mistral and the 7B version of Llama2.

- **Question Answering:**: We use the 7B version of Bloom; the 7B version of Mistral; the 7B version of Llama2; the 7.5B version of XGLM trained on 31 natural languages; mT5-base; the 8B version of Qwen3 and the 7B version of Mistral.

A.3 Language Details

In Table 4 we provide an brief overview of LRLs we include in our evaluation.

A.4 XLsum languages

In Table 5, we list all the languages included in XLSum benchmark dataset with some details including the iso code, language family, script and class.

B Token analysis

We report the token analysis for all models in Table 7. For all models the train_samples are 50000, analysis_samples are 10000, target_vocab_size is 30000, and min_frequency is 5.

Model	Type	Tokenizer	Pre-training data (Tokens)	Languages	Low-resource strength
XGLM	Decoder-only (GPT-style)	SentencePiece (BPE)	500B	30	Excellent (gen/QA)
Qwen3	Decoder-only	Custom BPE tokenizer (QwenTokenizer)	3T (estimated)	Partial	Moderate
Llama2	Decoder-only	SentencePiece (BPE)	2T	20 (English-heavy)	Weak
Bloom	Decoder-only	GPT2-style BPE	1.6T	46	Good
Mistral	Decoder-only	SentencePiece (BPE)	1–2T	Some	Limited
mT5-base	Encoder-decoder (Seq2Seq)	SentencePiece (Unigram)	250k	101	Strong
NLLB	Encoder-decoder (based on mBART)	SentencePiece	1.3T	200	Best (translation)
mBART-large	Encoder-decoder	entencePiece	250GB	25	Good (but limited)
mT5-multilingual-XLSum	mT5-base fine-tuned for summarization	SentencePiece (Unigram)	250k	45	Strong (summarization)

Table 6: Pre-training details of each model.

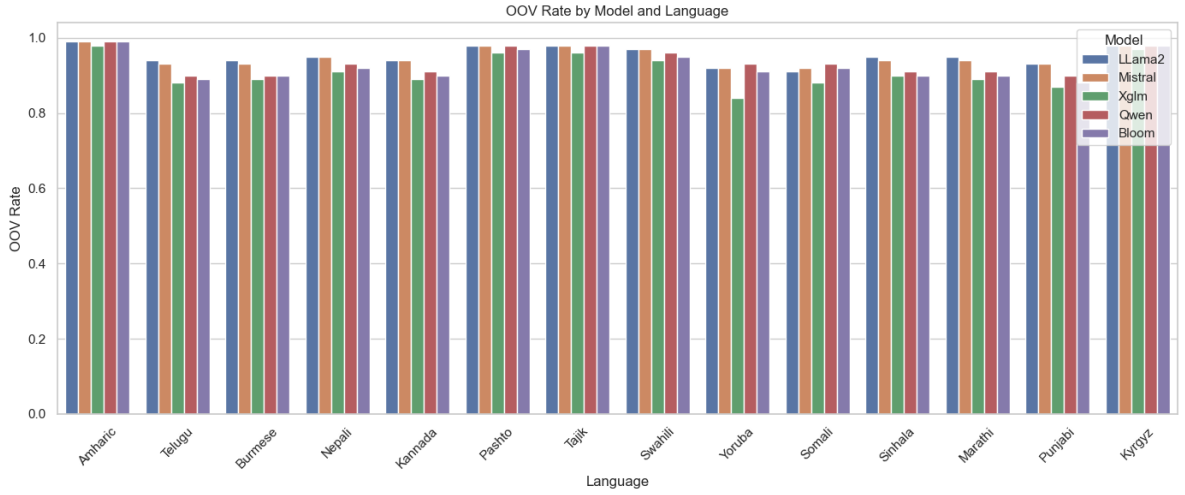

Figure 5: The avg. out-of-vocabulary rate by model and language.

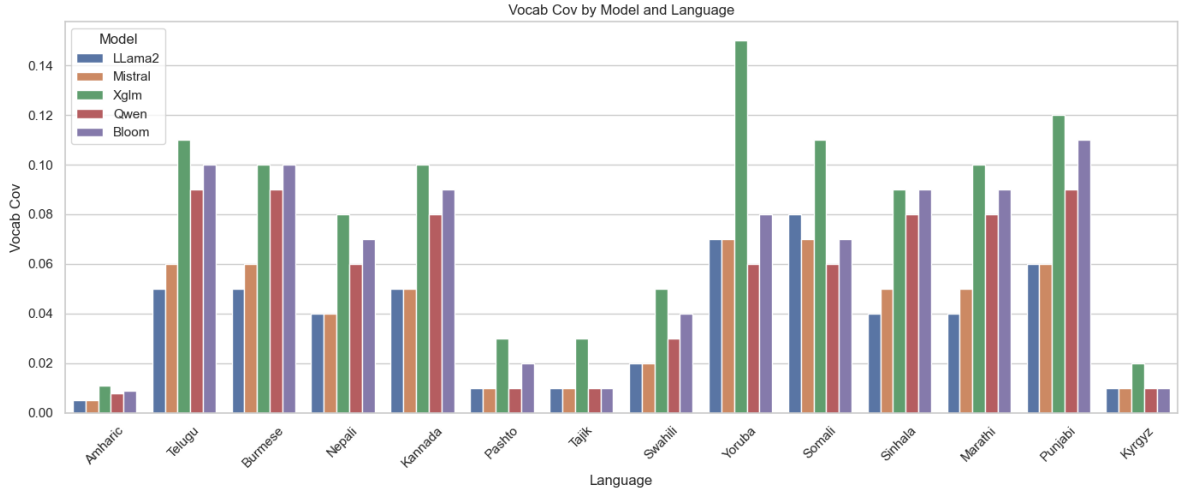

Figure 6: The avg. vocab coverage by model and language.

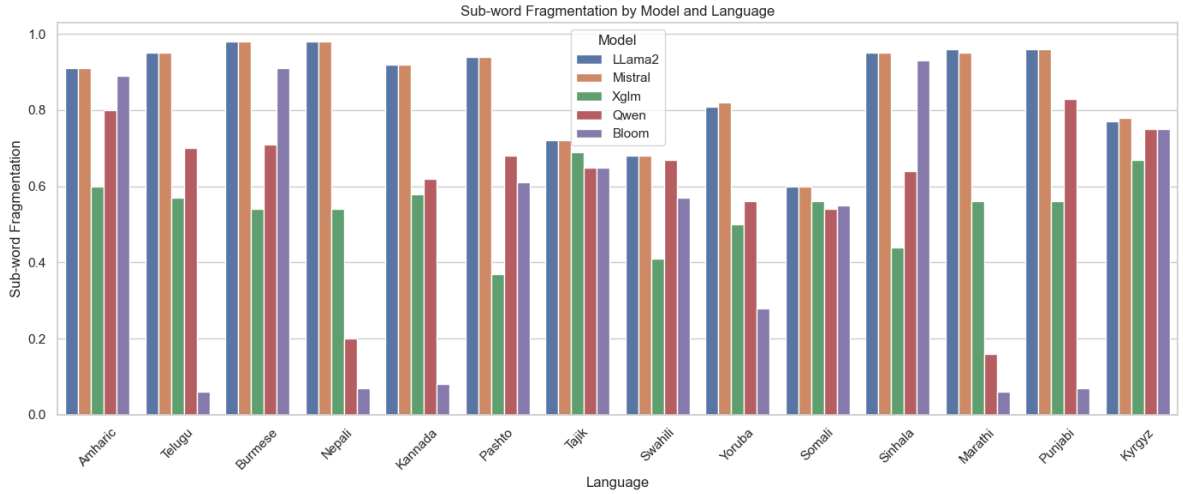

Figure 7: The avg. sub-word fragmentation by model and language.

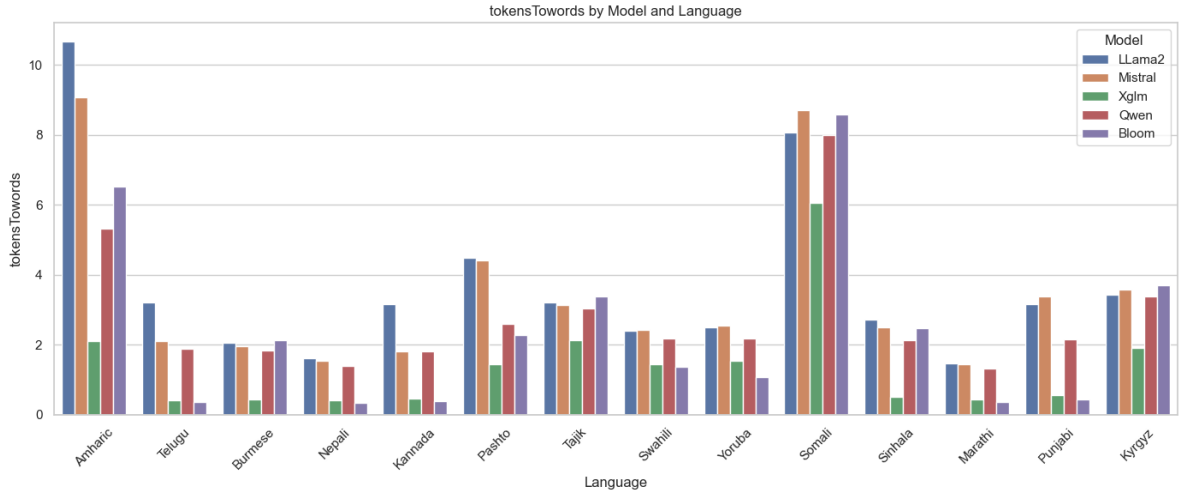

Figure 8: The avg. Tokens/Word by model and language.

Language	Tokenizer	actual_tokenizer_vocab_size	processed_analysis_texts	avg_tokens_per_word	avg_oov_rate	avg_vocab_coverage	avg_subword_fragmentation_rate
Amharic	LLama2	32000	10000	10.67	0.99	0.005	0.91
	Mistral	32000	10000	9.09	0.99	0.005	0.91
	Xglm	256008	10000	2.11	0.98	0.011	0.60
	Qwen	151669	10000	5.33	0.99	0.008	0.80
	Bloom	250680	10000	6.52	0.99	0.009	0.89
Telugu	LLama2	32000	10000	3.21	0.94	0.05	0.95
	Mistral	32000	10000	2.11	0.93	0.06	0.95
	Xglm	256008	10000	0.42	0.88	0.11	0.57
	Qwen	151669	10000	1.89	0.90	0.09	0.70
	Bloom	250680	10000	0.37	0.89	0.10	0.06
Burmese	LLama2	32000	10000	2.07	0.94	0.05	0.98
	Mistral	32000	10000	1.97	0.93	0.06	0.98
	Xglm	256008	10000	0.45	0.89	0.10	0.54
	Qwen	151669	10000	1.83	0.90	0.09	0.71
	Bloom	250680	10000	2.14	0.90	0.10	0.91
Nepali	LLama2	32000	10000	1.61	0.95	0.04	0.98
	Mistral	32000	10000	1.54	0.95	0.04	0.98
	Xglm	256008	10000	0.41	0.91	0.08	0.54
	Qwen	151669	10000	1.40	0.93	0.06	0.20
	Bloom	250680	10000	0.35	0.92	0.07	0.07
Kannada	LLama2	32000	10000	3.17	0.94	0.05	0.92
	Mistral	32000	10000	1.82	0.94	0.05	0.92
	Xglm	256008	10000	0.47	0.89	0.10	0.58
	Qwen	151669	10000	1.82	0.91	0.08	0.62
	Bloom	250680	10000	0.39	0.90	0.09	0.08
Pashto	LLama2	32000	10000	4.48	0.98	0.01	0.94
	Mistral	32000	10000	4.42	0.98	0.01	0.94
	Xglm	256008	10000	1.45	0.96	0.03	0.37
	Qwen	151669	10000	2.61	0.98	0.01	0.68
	Bloom	250680	10000	2.29	0.97	0.02	0.61
Tajik	LLama2	32000	10000	3.20	0.98	0.01	0.72
	Mistral	32000	10000	3.15	0.98	0.01	0.72
	Xglm	256008	10000	2.14	0.96	0.03	0.69
	Qwen	151669	10000	3.03	0.98	0.01	0.65
	Bloom	250680	10000	3.39	0.98	0.01	0.65
Swahili	LLama2	32000	10000	2.41	0.97	0.02	0.68
	Mistral	32000	10000	2.42	0.97	0.02	0.68
	Xglm	256008	10000	1.44	0.94	0.05	0.41
	Qwen	151669	10000	2.19	0.96	0.03	0.67
	Bloom	250680	10000	1.36	0.95	0.04	0.57
Yoruba	LLama2	32000	192	2.50	0.92	0.07	0.81
	Mistral	32000	192	2.54	0.92	0.07	0.82
	Xglm	256008	192	1.54	0.84	0.15	0.50
	Qwen	151669	192	2.17	0.93	0.06	0.56
	Bloom	250680	192	1.08	0.91	0.08	0.28
Somali	LLama2	32000	39	8.07	0.91	0.08	0.60
	Mistral	32000	39	8.70	0.92	0.07	0.60
	Xglm	256008	39	6.06	0.88	0.11	0.56
	Qwen	151669	39	7.99	0.93	0.06	0.54
	Bloom	250680	39	8.59	0.92	0.07	0.55
Sinhala	LLama2	32000	10000	2.72	0.95	0.04	0.95
	Mistral	32000	10000	2.51	0.94	0.05	0.95
	Xglm	256008	10000	0.50	0.90	0.09	0.44
	Qwen	151669	10000	2.14	0.91	0.08	0.64
	Bloom	250680	10000	2.48	0.90	0.09	0.93
Marathi	LLama2	32000	10000	1.47	0.95	0.04	0.96
	Mistral	32000	10000	1.45	0.94	0.05	0.95
	Xglm	256008	10000	0.43	0.89	0.10	0.56
	Qwen	151669	10000	1.32	0.91	0.08	0.16
	Bloom	250680	10000	0.36	0.90	0.09	0.06
Punjabi	LLama2	32000	10000	3.17	0.93	0.06	0.96
	Mistral	32000	10000	3.38	0.93	0.06	0.96
	Xglm	256008	10000	0.57	0.87	0.12	0.56
	Qwen	151669	10000	2.16	0.90	0.09	0.83
	Bloom	250680	10000	0.43	0.88	0.11	0.07
Kyrgyz	LLama2	32000	10000	3.44	0.98	0.01	0.77
	Mistral	32000	10000	3.57	0.98	0.01	0.78
	Xglm	256008	10000	1.91	0.97	0.02	0.67
	Qwen	151669	10000	3.39	0.98	0.01	0.75
	Bloom	250680	10000	3.70	0.98	0.01	0.75

Table 7: We report Token analysis of a batch of training samples by computing the average tokens per word, the average out-of-vocabulary rate, the average vocabulary coverage and average sub-word fragmentation rate of each of the language using Llama2, Mistral, Xglm, Qwen and Bloom tokenizers .