Do All Languages Cost the Same?
Tokenization in the Era of Commercial Language Models

Orevaoghene Ahia[◇] **Sachin Kumar**[♠♡] **Hila Gonen**[◇] **Jungo Kasai**[◇]
David R. Mortensen[♠] **Noah A. Smith**[◇♡] **Yulia Tsvetkov**[◇]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Language Technologies Institute, Carnegie Mellon University
[♡]Allen Institute for Artificial Intelligence
{oahia,jkasai,nasmith,yuliats}@cs.washington.edu, sachink@allenai.org
hilagnn@gmail.com, dmortens@cs.cmu.edu

Abstract

Language models have graduated from being research prototypes to commercialized products offered as web APIs, and recent works have highlighted the multilingual capabilities of these products. The API vendors charge their users based on usage, more specifically on the number of "tokens" processed or generated by the underlying language models. What constitutes a token, however, is training data and model dependent with a large variance in the number of tokens required to convey the same information in different languages. In this work, we analyze the effect of this non-uniformity on the fairness of an API's pricing policy across languages. We conduct a systematic analysis of the cost and utility of OpenAI's language model API on multilingual benchmarks in 22 typologically diverse languages. We show evidence that speakers of a large number of the supported languages are overcharged while obtaining poorer results. These speakers tend to also come from regions where the APIs are less affordable to begin with. Through these analyses, we aim to increase transparency around language model APIs' pricing policies and encourage the vendors to make them more equitable.

1 Introduction

Language models (LMs) have come to be known as general-purpose solutions capable of performing many tasks by following natural language instructions (Brown et al., 2020; Ouyang et al., 2022; Chung et al., 2022), and generalizing to new tasks at test time using a handful of demonstrations (Brown et al., 2020; Su et al., 2023). Motivated by their potential for commercial use, many industrial research institutions have moved away from openly releasing them (Abdalla et al., 2023). Instead, a new business model of LM as a Service (Sun et al., 2022) has emerged where LMs can be accessed for inference using (paid) web APIs. The majority of these models (Ouyang et al., 2022) offer multilingual capabilities, and the API providers charge the users proportionally to the number of tokens processed or generated.

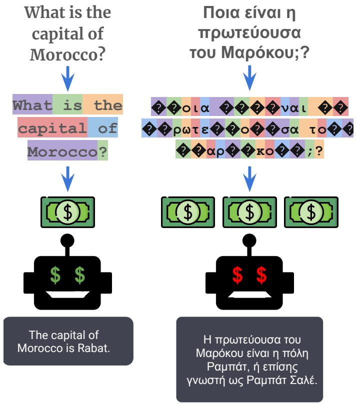

What is the capital of Morocco?
Ποια είναι η πρωτεύουσα του Μαρόκου;?
What is the capital of Morocco?
��οια ����ναι ��
��ρωτε��ο��σα το��
��αρ��κο��;?
$ $
$ $
The capital of Morocco is Rabat.
Η πρωτεύουσα του Μαρόκου είναι η πόλη Ραμπάτ, ή επίσης γνωστή ως Ραμπάτ Σαλέ.

Figure 1: We investigate the effects of subword tokenization in LMs across languages with different writing systems. Our findings highlight disparities in the utility of LMs, as well as socio-economic disparities and increased costs in using commercial APIs for speakers of underrepresented languages.[1]

In this work, we examine the fairness of this pricing model for different languages, based on how a "token" is defined in practice.[2] Most LMs rely on tokenizers that split text strings into chunks (subwords). Subword tokenizers (Sennrich et al., 2016; Kudo, 2018; Song et al., 2020) are typically data-driven and learn to split text based on frequency patterns of characters or bytes in some corpus. Prior work argued that, in multilingual settings, subword tokenizers lead to disproportion-

[1] OpenAI's tokenizer interface displays byte tokens absent from their vocabulary as "?".

[2] Code available at https://github.com/orevaahia/llm_tokenizer_cost

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9904–9923
December 6-10, 2023 ©2023 Association for Computational Linguistics

ate fragmentation rates for different languages and writing scripts (Zhang et al., 2022a; Rust et al., 2021; Muller et al., 2021). Many commercial LMs are multilingual, and text from languages that suffer from excessive fragmentation will be represented using more tokens. This directly increases cost of API usage for certain language speakers, even if they convey the same information as the others.

We highlight this unfairness through three stages of systematic analyses. First, we show evidence that tokenizers of popular LMs indeed over-fragment texts in certain language scripts and quantify the API cost disparity that this issue causes. We discover that the disparity is not caused just by data imbalance, but is rooted in the language properties or the ways they are represented in Unicode. Second, we show that languages with longer token lengths as a result of greater fragmentation derive less model utility with in-context learning (Brown et al., 2020). Finally, we find that languages that cost more and perform worse are often associated with populations of speakers for whom the APIs are less affordable on average, exacerbating the economic divide in the accessibility of NLP technology.

Through these analyses, we argue that commercial LM API vendors should revisit their processing and pricing strategies to be more equitable. In addition, we encourage the NLP community to pay better attention to tokenizers, an often neglected part of the LM pipeline.

2 Do All Languages Cost the Same?

2.1 Background

Language Model APIs Autoregressive LMs are trained to predict the next "token" given a previous context. Following the success of such models, many commercial LM web APIs have emerged and allow users to interface with the models using natural language instructions to perform various tasks with little to no exposure to the underlying workings of the models. The API providers often support dozens of languages and charge users[3] at a fixed rate based on the total number of input and generated tokens.[4] What constitutes a "token," however, is not a universally accepted definition but a design choice that the model developers make. The total token count is also not immediately obvious to users except through a tokenizer interface[5] separate from the chat interface.

Tokenization in LMs Tokenization—segmenting text into atomic units—is an active research area. Proposed approaches range from defining tokens as whitespace-delimited words (for languages that use whitespace) which makes the vocabulary extremely large, to defining tokens as characters or bytes, making the tokenized sequences extremely long in terms of number of tokens; see Mielke et al. (2021) for a detailed survey. A commonly-used solution now is to tokenize text into *sub*word chunks. With Sennrich et al. (2016), one starts with a base vocabulary of only characters adding new vocabulary items by recursively merging existing ones based on their frequency statistics in the data. Other approaches judge subword candidates to be included in the vocabulary using an LM (Kudo, 2018; Song et al., 2021). For multilingual models containing data in a variety of scripts, even the base vocabulary of only characters (based on Unicode symbols) can be very large with over 130K types. Radford et al. (2019) instead proposed using a byte-level base vocabulary with only 256 tokens. Termed byte-level byte pair encoding (BBPE), this approach has become a de facto standard used in most modern language modeling efforts (Brown et al., 2020; Muennighoff et al., 2022; Scao et al., 2022; Black et al., 2022; Rae et al., 2022; Zhang et al., 2022b). In this work, we investigate the impact this tokenization strategy has on LM API cost disparity as well as downstream task performance (i.e., utility) across different languages.

2.2 Investigating the Impact of Byte-level Subword Segmentation

There are hundreds of distinct writing systems in the world (Hockett, 1997). BBPE, by design, makes vocabulary construction script-agnostic, allowing (in principle) new scripts to be supported later on without modifying the vocabulary. However, not only are different scripts encoded differently, their distribution in the training corpora varies widely. To investigate the effects of this variation, we propose the following research questions as the main focus of this work.

[3] While most services also have free tiers, they limit daily usage to a small number of tokens.

[4] E.g. see OpenAI models' cost: https://openai.com/pricing.

[5] https://platform.openai.com/tokenizer

RQ1 (number of tokens): do all languages convey the same information with the same number of tokens? We analyze the fragmentation of sequences in different languages with different tokenizers. We find that among the supported languages in popular LMs, there is a large variance in the average number of tokens required to convey the same information with some languages requiring 5 times as many tokens than others. Previous work has shown that tokenization in multilingual models is usually biased towards high-resourced languages in the pretraining data (Ács, 2019; Rust et al., 2021); we observe that this is not always the case, but it could also be dependent on linguistic features or properties of language scripts.

RQ2 (cost): do non-uniform tokenization rates lead to LM API cost disparity for speakers of different languages? LM APIs like ChatGPT are available worldwide and have been widely claimed to have multilingual capabilities (Kasai et al., 2023; Lai et al., 2023).[6] We show that disparate fragmentation rates across languages lead to significantly high usage costs for less represented languages, and we argue for a more equitable API pricing system.

RQ3 (model utility): do non-uniform tokenization rates affect the models' utility? LMs have exhibited in-context learning capabilities, performing new tasks with few demonstrations as input (without parameter finetuning). This is highly desirable in any LM API as it avoids computational, annotation (and financial) costs. We show that high fragmentation rate of a language negatively affects the in-context learning performance in that language, resulting in reduced model utility.

RQ4 (socio-economic aspects): what are the socio-economic implications of the API's cross-lingual cost and performance disparity? Our analysis shows evidence that not only are LMs more expensive for certain languages, they are also less effective for them. To highlight the implications of these findings, we correlate those measurements with the socio-economic indicators of language speakers as a proxy for affordability of the APIs. This analysis indicates that *users who likely cannot afford high API costs are charged more for poorer service*, hindering uniform accessibility.

3 Experimental Setup

3.1 Models

Throughout this work, we focus on two LMs: ChatGPT (Ouyang et al., 2022; Brown et al., 2020) (gpt-3.5-turbo) and BLOOMZ (Muennighoff et al., 2022). Both of these models are trained and advertised as general-purpose models capable of following instructions and performing a wide range of tasks (Qin et al., 2023; Zhu et al., 2023; Ahuja et al., 2023; Huang et al., 2023).

ChatGPT (Ouyang et al., 2022) is a closed model only accessible through an API (with a premium tier) provided by OpenAI. Studies report that it supports as many as 90 languages (Ahuja et al., 2023). ChatGPT can handle a maximum sequence length of 4096 tokens (including both the prompt and generated tokens).

BLOOMZ (Muennighoff et al., 2022) is an open-source multilingual model trained on 46 natural languages and 13 programming languages. While training its tokenizer, sentences from different languages were sampled according to a multinomial distribution (Conneau et al., 2020), thereby increasing the number of tokens associated with low-resource languages. The best-performing version of this model has 175B parameters and is not feasible to be loaded on our academic servers; hence we rely on a free API of BLOOMZ hosted by Huggingface.[7] Although BLOOMZ was trained with ALiBi positional embeddings (Press et al., 2022) which allows the model to extrapolate to any length sequences during inference, the Huggingface API has a context limit of 1000 tokens.

3.2 Tasks and Datasets

To answer RQ1—whether the same information is conveyed with similar numbers of tokens in different languages—we use a validation set of FLORES-200 (Goyal et al., 2022), a multilingual parallel corpus containing examples in over 200 languages.[8] We tokenize each sentence in the FLORES-200 subset with ChatGPT's tokenizer[9] and compute the average number of tokens per sentence for each language. Using parallel data controls for the same information across languages. We consider that

[6] https://help.openai.com/en/articles/6742369-how-do-i-use-the-openai-api-in-different-languages

[7] https://huggingface.co/docs/api-inference/quicktour.

[8] We also experimented with WMT 2021 data (Akhbardeh et al., 2021) and found similar results. Note that the WMT data are focused on European languages.

[9] ChatGPT's tokenizer https://github.com/openai/tiktoken

language A is more efficiently tokenized than language B if it uses fewer tokens per sentence on average. While previous studies have computed fragmentation rates with fertility (Ács, 2019), we instead define it as the average number of tokens in a sequence for two reasons. First, our goal is to compare LLM API costs across languages that charge users based on the number of tokens. To control for content, we use a parallel corpus for this analysis. Second, many languages we analyze are understudied and do not have word tokenizers available which are required to compute fertility.

For RQ2 and RQ3, to clearly highlight the cost and utility disparities, we evaluate the models on NLP tasks that involve long-form texts either at input or output. We evaluate the models on diverse, challenging natural language generation and classification tasks on the following benchmarks:

Classification We evaluate on (1) XNLI (Conneau et al., 2018): a cross-lingual inference benchmark comprising of 11 typologically diverse languages. It involves two sub-tasks, passage selection and minimum answer span (Gold-P). We focus on the latter task in our experiments. (2) XFACT (Gupta and Srikumar, 2021): a multilingual fact verification dataset of naturally existing real-world claims covering 25 languages.

Span Prediction We use XQUAD (Artetxe et al., 2019): a crosslingual question-answering dataset where each example consists of a paragraph, a question, and the answer as a span in the paragraph.

Generation We evaluate on (1) Cross Sum (Hasan et al., 2021a): a cross-lingual abstractive summarization dataset comprising 1.7 million article-summary samples in 1500+ language pairs, and, (2) XLSUM (Hasan et al., 2021b): a summarization dataset covering 44 diverse languages. It comprises news articles and summaries in the same language as the article.

3.3 Prompting Formulation

We evaluate both models in a k-shot in-context learning setup where we also provide task instructions. We experiment with $0 \leq k \leq X$, where X is the maximum number of in-context examples that can be provided. Note that X is not a fixed value, but is determined by the LM API's limit on the number of input tokens and the fragmentation rate of the language.

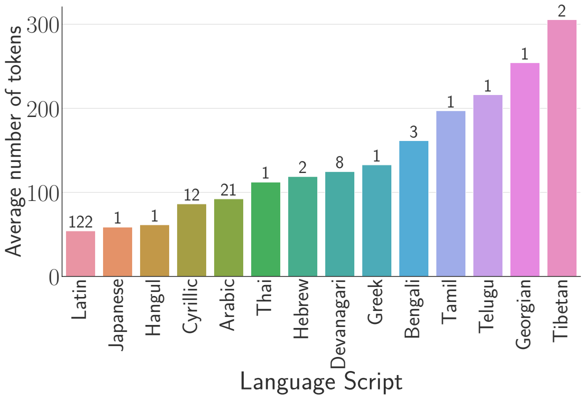


Figure 2: Average number of tokens by script after tokenizing the Flores dataset. The fragmentation rate is lower for Latin script languages and higher for other scripts. Number of languages per language group is indicated at the top of each bar.

For all tasks, we provide the instructions in English following Ahuja et al. (2023), who show that on several multilingual benchmarks, English instructions outperform the in-language prompts (see Table 2 in the Appendix for the prompting format for all tasks). For each task, we randomly sample at most 500 test examples for evaluation.

4 Results and Analysis

4.1 RQ1 (number of tokens): do all languages convey the same information with the same number of tokens?

In Figure 2 we show that Latin-script languages are represented with substantially fewer tokens compared to languages in other scripts. While Cyrillic and Japanese script languages come close to the Latin, languages with their own script (e.g., Telugu) require up to 5× more tokens to convey the same information. We hypothesize that this disparity is due to training data imbalance since ChatGPT's tokenizer was primarily trained on Latin-script languages, mainly English. The training details of ChatGPT are not available. However, we make a reasonable assumption that its training data has a similar proportion of languages as the publicly available large corpus CC100 (Wenzek et al., 2020). If we sort languages shown in Figure 2 based on their data size in CC100 (see Figure 14 in the Appendix), low-resourced languages of Latin script appear to be less fragmented compared to other mid-resourced languages of non-Latin scripts.

In Figure 15 in the Appendix, we present a similar analysis for BLOOMZ's tokenizer. We sort the languages based on their size in the pretraining data (ROOTS corpus; Laurençon et al., 2023). We observe that languages with fewer resources generally have a higher average token length. Arabic is an outlier here as it appears to have more tokens than some other mid-resourced languages.

What influences the non-uniformity of a tokenizer across languages? From our analysis above, we identify two influential factors: (1) the proportion of the language in the pretraining data, and (2) inherent properties of the language and its writing script. While we see some correlation between pretraining data size and fragmentation rate in BLOOMZ , with ChatGPT it is quite different as higher-resourced non-Latin script languages still get excessively tokenized.

To disentangle the effects of factors (1) and (2) we train BBPE tokenizers on a variety of languages with diverse scripts with vocabulary sizes ranging from 5,000 to 50,000, while controlling for content and data size. Specifically, we train the tokenizers on parallel corpora and include one language per script. We then use these tokenizers to tokenize the text they were trained on, and compute the average number of tokens per sentence.

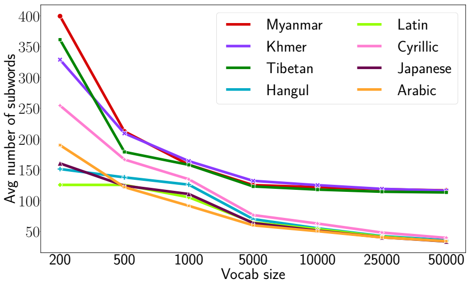


Figure 3: BBPE tokenizer trained on parallel text from different language scripts with varying vocabulary sizes. We display a larger version with 21 more scripts in Figure 19 in the Appendix.

As shown in Figure 3, even when controlling for the content, there is still a disparity in the tokenization rate at different vocabulary sizes. In particular, most scripts are very sensitive to small vocabulary sizes compared to Latin and Hangul scripts. We do not achieve uniform fragmentation rate across all language scripts even with large vocabulary sizes. We therefore conclude that uniformity of BBPE tokenizers across languages is not just determined by the proportion of text from language in the pretraining data but also by language/script properties.

4.2 RQ2 (cost): how do non-uniform tokenization rates affect LM API costs for different languages?

LM APIs charge users a fixed amount for a given number of input and generated tokens. Since the same information is expressed using different number of tokens in different languages, we aim to investigate the disparity in what users pay to use the API for different languages. From the results of our analysis in §4.1, we compute the estimated cost of API use per language as a function of the average sequence length derived in Figure 2. We report this on a subset of languages in Figure 16 in the Appendix and present a granular analysis of languages that share family and script in Figure 4.

Languages that are more heavily segmented have predictably higher costs of usage. Overall, we see that the API costs are biased towards (i.e., cheaper for) Indo-European and Latin script languages and against many non-Latin script languages. In most mid-resourced Indic languages with non-Latin scripts, we see close to a 5× increase in cost compared to English.

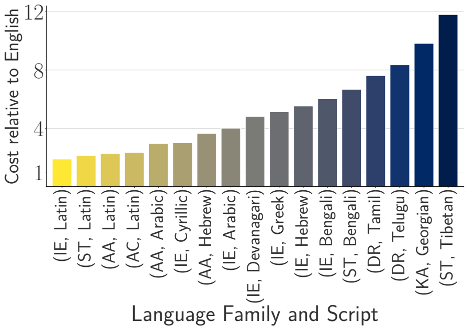


Figure 4: Estimated cost per language family/script, relative to English. The language families are abbreviated as follows: IE: Indo-European, ST: Sino-Tibetan, AC: Atlantic-Congo, AA: Afro-Asiatic, DR: Dravidian, KA: Kartvelian.

Next, we report the costs of running experiments relative to English. We report costs based on our zero-shot experiments across all tasks listed in §3.2. This is due to excessive tokenization in some languages for which we can only do zero-shot evaluations. For XLSUM, we show in Figure 5 that we spend up to 4× more for both prompting and generation in Telugu and Amharic. We observe similar findings in XFACT and CROSSUM, as displayed in Figure 11 in the Appendix.

While the majority of the commercial LMs are perhaps being optimized to perform well in many

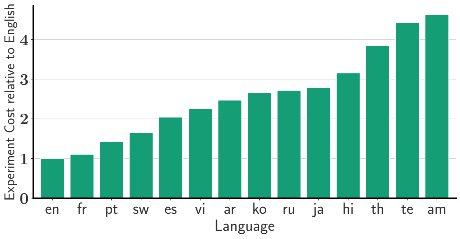


Figure 5: Average cost of prompt + generated tokens for XLSUM evaluations relative to English.

languages, we show that there is less focus on individual experiences of speakers of languages other than English. While LMs like ChatGPT might perform tasks in Telugu, for example, a user in Andhra Pradesh might pay 5× more than an English user in the US for an equivalent use of the model.

4.3 RQ3 – Model utility: do non-uniform tokenization rates affect the models' utility?

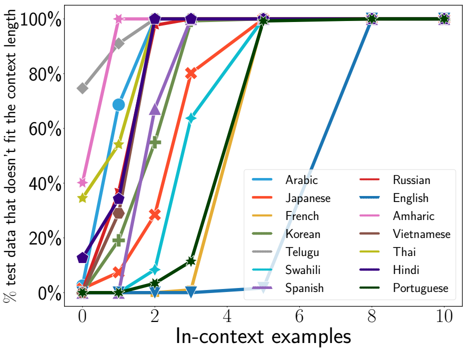


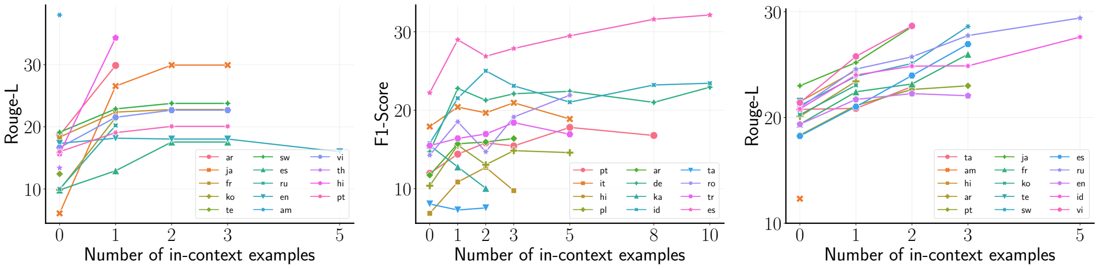
Figure 6: Percentage of test examples per language in XLSUM that do not successfully fit into the context length of ChatGPT. We can fit more few-shot examples in Latin script languages than in other languages.

LMs typically have an upper bound of the number of tokens they can handle, e.g., ChatGPT can process a maximum of 4,096 tokens. Hence, due to non-uniform fragmentation rates across languages, there is a disparity in the amount of information the models can process per language. In Figure 6 we plot the percentage of XLSUM test instances against the maximum number of in-context examples those instances can be accompanied with. For example, Telugu struggles to fit even one in-context example for the majority of the test set. Hence, the model can only do zero-shot prompting in this case.

To measure the impact of this issue on task performance, we evaluate ChatGPT and BLOOMZ with a k-shot learning setup on the 5 tasks on diverse languages as described in §3.2. Figure 7 shows ChatGPT's performance according to standard automatic metrics of all tasks. Note that the focus of this experiment is to illustrate the impact of tokenization in in-context learning settings. Therefore, we are interested not in the absolute value of the metrics or comparisons among languages but the relative improvement within the test sets of the same language as we increase the number of in-context examples. For all tasks and most languages, we see consistent performance improvements as we increase the number of in-context examples, from zero-shot to k (even for $k = 1$). For many languages such as Telugu and Thai, due to their high fragmentation rates, we were unable to fit even one complete demonstration and hence, only report zero-shot results. Based on trends from other languages, we suspect that these languages could also have benefitted from more demonstrations. Hence, as a result of unfair tokenization, ChatGPT's utility is much lower for speakers of those languages compared to better represented languages like English.

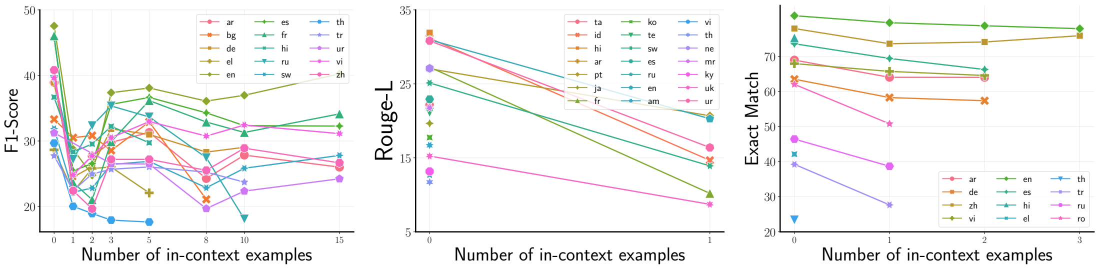
Figure 8 reports the results of the same experiment for BLOOMZ. Across all tasks we find that adding in-context examples does not help. In fact, in some cases, there is a performance drop even with one in-context example. Upon manual inspection of the generated outputs from the one-shot experiments, the model has a tendency to copy spans from the in-context example, presenting that as output and thus not successfully utilize demonstrations. Our hypothesis here is that BLOOMZ is better optimized for zero-shot prompting and is not as suitable for in-context learning.

Due to the limited number of tokens that BLOOMZ's inference API accepts, some examples in some languages cannot fit the 1000 token context length when doing zero-shot prompting. We experienced this with the XLSUM dataset as we couldn't fully fit news articles for some languages. Understandably, some of these languages are not even present in its pretraining data, and hence we do not expect them to be tokenized efficiently. For these examples that do not fit the context length, we feed in truncated news articles into the model. We therefore evaluate the generations for the fraction of examples that fit context and ones that do not fit the context separately. Figure 9 shows the performance comparison when we use truncated summaries in the prompt and when we use the full articles. While the performance drop is expected,



(a) XLSUM (b) XFACT (c) CROSS-SUM

Figure 7: Results from ChatGPT few-shot evaluations. In most tasks, we see an increase in performance as we increase the number of in-context examples.



(a) XNLI (b) XLSUM (c) XQUAD

Figure 8: Results from BLOOMz few-shot evaluations. The BLOOMz model is clearly better at zero-shot prompting than few-shot.

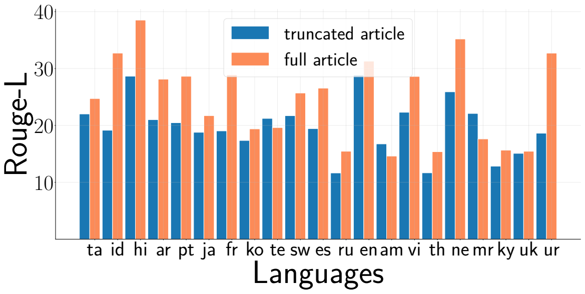


Figure 9: Zero-shot evaluation of BLOOMz on XL-SUM. Since we cannot fit the full article in the context length for some languages, we compare results on evaluating full articles vs. truncated articles.

our focus here is to highlight a consequence of differentiated tokenization in LMs.

4.4 RQ4 – Socio-economic aspects: what are the socio-economic implications of the API's cross-lingual cost and performance disparity?

In Figure 10, we plot the fragmentation rate per language against the Human Development Index in the country with the highest absolute number of speakers of that language. We find a strong negative correlation close to -0.5 showing that in most cases, the lower the HDI index, the higher the fragmentation rate and vice versa. Evidently, the model's vocabulary is biased towards users of more developed countries.

Task	Cost-HDI		HDI-Utility		Cost-Utility	
	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson
XFACT	**–0.41	**–0.60	*0.34	**0.38	**–0.61	**–0.55
XLSUM	**–0.42	**–0.43	**–0.44	**–0.57	*–0.23	*0.21
CROSS SUM	**–0.41	**–0.45	*–0.18	*0.24	*0.27	*–0.17

Table 1: Correlation between model utility, cost of API access and Human Development Index (HDI) for each task. We mark correlations with $p < 0.05$ with * and also mark correlations with $p < 0.0056$ (according to Bonferroni correction for multiple hypotheses) with **.

This bias is further validated by results shown in Table 1, where we mostly find negative correlations between pairs of each of the following variables: average financial cost of experiments, model utility (performance), and human development index of the country in which each language is spoken. We term this "double unfairness" as people from less economically developed countries are overcharged at a fixed rate per-token due to excessive tokenization, but often derive less utility from the model.

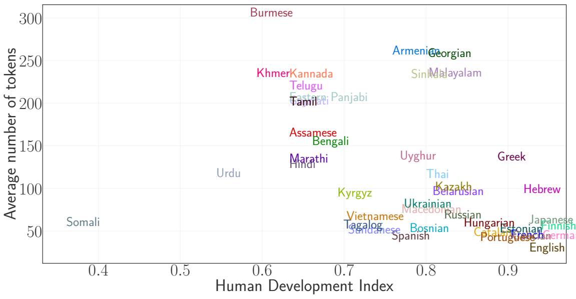


Figure 10: Fragmentation rate per language against the Human Development Index in the top country where the language is spoken.

5 What is the Way Forward?

Transparency in API limitations While the NLP community is aware of many of the issues we point out in this work, LM APIs are advertised to a general audience. Much like the policy of adding limitations to research papers, LM API providers ought to be more transparent about the flaws and biases in their models, especially when describing their multilingual capabilities. Many users are not privy to the inner workings of these models and will be unknowingly charged higher prices if they use the model in their native languages.

Rethinking the API pricing models Higher API costs for languages in underprivileged communities risks excluding many populations from using language technologies. A potential solution is to develop pricing policies based on languages/regions while also accounting for model performance on language-specific benchmarks. An alternative is to not charge by tokens at all. PaLM 2 API, for example, charges the users based on characters.[10] Hence, further analysis is needed to assess the fairness of character-based pricing. Huggingface also offers an inference service for their enterprise customers[11] relying on AWS instances and charging them at an hourly rate. Future work may compare this with a per-token rate we study in this work.

Open-source models vs. paid APIs Given issues with paid APIs, the natural next question might be: should the API users move to open-source models or train their own? In fact, in our experiments, we find BLOOMZ, an open-source model, to perform better in the zero-shot setting than ChatGPT performs in the few-shot setting, in most cases. However, first, most open-source models are distributed under an academic license whereas most developers are interested in integrating these technologies into their products for commercial use, which may incur licensing costs. Second, barring licensing issues, LMs tend to be large and resource-intensive to train and deploy and require dedicated expensive hardware to run at a commercial scale, which again might not be possible for most developers and users, even exceeding the cost of using the APIs. Research on reducing such hardware requirements (Dettmers et al., 2022; Park et al., 2023) could increase accessibility. Still, this requires a considerable level of technical expertise from developers and users which might be infeasible.

Technological improvements in LMs Several solutions proposed in recent work to improve language modeling performance can help alleviate the cost and utility issues we highlight. Tokenization is an active area of research and various solutions based on data balancing (Johnson et al., 2017; Conneau and Lample, 2019), optimal transport (Xu et al., 2021), fuzzy subwords (Provilkov et al., 2020), and many more (Chung et al., 2020; Tay et al., 2022) have been proposed. BLOOMZ, for instance, relies on data balancing to improve fragmentation rates across languages. Some works also focused on increasing the context lengths of language models (Bulatov et al., 2023; Press et al., 2022) which can help alleviate issues with utility by allowing more in-context examples as input.

6 Related Work

Analyzing tokenization methods The impact of tokenization on model performance (Ács, 2019; Rust et al., 2021; Zhang et al., 2022a; Klein and Tsarfaty, 2020; Bostrom and Durrett, 2020; Kamali et al., 2022), inference speed and memory usage of LMs in practical settings (Sun et al., 2023; Hofmann et al., 2022) has been widely studied. Ács (2019) observes that mBERT's vocabulary is largely dominated by Indo-European languages. Rust et al. (2021) find that monolingual LMs perform better than mBERT because some languages suffer from over-fragmentation. Zhang et al. (2022a) find that sentence-level MT models are not sensitive to language imbalance in their tokenizer training data. In contrast to prior work, our focus is on the cost and performance analysis of multilingual LM APIs across languages with regard

[10] Prior work has shown evidence that even the number of characters used to express the same information in different languages is

[11] https://huggingface.co/pricing#endpoints

to over-fragmentation and in-context learning.

Socio-economic impacts of language models Prior work show that unfairness in LMs is a consequence of many stages in the development pipeline (Cao and Daumé III, 2020; Talat et al., 2021). Efforts have tried to identify social biases in LM generations (Wolfe and Caliskan, 2021; Dev et al., 2022; Sheng et al., 2021; Chen et al., 2021; Hutchinson et al., 2020). Other works have surfaced the cultural and language disparity beyond and within multilingual LMs (Gururangan et al., 2022; Kreutzer et al., 2022; Virtanen et al., 2019). Talat et al. (2022) discuss challenges impacting bias evaluation in multilingual LMs. They examine power dynamics and consequences of training LMs emphasizing implications associated with advancement of such technologies. In this work, we study economic unfairness of LMs across different communities. Concurrent work (Petrov et al., 2023) analyses multilingual tokenizers focusing on financial cost, latency and context size. However, apart from cost, our analysis also covers model utility and socio-economic implications. Kasai et al. (2023) report unfair API costs as a result of tokenization differences between English and Japanese. We extend this to 21 more languages highlighting the pervasiveness of this issue.

7 Conclusion

By analyzing popular language model APIs on challenging multilingual benchmarks, we find that (a) API tokenizers disproportionately favor Latin scripted languages and over-fragment less represented languages and scripts, (b) the API pricing policy of charging based on the number of tokens is flawed and extremely unfair towards speakers of the over-fragmented languages, and (c) the API performs poorly on such languages compared to the less-fragmented counterparts. In the current NLP research landscape, where more and more industrial labs are building their own APIs, this is a concerning trend that may reduce the accessibility of these technologies to already marginalized communities. Hence, we encourage the vendors to be more transparent about their models' limitations and rethink their pricing policy.

Ethics Statement

This work sheds light on the consequences of unfair tokenization to users of commercial LM APIs that speak languages with scripts less represented in the pretraining data. With the recent widespread use of commercial LMs, we believe that our work is crucial to ensuring that language technologies are accessible to diverse users irrespective of the languages they speak.

There are different factors that contribute to non-uniform tokenization across languages. Whilst our analysis touches on the size of pretraining data and language writing systems we suspect that there might be other factors not yet uncovered; we leave that for future work. The lack of access to OpenAI's training data prevents us from making solid claims about all the languages that ChatGPT is optimized for; however, their models have been advertised and shown to work well in many languages. More work on large multilingual models should include the release of (details of) training data to further enable this kind of research.

Limitations

Translationese We conduct the analysis to answer RQ1 using a parallel corpus, FLORES-200 (Team et al., 2022), in order to control for the same information. This corpus consists of many examples that have been professionally translated. Prior studies have shown that translated texts in any language (referred to as translationese) may differ from original written text in many ways (Laviosa, 2002). These may have caused the information conveyed in different languages to not be exactly the same. We do not have a way to measure these differences. However, we expect them not to be so large as to meaningfully affect the trend of fragmentation rates.

Language statistics of ChatGPT training data ChatGPT is a closed model developed by OpenAI who have not released the training details of the model including any information of the languages it supports.[12] Hence, we cannot ascertain the actual statistics of all the languages in their training data. We use CC100 (Wenzek et al., 2020), a large multilingual corpus, to estimate these statistics.

Reproducibility One limitation with testing closed LMs is lack of reproducubility particularly

[12] The only official information they provide about ChatGPT's multilingual support is here: `https://help.openai.com/en/articles/6742369-how-do-i-use-the-openai-api-in-different-languages` Prior studies have speculated that ChatGPT was trained on at least 90 languages (Ahuja et al., 2023).

because the model weights are typically updated continually. However, this only affects the downstream evaluations as our cost analysis is reproducible, since the tokenizers we evaluate are open-source.

Acknowledgements

We thank the members of the Tsvetshop and Noah's ARK labs at the University of Washington for the valuable discussions and useful feedback. We thank Lucille Njoo for help with our illustrations. We also thank the reviewers and area chair for their valuable feedback. During the course of this study, S.K. was supported by Google Ph.D. Fellowship. We also gratefully acknowledge support from NSF CAREER Grant No. IIS2142739, the Alfred P. Sloan Foundation Fellowship, and NSF grants No. IIS2125201, IIS2203097, and IIS2113530. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif M Mohammad, and Karën Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of ACL*.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. 2023. Scaling transformer to 1M tokens and beyond with RMT.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias and under-representation in natural language processing across human languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 24–34, New York, NY, USA. Association for Computing Machinery.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Annual Meeting of the Association for Computational Linguistics*.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021a. CrossSum: Beyond English-centric cross-lingual abstractive text summarization for 1500+ language pairs. *ArXiv*, abs/2112.08804.

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

Charles F. Hockett. 1997. The world's writing systems. *Language*, 73(2):379–385. Accessed 17 Oct. 2023.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Danial Kamali, Behrooz Janfada, Mohammad Ebrahim Shenasa, and Behrouz Minaei-Bidgoli. 2022. Evaluating Persian tokenizers.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir R. Radev. 2023. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

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

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The BigScience ROOTS Corpus: A 1.6tb composite multilingual dataset.

Sara Laviosa. 2002. Corpus-based translation studies: Theory, findings, applications.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Rose Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir R. Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *ArXiv*, abs/2211.01786.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Gunho Park, Baeseong Park, Minsub Kim, Sungjae Lee, Jeonghoon Kim, Beomseok Kwon, Se Jung Kwon, Byeongwook Kim, Youngjoo Lee, and Dongsoo Lee. 2023. LUT-GEMM: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver?

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

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose 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 Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. 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, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali 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 L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. 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, 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 V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y 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 Franccois Lavall'ee, 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, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, 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, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim T Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully A. Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, R. Eisenberg, Robert Martin, Rodrigo L. 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 Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yun chao Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

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

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinying Song, Alexandru Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast WordPiece tokenization. In *Conference on Empirical Methods in Natural Language Processing*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *Proc. of ICLR*.

Jimin Sun, Patrick Fernandes, Xinyi Wang, and Graham Neubig. 2023. A multi-dimensional evaluation of tokenizer-free multilingual pretrained models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1725–1735, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *Proceedings of ICML*.

Zeerak Talat, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *ArXiv*, abs/2101.11974.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*.

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.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022a. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT reproduce human-generated labels? a study of social computing tasks.

Judit Ács. 2019. Exploring BERT's vocabulary. `http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html`.

A Prompt template

In Table 2, we provide the exact prompts we use for each respective task in our experiments.

B Cost analysis

In Figure 11 we present the experimental cost relative to English and Spanish for Crosssum and XFACT respectively. Figure 16 shows the estimated cost of GPT3.5 API access for all languages in CC100 relative to English .

C Extra analysis of fragmentation rate

In Figure 12 and Figure 13 we present the fragmentation rate across language families and scripts for both GPT3.5 and BLOOM respectively. Figure 17 shows the fragmentation rate for all languages in FLORES grouped by language script.

D Fragmentation rate vs pretraining data size

In Figure 14 we sort languages based on their size in CC100 corpus (Wenzek et al., 2020) and plot their fragmentation rate with GPT3.5 tokenizer. Figure 15 shows the same statistics for BLOOM's tokenizer based on language pretraining data size in (ROOTS corpus; Laurençon et al., 2023).

E Pretrained tokenizers on more languages

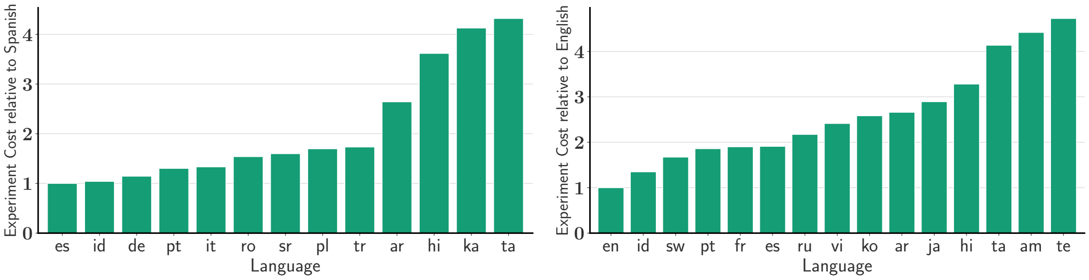
Figure 19 shows fragmentation rate across language scripts, when we train a BBPE tokenizer trained on parallel text in 30 languages.

F Fragmentation rate vs HDI

Figure 18 shows GPT3.5's fragmentation rate per language against the Human Development Index of the country with the largest amount of speakers of that language. We add more languages and countries here compared to the figure in the main paper.

Task	Prompt Template
XLSUM	Write a short summary sentence of the following text in {language} Article: { article} Summary:
XQUAD	Context: context Question: question Answer: Template
XNLI	{Premise} Question : {hypothesis} True, False, or Neither? Answer:
CROSSUM	Write a short summary sentence of the following text in English. Article: { article} Summary:
XFACT	Tell me whether the following claim is {label 1 } or {label 2 } or {label 3 } ... given evidence {evidence 1 }, {evidence 2 }, {evidence 3 }

Table 2: Prompt template used for each dataset.



(a) XFACT

(b) CROSSUM

Figure 11: Relative cost of prompt + generated tokens for XFACT and CROSS-SUM evaluations.

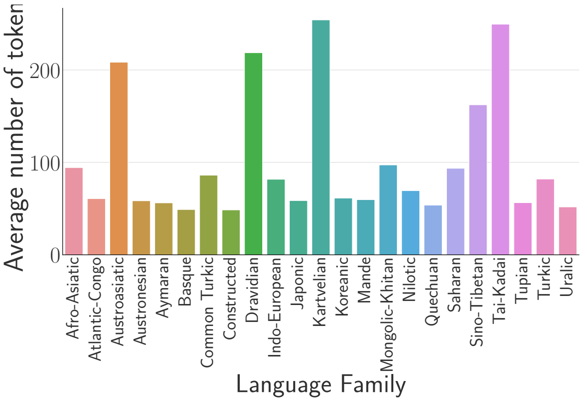


Figure 12: Average number of tokens per language family after tokenizing Flores dataset with GPT3.5 tokenizer. The fragmentation rate is lower for Latin script languages and higher for other scripts.

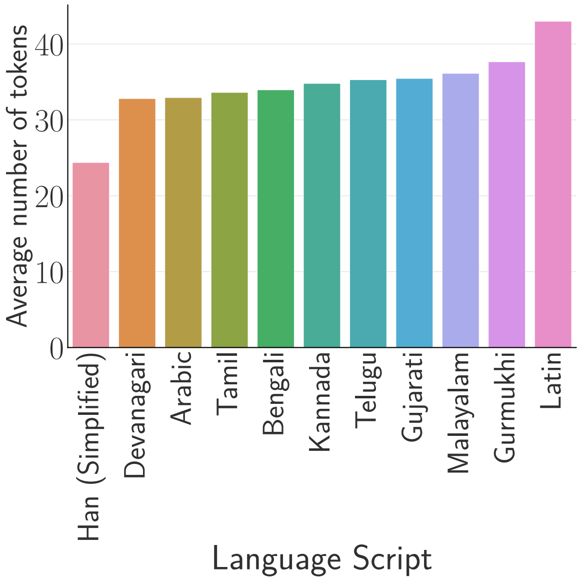


Figure 13: Average number of tokens per language script after tokenizing Flores dataset with BLOOM tokenizer. The fragmentation rate is higher on average for Latin script languages. This is because majority of the low-resourced languages are latin-script and have higher fragmentation rate.

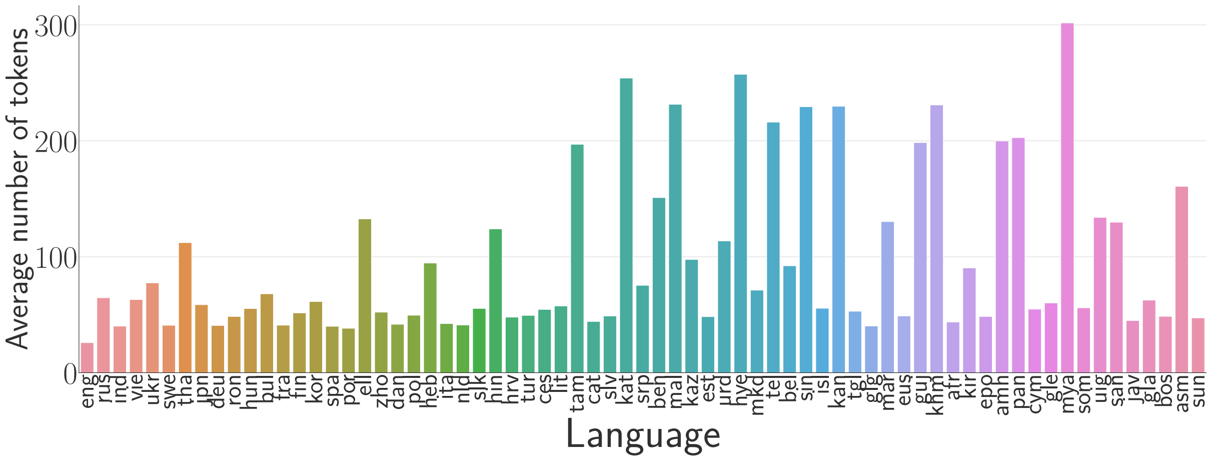


Figure 14: Average number of tokens per language after tokenizing FLORES with GPT3.5 tokenizer. Languages are arranged in descending order based on the size of pretraining data in Commoncrawl.

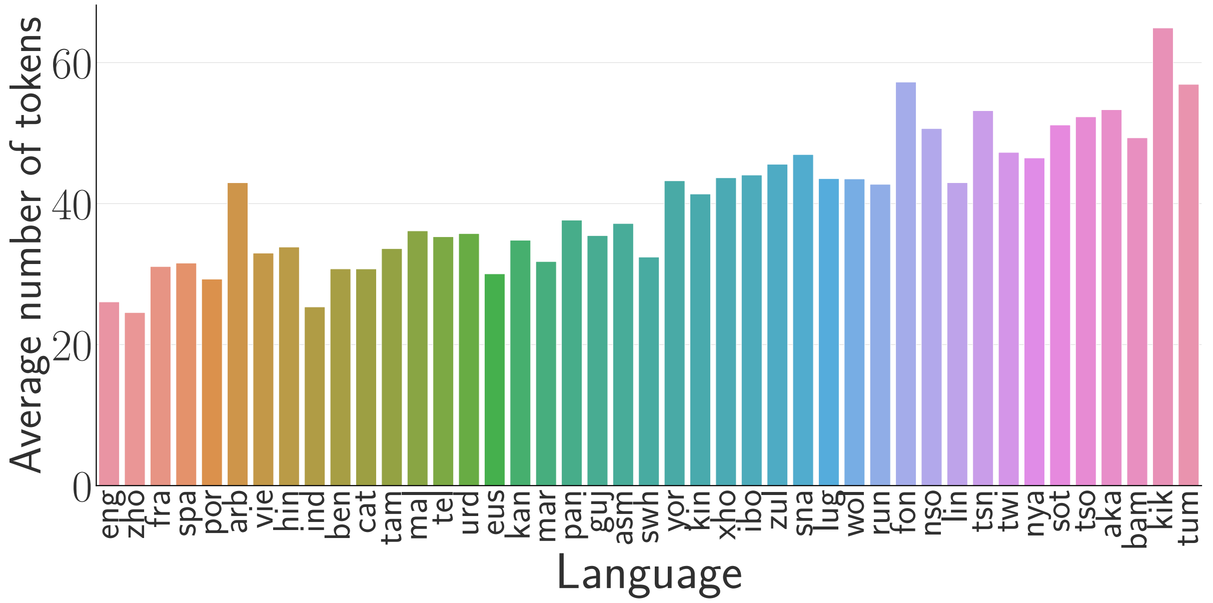


Figure 15: Average number of tokens per language after tokenizing FLORES with BLOOM tokenizer. Languages are arranged in descending order based on the size of pretraining data in the ROOTS corpus.

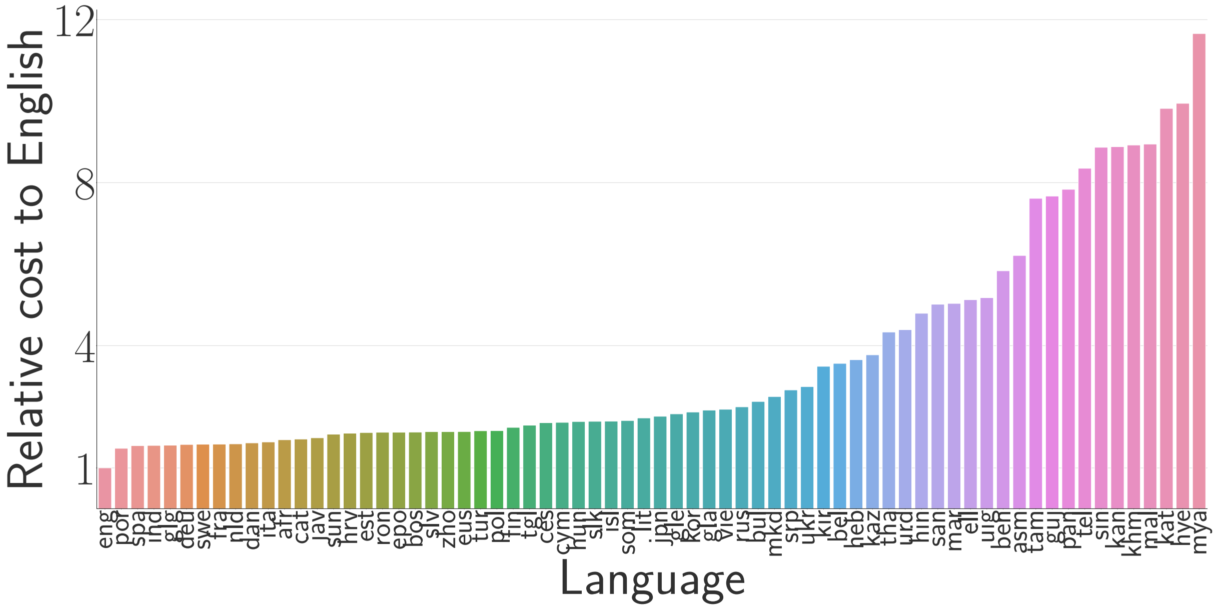


Figure 16: Estimated cost of GPT3.5 API access relative to English.

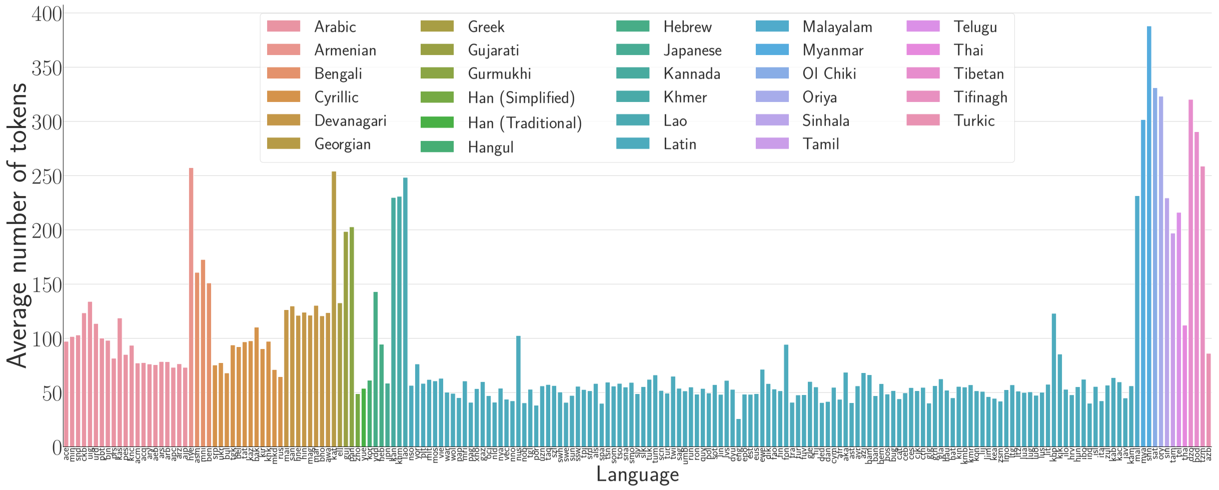


Figure 17: Average number of tokens by script after tokenizing all languages in the Flores dataset with GPT3.5 tokenizer.

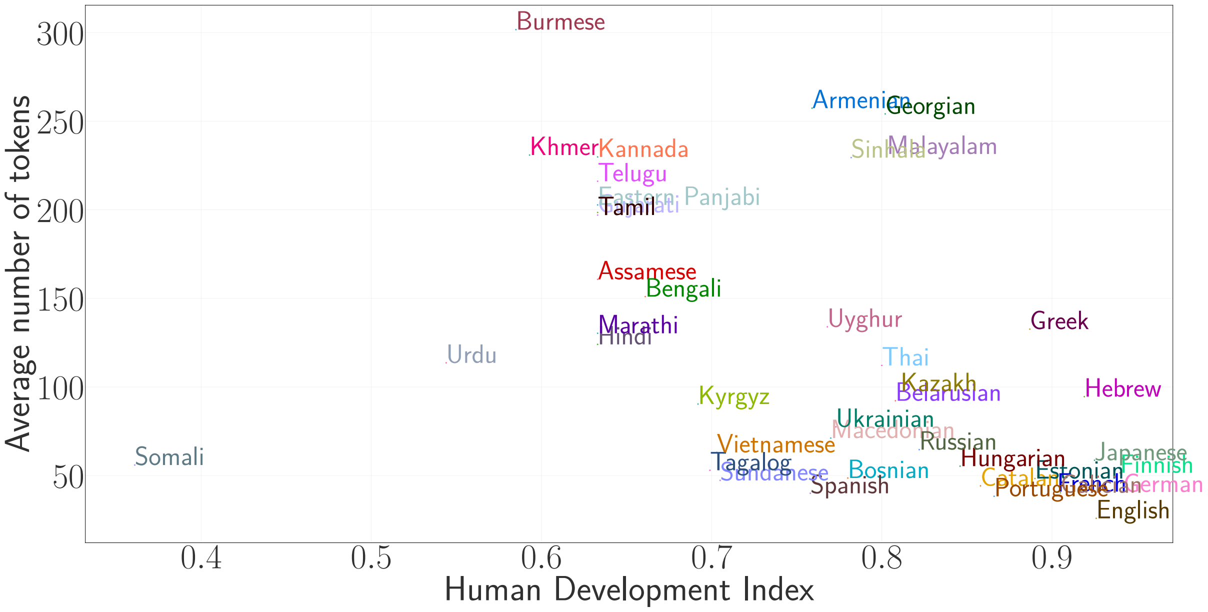


Figure 18: Fragmentation rate per language against the Human Development Index in the country with the largest amount of speakers of that language.

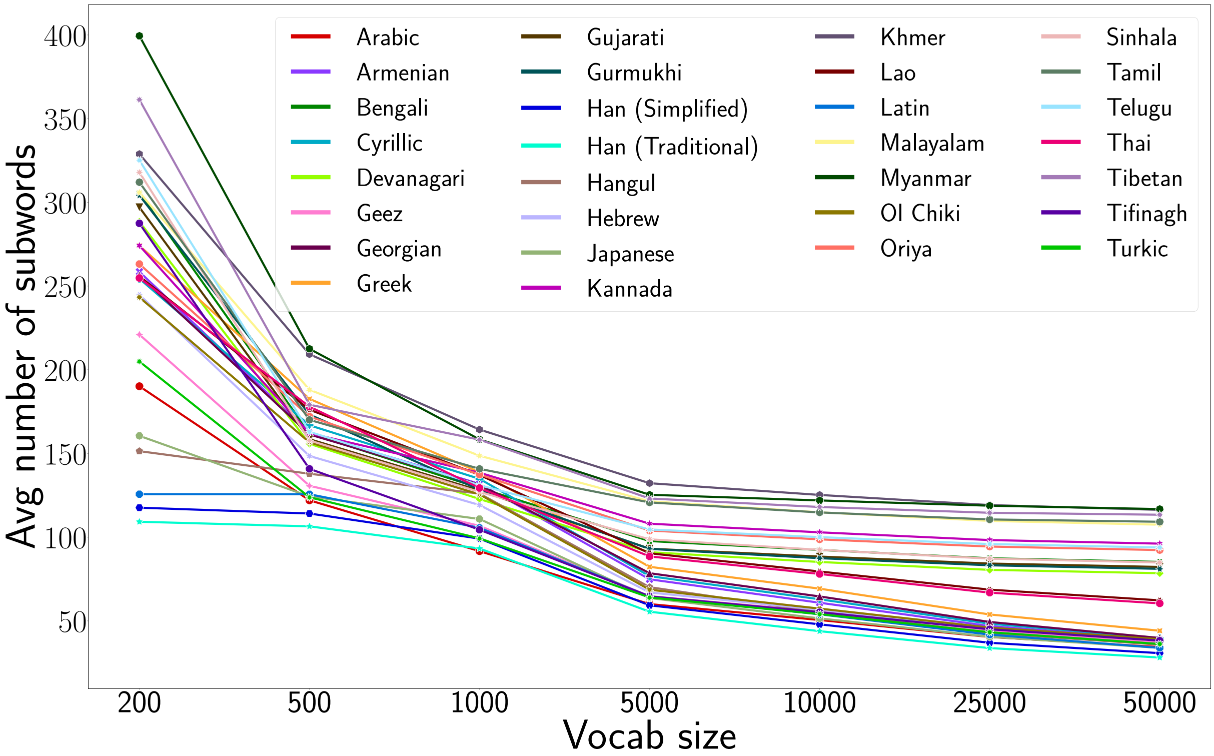


Figure 19: BBPE tokenizer trained on parallel text from 30 language scripts with varying vocabulary sizes. It is impossible to achieve uniform fragmnatation rate even when we have equal pretraining data sizes across all language scripts.