CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages

Thuat Nguyen[1], Chien Van Nguyen[1], Viet Dac Lai[1], Hieu Man[1],
Nghia Trung Ngo[1], Franck Dernoncourt[2], Ryan A. Rossi[2], Thien Huu Nguyen[1]
[1] Department of Computer Science, University of Oregon, USA
[2] Adobe Research, USA
nguyenhuuthuat09@gmail.com, {chienn,vietl@cs,hieum,nghian,thien@cs}@uoregon.edu
{franck.dernoncourt,ryrossi}@adobe.com

Abstract

Extensive training datasets represent one of the important factors for the impressive learning capabilities of large language models (LLMs). However, these training datasets for current LLMs, especially the recent state-of-the-art models, are often not fully disclosed. Creating training data for high-performing LLMs involves extensive cleaning and deduplication to ensure the necessary level of quality. The lack of transparency for training data has thus hampered research on attributing and addressing hallucination and bias issues in LLMs, hindering replication efforts and further advancements in the community. These challenges become even more pronounced in multilingual learning scenarios, where the available multilingual text datasets are often inadequately collected and cleaned. Consequently, there is a lack of open-source and readily usable dataset to effectively train LLMs in multiple languages. To overcome this issue, we present CulturaX, a substantial multilingual dataset with 6.3 trillion tokens in 167 languages, tailored for LLM development. Our dataset undergoes meticulous cleaning and deduplication through a rigorous pipeline of multiple stages to accomplish the best quality for model training, including language identification, URL-based filtering, metric-based cleaning, document refinement, and data deduplication. CulturaX is released in Hugging Face facilitate research and advancements in multilingual LLMs: https://huggingface.co/datasets/uonlp/CulturaX.

Keywords: Large Language Models, Open Training Data, Multilingual Learning

1. Introduction

Large language models (LLMs) have fundamentally transformed research and applications of natural language processing (NLP), significantly advancing the state-of-the-art performance for numerous tasks and revealing new emergent abilities (Brown et al., 2020; Wei et al., 2022). Based on the transformer architecture (Vaswani et al., 2017), three major variants of LLMs have been explored in the literature: the encoder-only models to encode input texts into representation vectors, e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019); the decoder-only models to generate texts, e.g., GPT (Radford et al., 2019; Brown et al., 2020); and the encoder-decoder models to perform sequence-to-sequence generation, e.g., BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). The remarkable capabilities of LLMs have primarily been propelled by the ever-expanding scale of model sizes and training datasets, which have been deemed essential for achieving optimal performance by the scaling laws (Hernandez et al., 2022). For instance, beginning with the BERT model, which had a mere few hundred million parameters (Devlin et al., 2019), recent GPT-based models have been expanded to encompass hundreds of billions of parameters (Shoeybi et al., 2019; Scao et al., 2022; Lieber et al., 2021; Chowdhery et al., 2022). Similarly, the training datasets for LLMs have grown exponentially, evolving from a modest 13GB of text data from Wikipedia and books used for BERT (Devlin et al., 2019; Liu et al., 2019) to consume terabytes of data for the latest models, such as Falcon (Penedo et al., 2023), MPT (MosaicML, 2023), LLaMa (Touvron et al., 2023), PolyLM (Wei et al., 2023) and ChatGPT[1].

As the field keeps progressing rapidly, pre-trained LLMs have typically been released to the public to foster further research and advancements. These models are obtainable either through commercial APIs, as illustrated by ChatGPT and GPT-4, or via open-source initiatives, exemplified by Falcon and LLaMa. Nevertheless, in contrast to the public accessibility of LLMs, the training datasets that underpin the state-of-the-art models have mostly remained closely guarded secrets, even in the case of open-source LLMs such as BLOOM, LLaMa, MPT, and Falcon. For example, Falcon (Penedo et al., 2023) and BLOOM (Scao et al., 2022) only provide a glimpse of their complete training data, whereas MPT's, LLaMa's and PolyLM's datasets (Touvron et al., 2023; Wei et al., 2023) remain inaccessible to the public. On one hand, the lack of transparency has impeded in-depth analysis and comprehension of LLMs, hindering crucial research into attributing and addressing fundamental issues stemming from the training data, such as hallucinations, biases, and toxic content (Tamkin et al., 2021; Weidinger

[1] https://openai.com/blog/chatgpt

LREC-COLING 2024, pages 4226–4237
20-25 May, 2024.

et al., 2021; Kenton et al., 2021; Bommasani et al., 2021). On the other hand, concealing the training data restricts the development of LLMs to a select few stakeholders with ample resources, thereby constraining the democratization and benefits of the technology and exacerbating its biases within broader society.

To attain transparency and democratization for LLMs, it is thus crucial to create large-scale and high-quality datasets for training high-performing LLMs while ensuring their public accessibility to foster deeper research and advancements. In the realm of LLMs, high-quality training datasets are often crafted through the application of extensive data cleaning and deduplication processes, aimed at eliminating noisy and redundant content from vast text collections (Allamanis, 2018; Penedo et al., 2023). To this end, there have been recent efforts from the community to develop such open-source datasets for LLMs, such as RedPajama with 1.21T tokens (Computer, 2023), SlimPajama[2] with 627B tokens, and AI2 Dolma[3] with 3T tokens. However, most of the existing open-source datasets for LLMs are tailored for the English language, which hinders the utilization and performance of the resulting LLMs when applied to non-English languages, particularly those with limited linguistic resources (Bang et al., 2023; Lai et al., 2023). This emphasis on English also restricts the capacity of open-source datasets to comprehensively tackle the research challenges and democratization concerns of LLMs across the diverse spectrum of over 7,000 languages spoken worldwide.

Simultaneously, some multilingual datasets have been developed and made available, providing text data for multiple languages. Nevertheless, their quality and scale fall short of meeting the requirements for training high-performing LLMs. Specifically, the multilingual text dataset sourced from Wikipedia, while of high quality, is regarded as relatively small when it comes to training LLMs (Conneau et al., 2020). The OSCAR datasets (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020; Abadji et al., 2021, 2022)[4] extract text data from CommonCrawl (CC) for more than 160 languages. However, these datasets lack document-level deduplication (i.e., removing similar documents in the dataset), leading to the inclusion of redundant information and impairing the performance of generative LLMs (Lee et al., 2022). Similarly, the mC4 (Xue et al., 2021), CCAligned (Conneau et al., 2020), WikiMatrix (Schwenk et al., 2021), and ParaCrawl (Bañón et al., 2020) datasets altogether support over 100 languages but suffers from less accurate language identification, introducing noise into the data (Kreutzer et al., 2022). These datasets are also not deduplicated at fuzzy and document levels, e.g., via MinHash (Broder, 1997). Additionally, the CC100 dataset (Wenzek et al., 2020; Conneau et al., 2020), employed in training the multilingual XLM-RoBERTa model across 100 languages, only considers the snapshots of CC in 2018, constraining its size and the availability of up-to-date information to train high-performing LLMs.

To address the aforementioned issues for open-source datasets, our work introduces a novel multilingual dataset, called CulturaX, for training LLMs in 167 languages. CulturaX merges the latest iteration of mC4 (version 3.1.0) with all available OSCAR corpora up to the current year, encompassing distributions 20.19, 21.09, 22.01, and 23.01. This amalgamation results in a large multilingual dataset, comprising 27 TB of text data with 6.3 trillion tokens and offering the most up-to-date data for LLM development. More than half of our dataset is dedicated to non-English languages to significantly boost the data size and enhance the feasibility of training models in multilingual scenarios. Importantly, CulturaX is extensively cleaned and deduplicated at the document level to produce the highest quality to train LLMs for multiple languages. In particular, our data cleaning process includes a comprehensive pipeline designed to eliminate low-quality data. This involves removing noisy text, non-linguistic content, toxic data, incorrect language identification, and more. Our data cleaning pipeline employs a variant of the Interquartile Range (IQR) method (Dekking et al., 2007) to select appropriate thresholds for various dataset metrics (e.g., stopword ratios, data perplexity, and language identification scores), which can be used to filter noisy outliers for the dataset. As such, we leverage the percentiles of the distributions computed over large samples of data to effectively guide the threshold selection process for each filtering metric and language. Finally, we perform extensive deduplication for the data of the languages within our datasets based on the near deduplication method MinHashLSH (Broder, 1997; Leskovec et al., 2020) and URLs, leading to high-quality data to train multilingual LLMs. Our dataset is available to the public to promote further research and development for multilingual learning: https://huggingface.co/datasets/uonlp/CulturaX. To our knowledge, CulturaX is the largest open-source multilingual dataset to date that is deeply cleaned and deduplicated for LLM and NLP applications.

[2] https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama

[3] https://blog.allenai.org/dolma-3-trillion-tokens-open-llm-corpus-9a0ff4b8da64

[4] https://oscar-project.org

2. Multilingual Dataset Creation

To develop a multilingual public dataset for LLMs, our strategy is to combine mC4 (Xue et al., 2021) and OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021, 2022), two largest multilingual datasets at our disposal. We then process the data with an extensive pipeline, involving two major steps of cleaning and deduplication, to produce an enormous and high-quality dataset for multilingual LLMs.

mC4 is a multilingual document-level dataset, originally created to train the multilingual encoder-decoder model mT5 (Xue et al., 2021) for 101 languages. This dataset is extracted from 71 monthly snapshots from CC by removing pages with less than three long lines (line length filter), pages with bad words, and duplicated lines across documents. Language identification for the pages in mC4 is done by the `cld3` tool (Botha et al., 2017)[5], which is a small feed-forward network (Xue et al., 2021). Any pages with a language confidence below 0.95% are excluded. mC4 is deduplicated with exact match at the document level; however, fuzzy document-level deduplication is not performed. We utilize the latest version of mC4 (version 3.1.0)[6] prepared by AllenAI in this work.

A notable aspect of our dataset pertains to the web-based origin of our selected datasets, mC4 and OSCAR, extracted from CC. This differs from certain previous work (Radford et al., 2019; MosaicML, 2023; Touvron et al., 2023) that has also relied on curated datasets like The Pile (Gao et al., 2020) and BookCorpus (Zhu et al., 2015) to train LLMs, presuming their higher overall quality. However, in the context of multilingual settings, we argue that web-scraped datasets can be a more suitable approach, as curated datasets of superior quality might not be available for various languages. Our strategy of using web-scraped data facilitates efficient data collection across multiple languages, contributing to enhanced training data scales. Furthermore, recent studies have demonstrated the effectiveness of cleaning web-scraped data to yield state-of-the-art LLMs (Raffel et al., 2020; Almazrouei et al., 2023). In total, the combination of mC4 and OSCAR provides us 13.5B documents for further processing. Figure 1 illustrates the distribution of the document counts for mC4 and the four available versions of OSCAR in our initial dataset.

2.1. Data Cleaning

Given the combination of the mC4 and OSCAR datasets, we first perform a comprehensive data cleaning procedure to remove noisy and bad content from the data, including language identification, ULR-based filtering, metric-based cleaning, and document refinement.

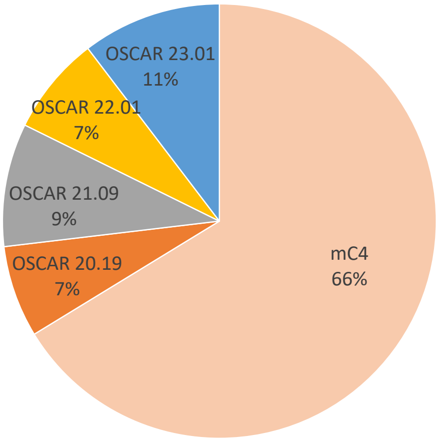

Figure 1: Distribution of document counts from mC4 and OSCAR in our initial dataset.

Language Identification: A particular issue concerns the use of two different language identification tools, i.e., `cld3` and FastText, for mC4 and OSCAR (respectively). It has been shown in previous studies that `cld3` is significantly worse than FastText, causing substantially more language detection errors for mC4 (Kreutzer et al., 2022). In fact, compared to several other language detectors, FastText has demonstrated state-of-the-art performance over benchmark datasets[7]. To this end, our first data cleaning step involves applying FastText to re-predict the languages for the documents in mC4. Documents whose predicted languages are different from the provided ones in mC4 will be removed from the dataset. The rationale is to avoid documents that are confusing for the language detectors `cld3` and FastText, thus potentially introducing noise for the data. Finally, to ensure the highest quality, we remove data for any language found in mC4 but not supported by FastText.

URL-based Filtering: In the next step, we aim to eliminate pages from the known toxic and harmful sources to reduce relevant risks from our data. In particular, we leverage the latest UT1 blacklist of URLs and domains provided by the University of Toulouse to support Internet use regulation for administrators at schools. This list involves sites from different topics, including pornography, grumbling, and hacking, that should be discarded for LLM training. Updated twice to thrice per week, the blacklist involves more than 3.7M records that are

[5] https://github.com/google/cld3

[6] https://huggingface.co/datasets/mc4

[7] https://modelpredict.com/language-identification-survey

contributed by both human and robots (e.g., search engines, known addresses and indexes) (Abadji et al., 2022). As such, we remove any page from our dataset whose associated URL matches a site in the blacklist. This step is helpful for our dataset as the blacklist is not employed before for the mC4 dataset. In addition, although OSCAR has already used this blacklist for data cleaning, our approach incorporates the most up-to-date information from the list, which might not be available for the current distributions of OSCAR.

Metric-based Cleaning: To enhance the dataset's quality, motivated by the data processing pipeline from the BigScience's ROOTS corpus for BLOOM (Laurençon et al., 2022; Scao et al., 2022), we further utilize the distributions for various dataset metrics to identify and filter outlying documents. Each metric provides a singular value for every document within the dataset, quantifying specific attributes such as *number_words*, *stopword_ratios*, and *perplexity_score* for each document. For each metric and its range of possible values within the dataset, a threshold will be determined to partition the range into two zones: a normal range and an abnormal range. The abnormal range is designated for documents exhibiting metric values significantly deviating from the norm, classifying them as outliers/noises, and consequently, these outliers are removed from our dataset. As such, we employ a comprehensive array of dataset metrics, which will be collectively employed to refine our dataset, as outlined below:

- Number of words
- Character repetition ratio
- Word repetition ratio
- Special character ratio
- Stop word ratio
- Flagged word ratio
- Language identification confidence
- Perplexity score
- Document length (number of characters)
- Number of lines
- Short line length ratio
- Short line ratio

The last four metrics are suggested by the OSCAR dataset while the others are inherited from the BigScience ROOTS corpus's pipeline to process OSCAR data. For the perplexity score, following the BigScience ROOTS corpus, we train a SentencePiece tokenizer (Kudo, 2018) and 5-gram Kneser-Ney language models as provided in the KenLM library (Heafield, 2011) using the 20230501 dumps of Wikipedia. Documents displaying high perplexity scores based on these KenLM models are considered notably different from Wikipedia articles. This indicates a level of noise that will be excluded from our dataset (Wenzek et al., 2020). The tokenizer will also be used to obtain the number of words/tokens in the documents for our metrics. We publicly release our KenLM models in HuggingFace[8] to faciliate future exploration.

Repeated information (e.g., words, paragraphs) can appear in the web-curated data due to crawling errors and low-quality sources, causing detrimental consequences for training LLMs (Holtzman et al., 2019). The character and word repetition ratios are thus designed to avoid documents with excessively repeated information. High frequencies of special characters, stop words, or flagged words can indicate noisy and low-quality documents. We thus utilize the stop word and flagged word lists for different languages to compute their ratios for document removal. In addition to the stop word and flagged word lists provided by BigScience ROOTS for their 13 languages, we further collect dictionaries for these types of words for other languages. We prioritize the lists that have been shared on personal GitHub accounts for various languages, as these are often crafted by native speakers and exhibit higher quality. Moreover, lower language identification confidence might also suggest noisy language structures for the data. For each document in the dataset, we thus obtain a language identification confidence via the probability that FastText assigns to its corresponding language to aid data filtering. Finally, for the short line-based criteria, we implement a threshold of 100 characters to classify lines as short, as used by OSCAR. Documents with excessive occurrence of short lines will not be retained in our dataset.

Threshold Selection: Given the set of dataset metrics, an important question concerns the selection of appropriate thresholds for each metric and language to generate high-quality multilingual data. In the BigScience ROOTS project (Laurençon et al., 2022), this selection process is carried out by native speakers of 13 languages. The resulting thresholds are employed for the rest of their 46 languages. The project offers a visualization interface that indexes a sample of a few thousand documents per language, enabling users to monitor data statistics as they adjust thresholds for the metrics. However, this process cannot be easily extended to different languages due to the requirement of experienced native speakers, which incurs significant costs. Furthermore, the limited sample sizes hinder the representativeness of the chosen thresholds for the full datasets. In our analysis, we observe that some selected thresholds for certain languages within BigScience ROOTS almost fall outside the value ranges for the entire dataset, leading to the deactivation of the corresponding metrics.

To address these issues, we leverage a variant

[8] https://huggingface.co/uonlp/kenlm

of the Interquartile Range (IQR) method (Dekking et al., 2007) to select appropriate thresholds for the filtering metrics for our dataset. For each metric and language, we generate a distribution of its possible values across the entire dataset for the language. There is an exception for languages with substantial amounts of data, such as Spanish and Russian, where only 25% of the data is used to calculate these distributions. Afterward, we compute the Q_1-th and Q_3-th percentiles of the distribution ($Q_1 < Q3$) and use them for the thresholds for our filtering metrics. In particular, the lower Q_1-th percentile will be chosen for the metrics that favor high values (e.g., language identification confidence), while metrics favoring low values (e.g., perplexity scores and document length) will utilize the upper Q_3-th percentile. We investigate different values for (Q_1, Q_3), considering $(25, 75)$, $(20, 80)$, $(15, 85)$, $(10, 90)$, and $(5, 95)$. The selection of $Q_1 = 10$ and $Q_2 = 90$ has achieved the best data quality for a sample of languages in our examination.

It is worth emphasizing that the utilization of percentiles for threshold selection enables our approach to efficiently draw upon more extensive data samples for each language compared to those employed in the BigScience ROOTS project. This results in more reliable thresholds for the full datasets over different languages. Specifically, concerning the large languages where only a 25% data sample is employed to compute the value distribution for a metric, we observe that the proportion of discarded data to the entire dataset closely aligns with that of the data sample when applying the same selected filtering threshold. This underscores the representativeness of the thresholds selected through our methodology. Finally, once the thresholds for the metrics in a given language have been determined, we will eliminate any document that surpasses a metric's threshold and enters the unfavorable range of the data.

Document Refinement: The previous cleaning steps are done at the dataset level, aiming to remove low-quality documents from the dataset. In this step, we further clean the retained documents to improve the quality. It is important to note that our prior metric-based filtering step plays a vital role in eliminating highly noisy documents, which, in turn, streamlines the process of developing effective document cleaning rules during this step. Notably, since the documents from mC4 and OSCAR are extracted from HTML pages crawled from the Internet, a significant portion of them may carry crawling and extraction errors, including long JavaScript lines and extraneous content. Consequently, filtering out these documents greatly simplifies our task of designing rules to clean the documents within our dataset.

As such, for each document, we eliminate its noisy or irrelevant portions via a series of operations. First, we remove any short lines located at the end of each document, as these lines typically contain footer details or unhelpful information from the websites. Second, we eliminate the lines that contain words from our list of JavaScript (JS) keywords (e.g., "`<script`") to avoid irrelevant and non-linguistic information. Here, we exclusively remove JS lines if the document contains just one line with JS keywords, and this particular line must also feature at least two different types of JS keywords. We adopt this approach as documents with more than two JS lines are likely coding tutorials in our data, which should be preserved to improve diversity. In addition, certain JS keywords are used in natural language, e.g., "`var`". By requiring at least two different types of JS keywords, we reduce the risk of inadvertently omitting helpful content and disrupting the document's structure.

2.2. Data Deduplication

Despite thorough data cleaning, the remaining dataset might still contain a substantial amount of repeated data due to various reasons, including information being reposted on the web, multiple references to the same articles, boilerplate content, and plagiarism. The duplicated data can thus cause memorization and significantly hinder generalization for LLMs (Lee et al., 2022; Hernandez et al., 2022). Although expensive, data deduplication is thus considered as a crucial step to guarantee the highest quality of data for training LLMs. To this end, we undertake a comprehensive deduplication procedure for our dataset, utilizing MinHash (Broder, 1997) and URLs. This deduplication process is carried out independently for each language. Furthermore, we restrict deduplication to languages that retain over 100K documents following our data cleaning procedures (i.e., 51.5% of our languages), aiming to promote smaller languages within our dataset.

MinHash Deduplication: For each language's dataset, we first apply the MinHashLSH method (Leskovec et al., 2020) to filter similar documents in the dataset. MinHashLSH is a near deduplication technique based on MinHash (Broder, 1997) with multiple hash functions for n-grams and the Jaccard similarity. Locality-Sensitive Hashing (LSH) is incorporated to improve efficiency by focusing on document pairs that are most likely similar. We leverage a variant of the Spark implementation of MinHashLSH in the `text-dedup` repo[9], employing 5-grams and a threshold of 0.8 to determine similar documents for the Jaccard similarity. Running MinHashLSH for each language's dataset, especially

[9] https://github.com/ChenghaoMou/text-dedup/tree/main

		#Documents (M)						#Tokens	
Code	Language	Initial	URL Filtering	Metric Filtering	MinHash Dedup	URL Dedup	Filtering Rate (%)	(B)	(%)
en	English	5783.24	5766.08	3586.85	3308.30	3241.07	43.96	2846.97	45.13
ru	Russian	1431.35	1429.05	922.34	845.64	799.31	44.16	737.20	11.69
es	Spanish	844.48	842.75	530.01	479.65	450.94	46.60	373.85	5.93
de	German	863.18	861.46	515.83	447.06	420.02	51.34	357.03	5.66
fr	French	711.64	709.48	439.69	387.37	363.75	48.89	319.33	5.06
zh	Chinese	444.37	444.03	258.35	222.37	218.62	50.80	227.06	3.60
it	Italian	406.87	406.04	254.72	226.42	211.31	48.06	165.45	2.62
pt	Portuguese	347.47	346.76	217.21	200.11	190.29	45.24	136.94	2.17
pl	Polish	270.12	269.73	170.86	151.71	142.17	47.37	117.27	1.86
ja	Japanese	247.67	247.19	137.88	114.64	111.19	55.11	107.87	1.71
vi	Vietnamese	182.88	182.72	118.67	108.77	102.41	44.00	98.45	1.56
nl	Dutch	238.92	238.56	148.19	125.51	117.39	50.87	80.03	1.27
ar	Arabic	132.88	132.65	84.84	77.65	74.03	44.29	69.35	1.10
tr	Turkish	183.65	183.47	109.94	99.18	94.21	48.70	64.29	1.02
cs	Czech	136.91	136.44	80.38	69.01	65.35	52.27	56.91	0.90
fa	Persian	118.55	118.50	70.26	62.42	59.53	49.78	45.95	0.73
hu	Hungarian	88.59	88.21	53.29	46.89	44.13	50.19	43.42	0.69
el	Greek	100.77	100.68	61.43	54.33	51.43	48.96	43.15	0.68
ro	Romanian	89.37	89.25	45.99	42.8	40.33	54.87	39.65	0.63
sv	Swedish	103.04	102.76	58.67	52.09	49.71	51.76	38.49	0.61
uk	Ukrainian	81.50	81.44	50.95	47.12	44.74	45.10	38.23	0.61
fi	Finnish	59.85	59.80	36.69	32.15	30.47	49.09	28.93	0.46
ko	Korean	46.09	45.85	25.19	21.17	20.56	55.39	24.77	0.39
da	Danish	53.16	52.99	28.67	26.48	25.43	52.16	22.92	0.36
bg	Bulgarian	47.01	46.90	28.09	25.45	24.13	48.67	22.92	0.36
no	Norwegian	40.07	40.01	20.69	19.49	18.91	52.81	18.43	0.29
hi	Hindi	35.59	35.50	22.01	20.77	19.67	44.73	16.79	0.27
sk	Slovak	40.13	39.95	22.20	19.56	18.58	53.70	16.44	0.26
th	Thai	49.04	48.96	26.20	21.93	20.96	57.26	15.72	0.25
lt	Lithuanian	27.08	27.01	15.87	14.25	13.34	50.74	14.25	0.23
ca	Catalan	31.13	31.12	18.99	16.46	15.53	50.11	12.53	0.20
id	Indonesian	48.08	48.05	25.79	23.74	23.25	51.64	12.06	0.19
bn	Bangla	20.90	20.85	13.82	13.22	12.44	40.48	9.57	0.15
et	Estonian	16.20	16.15	9.69	8.45	8.00	50.62	8.81	0.14
sl	Slovenian	15.46	15.39	8.00	7.60	7.34	52.52	8.01	0.13
lv	Latvian	14.14	14.09	8.37	7.48	7.14	49.50	7.85	0.12
he	Hebrew	10.78	10.77	5.90	4.77	4.65	56.86	4.94	0.08
sr	Serbian	7.80	7.75	4.80	4.25	4.05	48.08	4.62	0.07
ta	Tamil	8.77	8.75	5.27	4.94	4.73	46.07	4.38	0.07
sq	Albanian	9.40	9.38	5.96	5.04	5.21	44.57	3.65	0.06
az	Azerbaijani	9.66	9.65	5.73	5.24	5.08	47.41	3.51	0.06
Total (42 languages)		**13397.79**	**13366.17**	**8254.28**	**7471.48**	**7181.40**	**46.40**	**6267.99**	**99.37**
Total (167 languages)		**13506.76**	**13474.94**	**8308.74**	**7521.23**	**7228.91**	**46.48**	**6308.42**	**100.00**

Table 1: Data statistics for 42 languages with the percentages of tokens greater than 0.05% in our dataset. Columns grouped with the "#Documents (M)" label indicate the number of documents for each language after the corresponding cleaning and reduplication steps. The token counts are based on our final dataset (i.e., after all the cleaning and deduplication steps).

for languages with the largest data volumes like English, Russian, Spanish, and Chinese, represents the most computationally expensive operation in our dataset creation effort.

URL-based Deduplication: Finally, we eliminate all documents that share identical URLs with other documents in the dataset. This step is necessary to address situations where various versions of the same articles are linked to identical URLs but have been updated or modified during the publication process, effectively bypassing the near deduplication step. Some URLs for the articles in CC might only display their general domains due to crawling errors. To enhance accuracy, we refrain from removing URLs that only include their general domains.

We utilize 600 AWS c5.24xlarge EC2 instances to preprocess and deduplicate our multilingual dataset. Each instance is equipped with 96 CPU cores, 192GB of memory, and 1TB of disk space. The disk space can be used to replace memory when necessary (e.g., for data deduplication).

3. Data Analysis and Experiments

After completing all the cleaning and deduplication steps, our ultimate dataset comprises 6.3 trillion tokens spanning 167 languages. Table 1 provides an overview of the number of documents and tokens for the top 42 languages in CulturaX following each processing stage. We employ the SentencePiece tokenizer (Kudo, 2018) trained for the perlexity computation in previous step for this table. As can be seen, our data-cleaning pipeline can substantially reduce the number of documents in the original mC4 and OSCAR datasets for each language. The total number of removed documents accounts for 46.48% of our initial documents, suggesting the effectiveness of our approaches to filter noisy information for multilingual datasets. The three largest languages in our final dataset are English (3241M documents and 2.8T tokens), Russian (799M documents and 0.7T tokens), and Spanish (450M documents and 0.37T tokens). Our dataset involves 33 languages with more than 10M documents and 58 languages with more than 1M documents. It also has 32 languages with more than 10B tokens and 58 languages with more than 1B tokens. Finally, our dataset also provides substantial data for low-resource languages. For example, CulturaX contains 4GB, 8.1GB, 6.5GB, and 7.1GB of data for Afrikaans (af), Gujarati (gu), Khmer (km), and Burmese (my), whereas in CCNet (Wenzek et al., 2020), the data for these languages is only 160MB, 190MB, 154MB, and 440MB (respectively).

Data Quality Evaluation: To further evaluate the quality of our data processing pipeline for multiple languages, we obtain a 40GB sample from the initially collected documents in mC4 and OSCAR for each of the three languages: English, German, and Vietnamese. Afterward, we apply our full data processing pipeline (including cleaning and deduplication) over the sampled datasets for each language. This process yields clean datasets, amounting to data sizes of 20GB (5.6M documents and 4.6B tokens), 19GB (5.2M documents and 4.3B tokens), and 21GB (4.8M documents and 4.6B tokens) for English, German, and Vietnamese respectively. In the next step, we train a RoBERT base model over both the original and clean datasets for each language, following the suggested settings in the original RoBERTa paper (Liu et al., 2019). Finally, following the data evaluation in CCNet (Wenzek et al., 2020), we measure the performance of the resulting pre-trained RoBERTa models using the multilingual natural language inference (NLI) task. In particular, we fine-tune the pre-trained RoBERTa models for each language, utilizing the training data from the XNLI dataset (Conneau et al., 2018) that corresponds to the specific language. The performance of the fine-tuned RoBERTa models on the dev sets for each language is reported in Table 2.

Data	English	German	Vietnamese
Initial	81.0	75.0	74.9
Clean	81.9	76.2	76.3

Table 2: NLI dev accuracy for English, German, and Vietnamese of the RoBERTa models that are trained either on initial or clean datasets.

The most important observation from the table is that the clean datasets can produce significantly better performance (with $p < 0.01$) for the RoBERT model than the initial collected datasets. This is remarkable as the clean datasets are only about half the size of initial datasets, thus demonstrating the quality of the clean data and the effectiveness of the data cleaning pipeline in this work.

4. Related Work

Compared to other NLP tasks, language models can be trained with unlabeled data, enabling efficient data collection to produce gigantic scales for the training data. There are two primary types of data commonly used for training LLMs: curated data and web crawl data. Curated data typically consists of well-written and well-formatted text from targeted sources and domains, e.g., Wikipedia articles, books, newswire articles, and scientific papers, as used for the "The Pile" (Gao et al., 2020) and "BookCorpus" (Zhu et al., 2015) datasets. In contrast, web crawl data encompasses text gathered from a wide array of sources across the internet, varying significantly in terms of format and writing styles, e.g., blogs, social media posts, news articles, and advertisements. CommonCrawl (CC) is a widely-used web crawl repository that has collected petabytes of data over the Internet for 12 years. To this end, curated data is frequently considered to possess higher quality, which has resulted in its preference for training early LLMs, e.g., BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019). However, as the demand for larger models has grown, web crawl data has gained more attention as it contributes a substantial portion to the training data of recent LLMs, e.g., RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2020), GPT-3 (Rae et al., 2021), LLaMa (Touvron et al., 2023), MPT (MosaicML, 2023), and Falcon (Almazrouei et al., 2023). As such, different extractions of CC has been produced to train such LLMs, including C4 (Raffel et al., 2020), CC-News (Nagel, 2016), and STORIES (Trinh and Le, 2018).

Regarding the accessibility of training data, datasets used to train early LLMs are often made available to the public (Devlin et al., 2019; Raffel et al., 2020). However, in the case of the most recent state-of-the-art (SOTA) generative LLMs, their training datasets are not released fully, potentially due to commercial interests. This applies not only to proprietary models like ChatGPT and GPT-4 but also to models that claim to be open-source models such as LLaMa, MPT, Falcon, and BLOOM (Scao et al., 2022). To address the transparency issue with existing LLMs, recent efforts have been made to replicate and release the training datasets for the state-of-the-art LLMs, i.e., RedPajama (Computer, 2023), SlimPajama, and AI2 Dolma. The key distinctions for these datasets concern their large-scale text data that has been meticulously cleaned and document-level deduplicated to ensure high quality for training LLMs. Nonetheless, a common drawback of these open-source datasets is that they remain predominantly focused on English data, offering limited data for other languages.

To obtain a multilingual large-scale dataset for training LLMs, it is more convenient to exploit web-scrape datasets such as CC to enable efficient data collection with up-to-date information in multiple languages. In addition, to ensure high quality for high-performing LLMs, it is necessary to extensively clean and deduplicate the multilingual data to avoid noisy and irrelevant content, e.g., low-quality machine-generated text and adult content (Trinh and Le, 2018; Kreutzer et al., 2022; Raffel et al., 2020). As such, a typical data processing pipeline to generate high-quality datasets can involve multiple steps, as demonstrated by FastText (Joulin et al., 2016), CC-Net (Wenzek et al., 2020), the BigScience ROOTS corpus for the BLOOM models (Laurençon et al., 2022; Scao et al., 2022), the RefinedWeb dataset for the Falcon model (Penedo et al., 2023; Almazrouei et al., 2023), and the dataset to train the LLaMa models (Touvron et al., 2023). The first step necessitates in such pipelines language identification to appropriately assign data to their corresponding languages (Joulin et al., 2016). The next steps features various dataset-specific rules and heuristics to filter undesirable content according to the ratios of special characters, short lines, bad words, among others (Grave et al., 2018; Laurençon et al., 2022). The data can also be filtered via lightweight models, e.g., via the KenLM language models (Heafield, 2011), to avoid noisy documents (Wenzek et al., 2020). Finally, data deduplication should be performed to remove similar or repeated information (Laurençon et al., 2022; Penedo et al., 2023). An important step in this regard involves fuzzy deduplication at document level, e.g., via MinHash (Broder, 1997), to eliminate similar documents, thus mitigating memorization and improving the generalization for resulting LLMs (Lee et al., 2022).

To this end, while there are multilingual open-source datasets with text data in multiple languages, such as mC4 (Xue et al., 2021), OSCAR (Ortiz Suárez et al., 2019), CC100 (Wenzek et al., 2020; Conneau et al., 2020), and the BigScience ROOT corpus (Laurençon et al., 2022), their quality and scale do not meet the requirements for effectively training LLMs, particularly generative models such as GPT. For example, as highlighted in the introduction, both mC4 and OSCAR lack fuzzy deduplication for the data at the document level. mC4 also suffers from its poorer language identification due to the use of `cld3`. BigScience ROOTS only provides a small sample data for 46 languages while CC100 does not have information beyond 2018. Our dataset CulturaX thus comprehensively addresses the issues for the existing datasets, offering a multilingual, open-source, and large-scale dataset with readily usable and high-quality data to train LLMs.

5. Conclusion

We present CulturaX, a novel multilingual dataset with text data for 167 languages. Our dataset is cleaned and deduplicated via a comprehensive pipeline, producing 6.3 trillion tokens. CulturaX is thus a large-scale and high-quality dataset, which can be readily used to train high-performing LLMs for multiple languages. Our data is openly accessible to the public to promote further research and applications of multilingual learning.

Ethical Statement

Our work comprehensively processes the combination of the public versions of the mC4 and OSCAR datasets, producing a large-scale and high-quality dataset for multilingual LLM training. We also introduce the URL-based filtering component to remove documents from known toxic and harmful sources to mitigate the related ethical issues. As such, while our dataset undeniably represents a substantial improvement over previous public datasets concerning ethical issues, its vast scale of 6.3 trillion tokens presents a challenge for conducting a granular analysis, such as at the line level, to ensure a completely flawless dataset. Our data might thus still inherit some potential issues from the original datasets, such as hallucination, biases, and private information. To maximally minimize the impacts of these issues, our dataset will be released to enable comprehensive exploration and research from the community. We will also fully respect the policy of the underlying datasets mC4, OSCAR, and Common Crawl in our dataset to facilitate future research for LLMs while limiting the potential ethical issues for the society. Consequently, we do not believe our framework poses any greater societal risks than existing published multilingual datasets

(Xue et al., 2021; Ortiz Suárez et al., 2019; Abadji et al., 2021, 2022).

Acknowledgements

This research has been supported by the NSF grant CNS-1747798 to the IUCRC Center for Big Learning and the NSF grant # 2239570. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government.

Bibliographical References

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

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

Miltiadis Allamanis. 2018. The adverse effects of code duplication in machine learning models of code. *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*.

Ebtesam Almazrouei, Hamza Alobeidli, and Abdulaziz Alshamsi et al. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, and Ehsan Adeli et al. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan McDonald, and Slav Petrov. 2017. Natural language processing with small feed-forward networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2879–2885, Copenhagen, Denmark. Association for Computational Linguistics.

A. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*.

Tom Brown, Benjamin Mann, and et al. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Michel Dekking, Cornelis Kraaikamp, Hendrik Paul, and Ludolf Erwin Meester. 2007. A modern introduction to probability and statistics: Understanding why and how. In *Springer Texts in Statistics*.

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

Leo Gao, Stella Rose Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Danny Hernandez, Tom B. Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, T. J. Henighan, Tristan Hume, Scott Johnston, Benjamin Mann, Christopher Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. 2022. Scaling laws and interpretability of learning from repeated data. *ArXiv*, abs/2205.10487.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *ArXiv*, abs/1612.03651.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *ArXiv*, abs/2103.14659.

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

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. *ArXiv*, abs/2304.05613.

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, Gérard 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 Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. Mining of massive datasets. In *Cambridge University Press*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs.*

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pre-training approach. *ArXiv*, abs/1907.11692.

MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. `https://www.mosaicml.com/blog/mpt-7b`.

Sebastian Nagel. 2016. Cc-news. http: //web.archive.org/save/http: //commoncrawl.org/2016/10/news-dataset-available.

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

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116.

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

Jack Rae, Sebastian Borgeaud, and et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*.

Teven Scao, Angela Fan, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv*, abs/2102.02503.

Hugo Touvron, Thibaut Lavril, and Gautier Izacard et al. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Xiangpeng Wei, Hao-Ran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *ArXiv*, abs/2307.06018.

Laura Weidinger, John F. J. Mellor, and Maribeth Rauh et al. 2021. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359.

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.

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.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.