

# ChineseWebText: LARGE-SCALE HIGH-QUALITY CHINESE WEB TEXT EXTRACTED WITH EFFECTIVE EVALUATION MODEL

A PREPRINT

Jianghao Chen<sup>1,2\*</sup> Pu Jian<sup>1,2 \*</sup> Tengxiao Xi<sup>1,2 \*</sup> Dongyi Yi<sup>3 \*</sup> Qianlong Du<sup>1 \*</sup>  
Chenglin Ding<sup>3</sup> Guibo Zhu<sup>1,2,3✉</sup> Chengqing Zong<sup>1,2</sup> Jinqiao Wang<sup>1,2,3</sup> Jiajun Zhang<sup>1,2,3✉</sup>

<sup>1</sup> Institute of Automation, Chinese Academy of Sciences

<sup>2</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>3</sup> Wuhan AI Research

{chenjianghao2022, jianpu2023, xitengxiao2022}@ia.ac.cn

{qianlong.du, gbzhu, cqzong, jqwang, jjzhang}@nlpr.ia.ac.cn

## ABSTRACT

During the development of large language models (LLMs), the scale and quality of the pre-training data play a crucial role in shaping LLMs' capabilities. To accelerate the research of LLMs, several large-scale datasets, such as C4 [1], Pile [2], RefinedWeb [3] and WanJuan [4], have been released to the public. However, most of the released corpus focus mainly on English, and there is still lack of complete tool-chain for extracting clean texts from web data. Furthermore, fine-grained information of the corpus, e.g. the quality of each text, is missing. To address these challenges, we propose in this paper a new complete tool-chain **EvalWeb** to extract Chinese clean texts from noisy web data. First, similar to previous work, manually crafted rules are employed to discard explicit noisy texts from the raw crawled web contents. Second, a well-designed evaluation model is leveraged to assess the remaining relatively clean data, and each text is assigned a specific quality score. Finally, we can easily utilize an appropriate threshold to select the high-quality pre-training data for Chinese. Using our proposed approach, we release the largest and latest large-scale high-quality Chinese web text **ChineseWebText**, which consists of 1.42 TB and each text is associated with a quality score, facilitating the LLM researchers to choose the data according to the desired quality thresholds. We also release a much cleaner subset of 600 GB Chinese data with the quality exceeding 90%. The data, codes and the tool-chain are available in this website <sup>2</sup>.

## 1 Introduction

Recent years have witnessed the rapid progress of large language models (LLMs). The models, such as GPT-3[5], BLOOM[6], LLaMA[7], Falcon[3], PaLM[8] and GPT-4[9], become more and more powerful, even performing better than humans in some natural language understanding and generation tasks. During the development, it is evident that the scale and quality of pre-training data play a crucial role on LLM's capability. A large-scale and high-quality dataset is the foundation of LLMs and is the source of all the LLM's amazing capabilities.

In order to expedite the research on LLMs, several large-scale datasets have been made publicly available in recent years, such as C4 [1], Pile [2], RefinedWeb [3] and WanJuan [4]. Previous studies usually collect the raw texts at first from various sources, such as Wikipedia, GitHub, ArXiv, Stack Exchange, and CommonCrawl, in which CommonCrawl data often accounts for the vast majority. Then, handcrafted rules are designed to filter out the raw data in three steps: extracting the data in the language of interested, filtering out the noisy texts with language-specific rules and data deduplication. It should be noted that, most of the previous studies mainly focus on the collection of English-centered texts, and there is lack of a complete tool-chain for extracting clean data centered in other languages, e.g. Chinese. Furthermore, previous work usually directly release the final data, without giving the fine-grained information of the

\*Equal Contribution.

<sup>2</sup><https://github.com/CASIA-LM/ChineseWebText>

text, such as the quality of each text, limiting the potential that assists LLM researchers to re-filter the data according to their desired quality threshold.

To address these problems, in this paper we introduce a new complete tool-chain **EvalWeb**, which could extract high-quality Chinese texts from raw web data. The whole process can be divided into two parts. The first part is similar to previous studies and mainly utilizes manually designed rules to filter out explicit noisy data, generating the initial Chinese clean data. This part processes the web texts with two modules: preparation module and processing module. The preparation module first employs a language identification model to extract Chinese data, and then adopts a hash-based deduplication algorithm to remove duplicate texts. The preprocessing module then handles the resulting data with well-designed rules, including length filtering, sensitive words filtering and filtering with Chinese character ratio. Previous studies usually stop after the usage of these rules. In contrast, we introduce the second part which is a quality evaluation module. Due to the diversity of web texts, the remaining dataset after the usage of filtering rules still contains a large number of low-quality text, which cannot be cleared using manually crafted rules. Consequently, we propose to design a BERT-based quality evaluation model for assessing all the remaining data, thereby generating a quality score for each text. Finally, we can select the high-quality Chinese data in the dataset according to the quality threshold. Using our complete tool-chain **EvalWeb**, we release the latest and largest Chinese dataset **ChineseWebText**, which consists of 1.42 TB data and each text is assigned a quality score, facilitating LLM researchers to select data according to a new quality threshold. We also release a much cleaner subset of 600 GB Chinese texts with quality exceeding 90%.

Our contributions can be summarized as follows:

- (1) In this paper, we propose a new complete tool-chain **EvalWeb**, which could extract high-quality Chinese pre-training data from noisy web texts.
- (2) In this paper, we release the latest and largest Chinese dataset consisting of 1.42 TB, and each text in this dataset is assigned a quality score according to our quality evaluation module. We further release a much cleaner subset of 600 GB Chinese texts with quality exceeding 90%.

## 2 Related Work

**Rule-based Text Filtering.** Rule-based text filtering methods are the dominant paradigm to identify content-rich and semantically coherent data from collected raw datasets with handcrafted rules. During the collection of pre-training data, there are a large number of text data on the web. However, these data include a lot of noise, such as violence, pornographic, advertisement and error characters. Consequently, in order to extract high-quality data, several rule-based methods have been proposed to explore how to automatically filter undesired content from noisy web data. In these work, deduplication[10] methods are employed to remove duplicate text from the data, while some handcrafted rules [11; 12] are adopted to filter out violence, pornographic, advertisement and other explicit noisy data. Besides, perplexity [13] is also usually used to evaluate the fluency of the texts. However, these work mainly focus on English and lack a complete tool-chain for Chinese.

**Text Classification Model.** Different from rule-based text filtering methods, text classification model is an alternative approach to identify high-quality data with a well-designed classifier. The simplest text classification model is logistic regression[5], which uses the logistic function to calculate the probability values for each text, and then classifies them into positive or negative with a designed threshold. Currently, BERT[14] and FastText[15] are both commonly used text classification models. BERT is a transformer-based[16] pre-training language model that has achieved remarkable performance in various text classification and understanding tasks. Through pre-training on masked language model and next sentence prediction tasks with a large dataset, this model learns powerful language understanding and representation abilities, which makes it perform well on text classification tasks. FastText[15] is also a neural network based approach which is similar to CBOW[17]. It is characterized by its ability to train efficiently and quickly on large-scale data, while achieving competitive classification performance. In this paper, both of these two approaches will be employed to evaluate the qualities of the Chinese texts.

**Datasets for Pre-training.** In recent years as the scale of pre-trained language models expands, there is a concomitant increase in the demand for large-scale pre-training datasets. Due to the convenience of acquisition and cost-efficiency associated with web-scraped data, it has progressively emerged as a pivotal source for pre-training datasets[18]. In these work, Gao et al. (2020) [19] build a 825 GB English corpus by mixing established natural language processing datasets and several newly introduced ones. This dataset covers 22 diverse high-quality subsets which derive from academic or professional sources, including PubMed Central, the FreeLaw Project, Stack Exchange, Books3[20], OpenSubtitles[21] and so on. Different from the work of Gao et al. (2020) [19], Penedo et al. (2023)[22] demonstrate that properly filtered and deduplicated web data alone can also train a powerful model, even outperforming the LLMs trained on

curated corpora. They use a pipeline approach to filter and deduplicate web data from CommonCrawl at very large scale and then release an English dataset which have 600 billion tokens. In addition, with data from web and synthetically generated textbooks and exercises with GPT-3.5, Gunasekar et al. (2023)[23] build a code dataset which has 7B tokens. They mainly focus on the coding functions of LLMs, especially the writing Python functions. Moreover, He et al. (2023) [24] release a comprehensive multimodal dataset, which also contains texts in both Chinese and English, and is collected from a wide range of web sources. These public datasets mainly focus on English, and lack a complete tool-chain for extracting Chinese clean data from web sources. Furthermore, they only contain the cleaned texts, while missing the corresponding fine-grained information (e.g. the quality of each text) which could help LLM researchers to re-filter texts with new desired quality thresholds.

### 3 Data Construction

Due to the presence of substantial noise and irrelevant information on the web, extracting high-quality Chinese data from the web poses a significant challenge. In order to extract high-quality Chinese text from web effectively, in this paper we propose a pipeline system **EvalWeb**, which integrates manual crafted rules and evaluation models. With this approach, we can effectively filter undesirable content such as offensive speech, advertisements and idle chatter, and then extract high-quality Chinese texts. As in Figure 1, it illustrates the overview of our proposed approach. For the crawled data from web, we first use a preparation module to process them, and then extract the monolingual Chinese data. After that, a preprocessing module will be used to further filter them with manual crafted rules, including data length, sensitive words, proportion of Chinese characters and so on. Finally, a BERT-based evaluation model will be employed to assess the qualities of filtered data. By this way, we can generate a quality score for each of the text, and then use an appropriate threshold to extract the high-quality data as we required. Furthermore, considering computational cost and efficiency, we further propose to leverage knowledge distillation[25] techniques to train a FastText classifier, which can achieve similar performance with faster efficiency and lower computational costs.

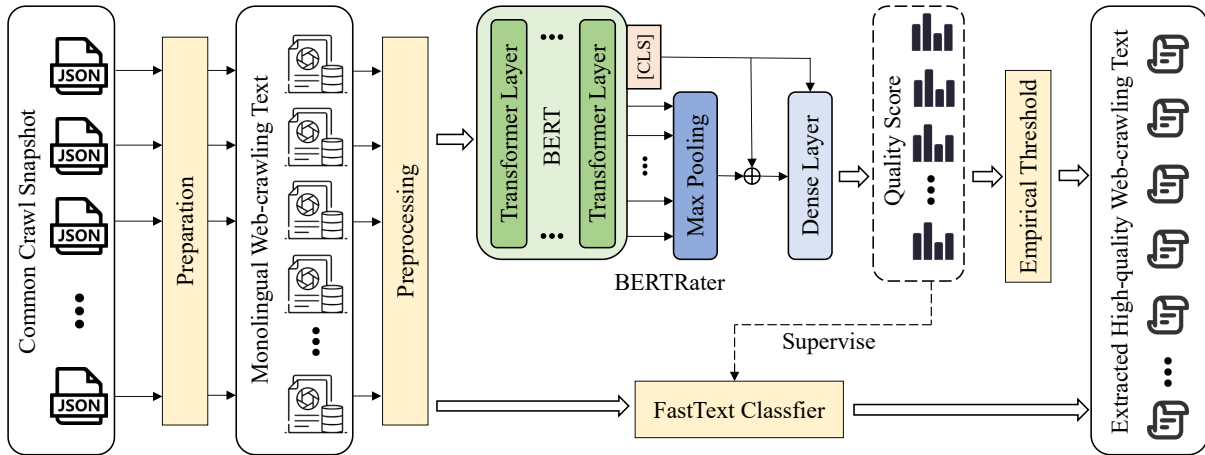


Figure 1: The architecture of our EvalWeb approach.

#### 3.1 Data Collection and Preparation

As a publicly accessible web scrape dataset, CommonCrawl has been running for 12 years and has accumulated petabytes of web data. Consequently, we regard it as the source of our web data. In this paper, we collect nine latest CommonCrawl snapshots from the internet, including "2021-43", "2022-05", "2022-21", "2022-27", "2022-33", "2022-49", "2023-06", "2023-14" and "2023-23". These obtained snapshots are compressed plain text, each of which is 8-10 TB in size (approximately 3 billion web pages). Each snapshot is regrouped into JSON format shards of 5 GB, where each item corresponds to a web page.

Since the original CommonCrawl file is typically quite diverse and contains texts in many languages, to efficiently extract Chinese text data from it, we first employ a deduplication and language identification (LID) module to perform preliminary cleaning on the collected datasets. Following the work of CCNet[13], in this module a Hash-based inter-string deduplication method is employed to remove duplicate text from different snapshots. Additionally, a well-trained

language identification model[26], which could support 157 languages, is applied to select Chinese data. By this way, we can obtain all the monolingual Chinese text data we required.

### 3.2 Preprocessing

Table 1: Examples for different filtering rules.

Filtering Operation	Example
Text Extraction	{ "url": "http://sarahokane.com/ywly/index.aspx", "date_download": "2022-05-28T10:17:39Z", "length": 854, "nlines": 17, "source_domain": "sarahokane.com", "title": "合肥市建设投资控股(集团)有限公司", "raw_content": "乡村振兴和现代农业板块乡村振兴与现代农业...注册资本4.39亿元。", ... "language": "zh", "bucket": "head" }
Length less than 200	"德国黑森州法兰克福AMazon数据中心"
Average line length less than 10	"汽车资讯\n汽车制造商\n学车租车\n俱乐部,汽车网,汽车报价,新车,汽车图片\n汽车网,汽车报价,新能源,新车,汽车图片\n南方网: 汽车频道\n维修,改装,车模,新车,用车\n ....."
Traditional Chinese characters	"有看過裝潢中的便利商店嗎？它可能內部還沒整備好，甚至根本是空盪盪的一片。但店門一定會上紅布條，寫著「XXX便利商店在此為您服務」。去過外縣市旅遊嗎？當你開著車要找某家店時，你大概不會開到正門口才看是不是你要找店。"
Proportion of Chinese characters fewer than 30%	"线上买球平台\u0006\u0007、精准\u0007\u0005、可靠的传感技术解决方案及产品\n\nXSENS MTi 600系列 ... \n\n7 GNSS/INS\n线上买球平台\u0006，具有多个GN.....</p>\nMTi-G-710 GNSS/I...<p><span style="font-size: 12px;">....."
Occurrence of sensitive words more than 0.5 per line	"宝博体育强化创新引领、坚持“科技宝博体育”战略，构建“以企业为主体、市场为导向、产学研相结合”的科技创新体系。 \n2019年度大事记\n友情链接: 玩球直播nba 真钱滚球真人 金花三张牌赢钱 55直播nba 买球 手机轮盘app 亚博app在线登录 亚博app登录....."
Internal duplication ratio greater than 50%	"“山东省民间融资机构宣传月活动”.....活动期间吸引了当地市民\n2018年11月9日，“山东省民间融资机构宣传月活动”.....活动期间吸引了当地市民的广泛关注，现场发放宣传资料600余份，解答市民疑问100余人次，德州广播电视台，德州日报，德州晚报全程采访报道。"

After getting the monolingual Chinese web data, in this section we will focus on how to extract high-quality Chinese texts from them. Given the prevalence of violent, pornographic, advertising, and error characters in web data, we will first employ some manually crafted rules to filter out these noisy data. The details of these crafted rules are presented in the following.

**Text Extraction** After the data preparation stage, there exists a substantial amount of redundant content, which holds little substantive value for subsequent analysis, such as irrelevant key-value pairs. To ensure the accuracy and efficiency of data analysis, we initially undertake the task of extracting all textual content from the entire dataset.

**Data Length** In web texts, a substantial portion of data consists of documents with short text lines which are separated by '\n'. And the texts in different lines do not have significant semantic relevance to each other, which results in that these data are not useful for the training of language models. To eliminate excessively documents with short text lines,

we will calculate the average text line length for each document and then remove documents with an average line length of fewer than 10 characters. Besides, during pre-training procedure, short text data usually contains limited information, making it ineffective in providing context and contextual information. Consequently, we will remove texts whose length is less than 200.

**Proportion of Characters** The objective of this paper is to create a high-quality simplified Chinese dataset sourced from web data. Therefore, we firstly eliminate text data composed of traditional Chinese characters. Additionally, we found that some data exhibit a notably low percentage of Chinese characters. And the rest of these data are filled with some other language characters, non-essential characters, special symbols, irrelevant markers, and so on. These data are unhelpful for the training of large language models. Consequently, we will remove the texts with fewer than 30% Chinese characters.

**Sensitive Words** In web data, there is a large amount of harmful texts, including sex, gambling, violence, discrimination, drugs, religion, and so on. These texts can make large language models generate toxic contents, which have a big negative influence on society, nations and individuals. To avoid these issues, it is necessary to filter out harmful content from the web texts. Firstly, we collect a lot of harmful words and build a sensitive word list. After that, we count the sensitive words in each line of the texts. For one text, if the occurrence of sensitive words exceeds 0.5 per lines, we will regard it as a toxic text and will remove it from our dataset.

**Internal duplication** In the training of large language models, duplicate texts can significantly impact training efficiency and model performance. Although we have conducted deduplication in the first stage, subsequent analysis revealed that some duplicate information still exists in the texts. Therefore, we adopt a granularity of 13-gram for analysis, quantifying the proportion of repetitive 13-gram character sequences across all data. When the proportion of repeated 13-gram characters in a data sample exceeds 50%, we opt to filter it out.

Through the aforementioned rigorous data preprocessing steps, a substantial amount of low-quality data is filtered out. After that, a quality evaluation model will be employed to evaluate the quality scores of the remaining data and then extract the high-quality data with a desired threshold.

### 3.3 Quality Evaluation

#### 3.3.1 BERTEval

In preprocessing procedure, we have used some handcrafted rules to remove the explicit noisy texts from our dataset. However, within the remaining data, there is still a considerable amount of low-quality text data, which cannot be filtered out with handcrafted rules. In order to extract the data of higher quality from them, in this section we further propose to design an evaluation model. In this approach, we will develop a BERT-based classification model to generate a quality score for each text, and then filter the high-quality data with a threshold. The details of the classification model are presented in the following.

**Training Data Composition** While the evaluation in our current experiment targets CommonCrawl data, we believe the positive training samples should encompass a variety of text types, such as Wikipedia, e-books, poetry, news, and Q&A data, to prevent the model from exhibiting bias toward deeming any specific text type as high quality. Since CommonCrawl data has a relatively high noise level overall, we directly sampled from CommonCrawl and used the sampling results as negative examples. Table 2 presents the detailed composition and quantity of the training data.

Table 2: Composition of BERTEval Training Data.

Type	Source	Size ( $\times 10^4$ )
Positive samples	Wikipedia	12.50
	Sina News	12.50
	Cbooks	12.50
	Zhihu	12.40
	WikiQA	0.90
	Law	0.40
	Poetry	0.20
	GovReport	0.13
Negative sample	CC-sampling	55.00

**BERTEval Architecture** We utilized Tran-BERT-MS-ML-R[27], an effective AES model based on the BERT-base architecture, to evaluate the quality of the text obtained from web crawling. To reduce computational complexity, we opted to exclude the sub-document scale representation in Tran-BERT-MS-ML-R, which employs text segmentation at various scales as input. Instead, we focused solely on the text-scale representation, utilizing the  $[CLS]$  embedding to extract pertinent information and structural features from the broadest perspective of the text. Simultaneously, the token-scale representation was derived from the sequence outputs of BERT. We anticipate that this token-scale representation will be instrumental in identifying and filtering out texts containing offensive language, sexually explicit terms, and frequent nonsensical vocabulary, typically absent in high-quality corpora. Let  $x$  represent the text input. After the application of Max Pooling, the token-scale representation is concatenated with the text-scale representation and passed through a Dense Layer with Sigmoid activation, producing a text quality score  $f(x|W)$  that falls within the  $(0, 1)$  range.

**Loss Function** In addition to the MSE loss[28], we used the following two loss functions: Margin Ranking ( $MR$ ) loss [29] and Cosine Similarity ( $CS$ ) loss [27]. Let  $D$  denote the CommonCrawl corpus. Each negative sample  $x_n$  is sampled from CommonCrawl associated with a fixed low score label  $y_n$ , which formed  $D_n$ . The positive sample,  $x_p$ , denotes curated corpora with ideal quality with a constant score  $y_p$ , which formed  $D_p$ . Throughout the training process, the labels for positive and negative samples are persistently  $y_p$  and  $y_n$ , respectively. Moreover, it’s noteworthy that a significant portion of high-quality texts is present within the CommonCrawl corpus. Given that the supervision employed is coarse-grained even somewhat inaccurate, it would be imprudent to solely rely on MSE loss to rigidly compel the model to fit these labels. For the  $ML$  loss, losses only emerge when the ranking of quality scores for samples within each batch doesn’t align with their respective labels. The  $CS$  loss evaluates the correlation between quality scores and their supervision, rather than their absolute differences. Therefore, the combined loss is

$$\mathcal{L}(\mathbf{Y}, f(\mathbf{X}|W)) = \alpha MSE(\mathbf{Y}, f(\mathbf{X}|W)) + \beta MR(\mathbf{Y}, f(\mathbf{X}|W)) + \gamma CS(\mathbf{Y}, f(\mathbf{X}|W)), \quad (1)$$

where  $\mathbf{Y}$  and  $f(\mathbf{X}|W)$  are the quality score labels and predictions of a batch, respectively. Given the supervision is coarse-grained, we contend that the combined loss function offers valuable insights for enhancing our capability of BERTeval to assess the relative quality of texts. The BERTeval training process consists of the following two stages, which are shown in Fig 2.

**Pre-training Stage** At this stage, we extracted positive and negative samples at a 1:1 ratio from the CommonCrawl corpus and the curated corpora. We trained BERTeval based on the loss functions previously described. After this training stage, BERTeval acquired a preliminary ability to discern the quality of web-scraped texts, which will be elaborated in detail in the subsequent experiments section.

**Self-training Stage** As mentioned before, there is a considerable proportion of texts with desired quality in the Common corpus  $D_n$ , which might introduce inaccurate supervision, resulting in neither increasing the epoch nor scaling up the training set leading to a detectable improvement in BERTeval. To ameliorate this problem, we adopt a self-training approach [30]. Let  $S_n$  denote a randomly sampled subset of  $D_n$ . In each self-training iteration, the parameters of BERTeval from the previous iteration,  $W^t$ , are used to generate the pseudo labels of  $S_n$ , and then BERTeval is retrained on sampled data in  $S_n$  with pseudo labels to learn the new parameters  $W^{t+1}$ [31]. It’s worth noting that since the positive samples originate from large-scale, high-reliability, curated corpora that do not require pseudo-labeling, we exclusively sample instances with pseudo-labels being  $y_n$ . The self-training stage can be formulated as:

$$W^{t+1} = \arg \min_W \mathbb{E}_{\mathbf{X}_p^l \subset D_p} \mathbb{E}_{S_n \subset D_n} \mathbb{E}_{\mathbf{X}_n^l \sim p(x_n|W^t), \mathbf{X}_n^l \subset S_n} \{\mathcal{L}(\mathbf{Y}_p^l \oplus \mathbf{Y}_n^l, f(\mathbf{X}_p^l \oplus \mathbf{X}_n^l|W))\}. \quad (2)$$

where  $\mathbf{X}_p^l$  and  $\mathbf{X}_n^l$  denote the vectors consisting of  $l$  randomly sampled samples from  $D_n$  and  $S_n$ , respectively, and  $\mathbf{Y}_p^l$  and  $\mathbf{Y}_n^l$  are the corresponding constant vector labels. Since the output layer is activated by Sigmoid, the quality score  $f(x|W)$  can also be regarded as a probability of positive samples  $p(y_p|x; W)$ . Based on the idea of preferring pseudo-labels with high confidence, in iteration  $t$ , the probability of selecting each sample  $x_n \in S_n$  is denoted by

$$p(x_n|W^t) = \frac{p(y_n|x_n; W)}{\sum_{x \in S_n} p(y_n|x; W)}, \quad (3)$$

which is the normalization of  $p(y_n|x_n; W)$ . Heuristically, we avoid sampling the web-crawled texts where  $p(y_n|x_n; W)$  falls within the last  $K$  proportions. The value of  $K$  is informed by our sampling observations from the CommonCrawl dataset.

### 3.3.2 FastText-based Evaluation Model

To further enhance data processing efficiency and reduce hardware resource requirements, in this paper we also develop a text evaluation model based on FastText<sup>3</sup> in addition to BERTeval. FastText is library for efficient learning of word

<sup>3</sup><https://github.com/facebookresearch/fastText>

Table 3: Composition of FastText Training Data.

Type	Source	Size ( $\times 10^4$ )
Positive samples	Baike	20
	Cbook	20
	Zhidao	20
	China News	20
	Zhihu	20
	WikiQA	10
	other news	10
	BERT-positive	40
Negative sample	BERT-negative	160

representations and text classification. Compared to other classification models, such as SVM, logistic regression, and BERT, FastText could significantly reduce training and inference time while maintaining classification performance.

In the last section, we have built a BERT-based evaluation model BERTEval which performs a good performance on the quality evaluation of Chinese texts. With this model, we can classify the preprocessed web data into high-quality texts (positive) and low-quality texts (negative). Inspired by the idea of knowledge distillation, we will use these classified texts to guide the training of our FastText model. In our approach, we select 400,000 high-quality texts classified by BERTEval as our positive data, while choosing 1,600,000 low-quality texts as our negative data. In order to increase the diversity of training data, our positive data also include some high-quality Chinese data from some other websites and books, such as Baidubaike, Zhihu, Cbook, ChinaNews and so on. These data have been manually proofread and processed. In this way, we can build a good training dataset with 3200K samples. As shown in Table 3, it presents the composition of our training data.

After collecting these training data, we will use a word segmentation tool to process all the texts, and then input the processed data into FastText to train the model. Through this approach, we can obtain a more efficient quality evaluation model.

### 3.3.3 Evaluation Model Comparison

In order to compare the performance of different evaluation models, in this section we evaluate them on a test set which includes 300 samples. Here we first list other two baseline quality evaluation models: regression-based approach and perplexity-based approach.

#### Regression-based Evaluator

Following the work of Gururangan et al. (2022) [32], we combine logistic regression with a word frequency-based vectorization method to conduct text classification on the testset. In this approach, logistic regression is used to calculate a probability value for each sample, and then a threshold is adopted to determine whether the data point should be classified into positive or negative.

#### Perplexity-based Evaluator

Perplexity could effectively measure the difficulty of a language model in predicting tokens and reflect the fluency of the input texts. Following the work of Wenzek et al. (2020) [13], we utilize a well-trained language model to calculate the perplexity of the texts and classify them with a threshold based on perplexity values. The samples with lower perplexity values will be classified into positive.

### Comparison Results

During testing procedure, we will classify the samples of testset with Regression, Perplexity, BERTEval and FastText models respectively, and then compute the precisions of them on positive data. As in Table 4, it shows the precisions of different evaluation models on the testset. TP represents the number of "True Positive" samples, while FP represents the number of "False Positive" samples. From this table, we can see that our BERTEval evaluation model gets a much better performance than the regression and perplexity approaches. Besides, benefiting from the good classified results of our BERTEval model, the FastText-based model could further improve the classification precision. This result indicates that using BERTEval to guide the construction of the FastText-based evaluation model is effective. And with this FastText-based evaluation model, our EvalWeb tool-chain could achieve a better performance while effectively improving processing efficiency and resource utilization.

Table 4: Classification results of different evaluation models.

Model	Precision(%)	TP+FP	TP
Regression	49.57	234	116
Perplexity	63.27	245	155
BERTEval	73.79	103	76
FastText	<b>81.58</b>	76	62

### 3.4 Quality Control

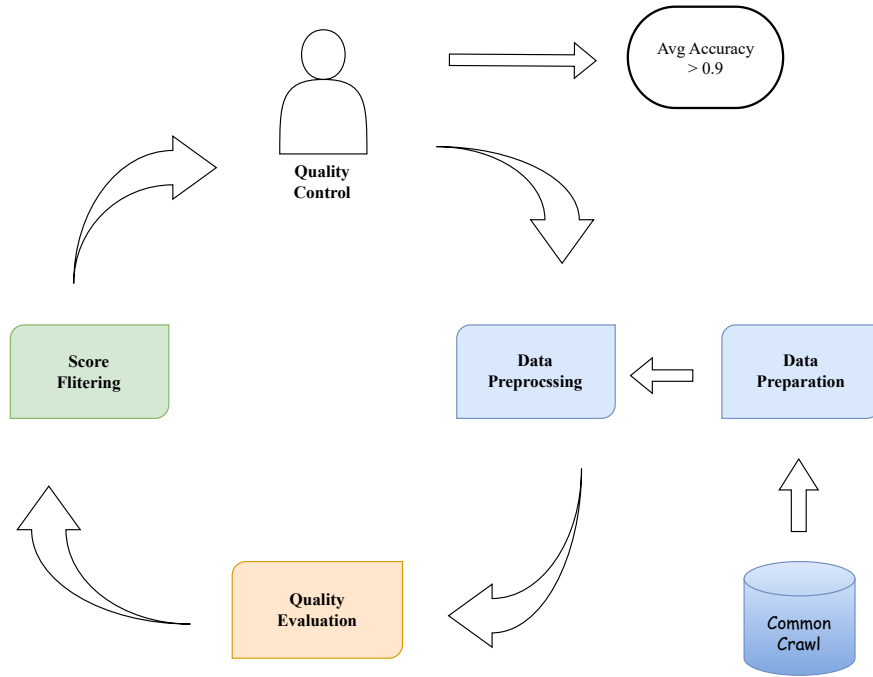


Figure 2: Quality Control.

In section 3.3, after filtering data with a desired quality threshold, we can obtain a Chinese text dataset. In order to ensure the quality of this dataset, We will hire some human evaluators to evaluate its quality. In this method, we will randomly sample 1000 examples from the dataset for three times. After that, three human evaluators will be hired to assess the quality of these data respectively, and the quality of these data is required to be evaluated from the following four aspects:

- **Informativeness:** Whether the text contains enough knowledge and information, or is just meaningless crap.
- **Fluency:** Whether the text has formatting issues, capitalization mistakes, or evident grammatical errors that impair readability.
- **Coherence:** Whether the text progressively forms a coherent body of information on a topic through its successive sentences.
- **Toxicity:** Texts used for pre-training should endeavor to exclude offensive remarks, sexually explicit content, and politically sensitive statements to mitigate potential generative risks.



Table 5: Data samples in json format. The higher score of Sample 1 versus the lower score of Sample 2 demonstrates their differing text qualities, with Sample 1 having better quality than Sample 2.

Key	Value
<title>	"潍坊银行2021年上半年净利润同比增长29.57% 不良率降至1.10%_财经_中国网"
<score>	0.95
<text>	"潍坊银行2021年上半年净利润同比增长29.57% 不良率降至1.10%\n中国网财经8月24日讯 潍坊银行昨日披露2021年二季度信息报告显示，截至2021年6月末，潍坊银行资产总额1920.44亿元，较上年末增长9.34%；负债总额1789.16亿元，较上年末增长10.54%。2021年上半年，潍坊银行实现净利润6.09亿元，同比增长29.57%。资产质量方面，截至2021年6月末，潍坊银行不良贷款率1.10%，较上年末下降0.13个百分点。资本金方面，截至2021年6月末，潍坊银行资本充足率、核心一级资本充足率、一级资本充足率分别为11.66%、7.89%、10.13%，分别较上年末下降1.89、0.89、1.15个百分点。"
<url>	http://finance.china.com.cn/news/special/2021bnb/20210824/5638343.shtml
<source_domain>	finance.china.com.cn
<title>	"上海巨也仪器设备有限公司"
<score>	0.19
<text>	"石子冲击试验仪\n现货提供石子冲击试验机\n现货提供石子冲击试验机符合SAE、ASTM、VDA、GM、Ford、Mazda、JIS、Nissan、及Toyota等的测试要求。石子冲击试验机（欢迎实地考察）满足：大众、神龙、通用、日产、马自达、丰田、本田、福特等汽车厂家试验。耐石子冲击性能试验机\n耐石子冲击性能试验机主要用于、德系日系、美系汽车厂家试验方法，它能准确再现由飞溅的砂砾造成的破化现象，适用于外涂层粘聚性破坏试验、涂层系统中不同层间粘合性破坏试验、抗剥落的*涂膜厚度、塑料及玻璃的抗剥落、抗碰撞、抗磨损测试等相关试验。石子冲击试验机（欢迎实地考察）\n石子冲击试验机（欢迎实地考察）符合SAE、ASTM、VDA、GM、Ford、Mazda、JIS、Nissan、及Toyota等的测试要求。满足：大众、神龙、通用、日产、马自达、丰田、本田、福特等汽车厂家试验。漆膜抗石击试验仪\n漆膜抗石击试验仪符合SAE、ASTM、VDA、GM、Ford、Mazda、JIS、Nissan、及Toyota等的测试要求。石子冲击试验机/抗石子冲击仪\n巨也仪器！有大量现货提供，欢迎客户随时来厂参观与指导！"
<url>	http://www.juyesh.com/SonList-1094890.html
<source_domain>	www.juyesh.com

During evaluation procedure, each text is assigned a label of either "True" or "False." "True" indicates that the data meets the quality requirement of pre-training in all four aspects, while "False" signifies that the text is noisy to some extent. After completing all the evaluations, we will calculate the average accuracy of these three evaluators. If the average accuracy could exceed 0.9, the filtered Chinese dataset is considered to be a high-quality dataset. Otherwise, we believe that there is still some noisy text in the dataset, and we need to optimize the preprocessing and evaluation modules again, and reprocess the dataset until it meets the quality requirements. The architecture of quality control process is illustrated in Figure 2.

### 3.5 Dataset Statistics and Comparison

After processing the collected CommonCrawl data with preprocessing and quality evaluation modules, this paper constructs a clean Chinese dataset ChineseWebText, which consists of 1.42 TB data. As shown in Table 5, each text in this dataset is assigned a quality score which is generated by the quality evaluation model BERTEval. In this table, a larger quality score signifies a higher text quality. With these quality scores, LLM researchers could further select data according to a desired quality threshold. In addition to ChineseWebText, this paper also release a much cleaner subset of nearly 600 GB Chinese texts, which is built by choosing data from ChineseWebText with quality scores in the top 40%. Through manual evaluations with three evaluators, the accuracy of this cleaner subset reaches 90%. Table 6 shows the details of our datasets, which are extracted from nine CommonCrawl snapshots.

Table 6: Overview of output datasets.

Snapshot	Data Size(GB)		
	Monolingual Chinese Data	ChineseWebText Dataset	Cleaner Subset
<b>2021-43</b>	505.92	187.57	78.95
<b>2022-05</b>	442.47	164.96	69.44
<b>2022-21</b>	443.57	166.75	70.19
<b>2022-27</b>	417.95	149.41	62.70
<b>2022-33</b>	369.56	123.70	51.98
<b>2022-49</b>	445.29	160.87	67.76
<b>2023-06</b>	396.40	173.47	74.19
<b>2023-14</b>	441.46	150.04	63.33
<b>2023-23</b>	371.96	143.93	61.28
<b>Total</b>	3834.58	1420.70	599.82

Table 7: The comparison of different pre-training datasets.

Dataset	Lang.	Availability	Publ c Size	Scoring
C4[1]	EN	Public	807GB	NO
The Pile[2]	EN	Public	825GB	NO
REFINEDWEB[3]	EN	Public	2.8TB	NO
WuDaoCorpora[33]	ZH	Pratly Public	200GB	NO
ROOTS-zh[34]	ZH	Public	265GB	NO
WanJuan1.0-zh[4]	ZH	Public	550GB	NO
<b>ChineseWebText (Ours)</b>	ZH	Public	1.4 TB	YES
<b>Cleaner Subset (Ours)</b>	ZH	Public	600 GB	YES

In Table 7, we compare our datasets with some other public pre-training corpora. In these work, the researchers first collect raw data from different sources, such as BookCorpus, Github, Arxiv, PubMed Central, CommonCrawl and so on. And then they clean them with some well-designed rules and algorithms. Specifically, C4[1], The Pile[2] and REFINEDWEB[3] are three public English datasets, while WuDaoCorpora[33], ROOTS-zh[34] and WanJuan1.0-zh[4] are three corpora for Chinese. From this table, we can see that our datasets are the latest and largest Chinese datasets. Besides, different with these previous datasets, each text in our datasets is also assigned a quality score, which could allow LLM researchers to choose data according to a new quality threshold.

## 4 Data Analysis

### 4.1 Removal Rate for Different Stages

To more precisely introduce our data processing workflow, we show Table 8, which details the remaining data size and its corresponding filtering ratio for each preprocessing step and quality evaluation module. In addition, we further depict the processing workflow and the removal rate of each step in Figure 3, thereby providing a high-level overview of the entire process. In each step, we show the removal ratio of data from the previous step and the absolute percentage of the remaining data from the original CommonCrawl. This facilitates readers in conveniently tracking the various processing stages from the raw data to the final data.

Specifically, since the proportion of Chinese data is relatively low in the original CommonCrawl dataset, a large amount of data is filtered out during the preparation stage, retaining only about 4.65% of the original data. In preprocessing stage, data is filtered in several steps. In the step of text extraction, we aim to extract all the text content and remove redundant content generated in the preparation stage, such as useless key-value pairs. Based on this, a variety of manually defined criteria are employed to further refine the dataset, targeting the elimination of texts that either possess limited informative value, contain sensitive or inappropriate content, exhibit a low percentage of Chinese characters, or display redundant characteristics. Due to the presence of numerous entries containing traditional Chinese characters,

the step of filtering based on character proportion results in a large proportion of data being cleaned up. After the preprocessing stage, we score each text of the remaining data using our evaluation model, and then construct the ChineseWebText dataset of 1.4 TB. Finally, We select the top 40% of data based on the quality scores to construct a higher-quality subset of 600GB, which accounts for only 0.73% of the original CommonCrawl data.

Table 8: The remaining data size and filtering ratio for each preprocessing step and quality evaluation module.

Snapshot	Size After filtering operation(GB)						
	Monolingual Chinese Data	Text Extraction	Data Length	Proportion of Characters	Sensitive Words	Internal Duplication	Quality Evaluation
<b>2021-43</b>	505.92	424.43	409.68	217.52	192.84	187.57	78.95
<i>removal rate</i>	-	-16.11%	-3.48%	-46.90%	-11.35%	-2.73%	-57.91%
<b>2022-05</b>	442.47	375.64	362.34	182.88	169.01	164.96	69.44
<i>removal rate</i>	-	-15.10%	-3.54%	-49.53%	-7.58%	-2.40%	-57.90%
<b>2022-21</b>	443.57	363.33	348.51	178.16	170.09	166.75	70.19
<i>removal rate</i>	-	-18.09%	-4.08%	-48.88%	-4.53%	-1.96%	-57.91%
<b>2022-27</b>	417.95	340.65	326.52	158.83	152.33	149.41	62.7
<i>removal rate</i>	-	-18.50%	-4.15%	-51.36%	-4.09%	-1.92%	-58.03%
<b>2022-33</b>	369.56	293.07	280.58	131.39	125.84	123.70	51.98
<i>removal rate</i>	-	-20.70%	-4.26%	-53.17%	-4.22%	-1.70%	-57.98%
<b>2022-49</b>	445.29	367.73	352.59	173.86	164.34	160.87	67.76
<i>removal rate</i>	-	-17.42%	-4.12%	-50.69%	-5.48%	-2.11%	-57.88%
<b>2023-06</b>	396.40	275.04	263.59	211.10	177.44	173.47	74.19
<i>removal rate</i>	-	-30.62%	-4.16%	-19.91%	-15.95%	-2.24%	-57.23%
<b>2023-14</b>	441.46	368.40	354.18	161.54	153.27	150.04	63.33
<i>removal rate</i>	-	-16.55%	-3.86%	-54.39%	-5.12%	-2.11%	-57.79%
<b>2023-23</b>	371.96	305.10	292.58	152.20	146.90	143.93	61.28
<i>removal rate</i>	-	-17.98%	-4.10%	-47.98%	-3.48%	-2.02%	-57.42%
<b>Total</b>	3834.58	3113.39	2990.57	1567.48	1452.06	1420.70	599.82
<i>removal rate</i>	-	-18.81%	-3.94%	-47.59%	-7.36%	-2.16%	-57.78%

## 4.2 Data Quality Distribution

To investigate the relationship between data quality and data quantity, in this section, we adopt different quality thresholds to select data from our ChineseWebText dataset. As shown in Table 9, it presents the high-quality data size for different threshold values. In this table, the value of threshold represents the proportion of selected data in the overall dataset. Then, we hire three human evaluators to assess the quality of the selected data for each threshold. The evaluation criteria has been outlined in section 3.4.

Table 9: The data quality distribution with different quality threshold.

Threshold	High Quality Data Size	Accuracy			
		#1	#2	#3	Average
25%	376.90 GB	92.80%	94.80%	94.90%	94.17%
35%	525.59 GB	93.60%	93.60%	93.10%	93.43%
40%	599.82 GB	90.30%	91.90%	89.50%	90.57%
45%	672.68 GB	84.60%	85.50%	85.59%	85.33%

From Table 9, we can observe that a lower threshold leads to higher data quality, but results in smaller data size. For example, when we keep top 25% of our ChineseWebText, the quality can be higher than 94%, but the remaining data only accounts for 376.9 GB. The threshold of 40% seems to be a good choice, and it can balance the data scale and data quality. The data size can reaches about 600 GB and the average data quality with human evaluation can be higher than 90%. Therefore, this threshold is selected to construct the cleaner subset which is released along with our

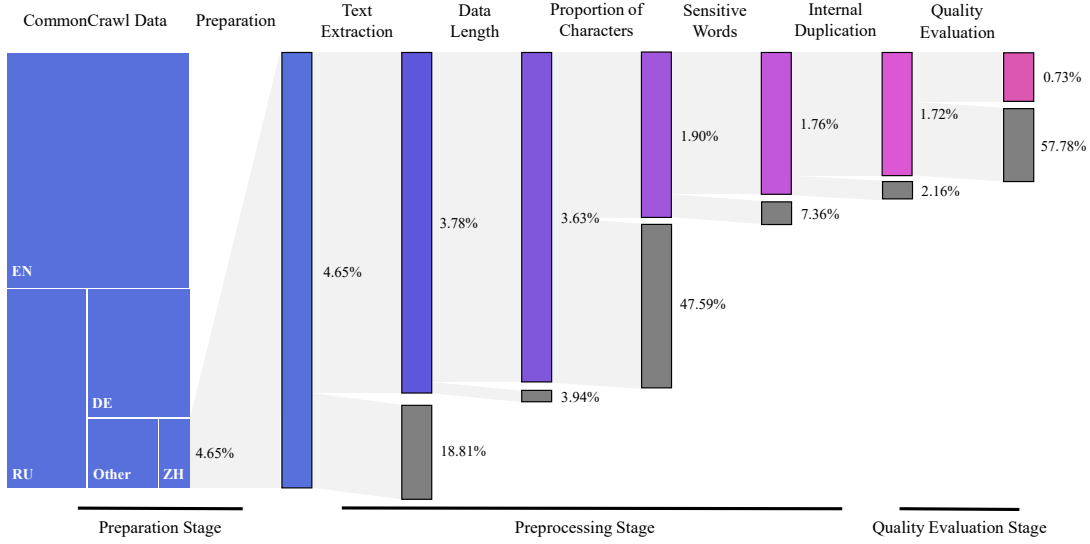


Figure 3: Removal rate for different stages. Grey represents the removal rate with respect to each previous step, while other colors represent the kept rate of all data.

ChineseWebText. Anyway, our ChineseWebText could facilitate the LLMs researchers to choose their own high-quality dataset with their desired threshold.

### 4.3 Data Length Distribution

During the training procedure of LLMs, longer texts can provide more abundant knowledge and information, making it easier for the models to understand complex relationships in the text, and learn more knowledge. In this section, we will analyze the length distribution of the texts in our cleaner dataset. As shown in Figure 4, it illustrates the distribution of text lengths within our cleaner Chinese subset. From this figure, we can observe that the majority of text lengths are mainly distributed within 1000 characters or less. Among them, the most significant proportion is observed within the length interval of 300 to 500 characters. Texts exceeding 1000 characters account for a relatively small portion, and there is a long tail of very long texts. After analysis, we found that the maximum text length in this dataset can reach 300,000 characters. However, they are considered to be outliers and excluded from this figure. The text length distribution in this figure could provide valuable insights into the structure and characteristics of our cleaner subset, thereby help researchers in understanding the composition of the processed dataset and facilitating the utilization of the dataset.

## 5 Conclusions and Future Work

In order to extract large-scale and high-quality Chinese pre-training data from the web, we have proposed a new pipeline approach which filter the raw crawled web data with both handcrafted rules and well-designed quality evaluation model. The rules are employed to first extract the Chinese texts and remove duplicate documents, and then filter out the explicit noisy contents such as toxic texts and advertisements. The quality evaluation model is designed based on BERT and can assign each text with a quality score. With the proposed approach, we release the latest and largest Chinese dataset of 1.4 TB, each of which is associated with a quality score, facilitating the LLMs researchers to re-filter the data with desired quality thresholds. We further release a much cleaner subset of 600 GB Chinese data with the quality exceeding 90% by human evaluation. We also release the complete tool-chain that processes the raw data into the clean texts.

In the future, we will continue to enlarge the Chinese dataset with newly incoming web data. Meanwhile, we are going to explore better algorithms and strategies for data filtering. For example, we can design quality evaluation models for each kind of data noise.

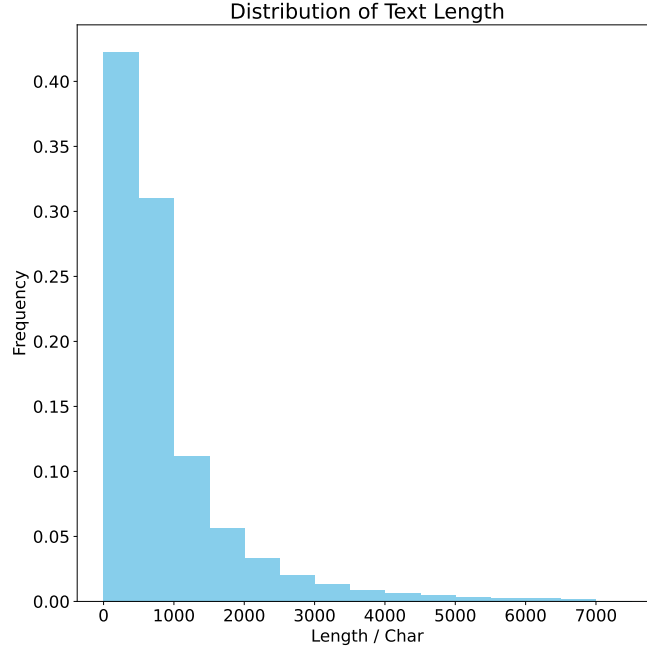


Figure 4: Length distribution of Data.

## References

- [1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [2] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [3] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023.
- [4] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [6] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.
- [8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [9] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

- [10] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 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, 2022.
- [11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. ISBN: 1532-4435 Publisher: JMLRORG.
- [12] Alexandra Luccioni and Joseph Viviano. What’s in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online, August 2021. Association for Computational Linguistics.
- [13] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, 2020.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [15] Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017.
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [17] Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.
- [18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [19] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [20] Shawn Presser. Books3, 2020.
- [21] Jörg Tiedemann. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3518–3522, 2016.
- [22] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- [23] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [24] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023.
- [25] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34:6906–6919, 2021.
- [26] Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [27] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States, July 2022. Association for Computational Linguistics.

- [28] Mohsen Mesgar and Michael Strube. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4328–4339, 2018.
- [29] Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, Haiying Wu, and Xiaojie Yuan. Temp: taxonomy expansion with dynamic margin loss through taxonomy-paths. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3854–3863, 2021.
- [30] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.
- [31] Subhabrata Mukherjee and Ahmed Awadallah. Uncertainty-aware Self-training for Few-shot Text Classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21199–21212. Curran Associates, Inc., 2020.
- [32] Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 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, 2022.
- [33] Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68, 2021.
- [34] 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. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023.