CCI3.0-HQ: A LARGE-SCALE CHINESE DATASET OF HIGH QUALITY DESIGNED FOR PRE-TRAINING LARGE LANGUAGE MODELS

Liangdong Wang[*], **Bo-Wen Zhang**[*], **Chengwei Wu**[*], **Hanyu Zhao**[*],
Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, TengFei Pan, Guang Liu[†]

Beijing Academy of Artificial Intelligence(BAAI)

ABSTRACT

We present CCI3.0-HQ[3], a high-quality 500GB subset of the Chinese Corpora Internet 3.0 (CCI3.0)[4], developed using a novel two-stage hybrid filtering pipeline that significantly enhances data quality. To evaluate its effectiveness, we trained a 0.5B parameter model from scratch on 100B tokens across various datasets, achieving superior performance on 10 benchmarks in a zero-shot setting compared to CCI3.0, SkyPile, and WanjuanV1. The high-quality filtering process effectively distills the capabilities of the Qwen2-72B-instruct model into a compact 0.5B model, attaining optimal F1 scores for Chinese web data classification. We believe this open-access dataset will facilitate broader access to high-quality language models.

Keywords Chinese Dataset · Pre-Training · Large Language Models

1 Introduction

The success of Large Language Models (LLMs) [1][2] is primarily attributed to the availability of extensive, high-quality pre-training corpora, which underpin their foundational knowledge and reasoning capabilities for a variety of tasks, from creative writing to complex problem-solving. Among them, the Open-source datasets, such as The Pile[3] and Common Crawl[4], have been instrumental in propelling LLM development, fostering collaboration and establishing benchmarks for innovation.

Existing Researchers focus more on scaling high-quality data. Recently the demand for pre-training data has exceeded 10 trillion tokens [1][5][6], underscoring two key trajectories in English pre-training: scaling data and improving its quality. Open-source datasets have rapidly expanded, evolving from collections like the Pile(825GB) to larger datasets such as FineWeb(15TB)[7], which draw extensively from Common Crawl. Simultaneously, the focus has shifted from rule-based filtering methods, as seen in early projects like Redpajama[8], to model-driven approaches exemplified by FineWeb-Edu[7].

Despite the rapid advancement of English open-source datasets, Chinese data remains significantly underrepresented on the global web. Existing open-source Chinese datasets, such as WuDao [9], SkyPile150B [10], and WanjuanV1 [11], are constrained in scale due to a scarcity of Chinese data sources online. Furthermore, there is limited research focused on improving quality classification for Chinese web data, resulting in suboptimal data quality. These challenges present substantial barriers to the development of high-performance Chinese language models, underscoring the urgent need for more effective data filtering and quality classification methodologies.

CCI3.0-HQ: a large-scale Chinese dataset of high quality designed for pre-training large language models
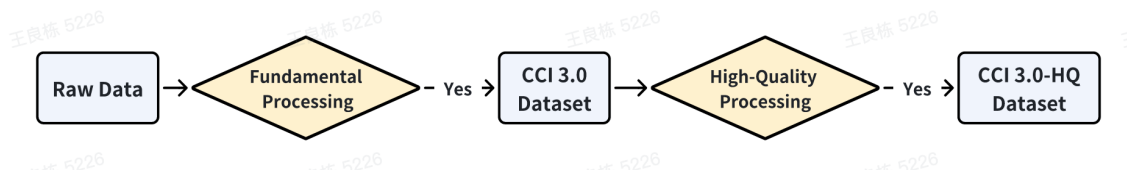

Figure 1: Dataset Curation Pipeline

To address the identified challenges, we present CCI3.0-HQ, a large-scale Chinese pre-training dataset created through a two-stage hybrid filtering strategy: Fundamental Processing and High-Quality Processing. The fundamental stage encompasses standard web data curation practices, including safety filtering, text extraction, deduplication, and initial quality assessment using a basic model score. The second stage further enhances data quality by employing Qwen2-72B-Instruct [2] to identify high-quality samples, resulting in a training set of 140k samples and a testing set of 14k samples. Our analysis shows that these annotations consistently align with GPT-4 annotations at approximately 80%. Consequently, we train a 0.5B quality classifier on the 140k training samples to efficiently filter CCI3.0, producing a high-quality dataset.

To evaluate the impact of our dataset on training LLMs, we conducted a series of experiments using a 0.5B model trained from scratch on a 100B token with a specific data mix, assessing its performance on 10 benchmarks under a zero-shot setting. The extensive results demonstrate that CCI3.0-HQ significantly outperforms competing Chinese datasets, such as SkyPile and WanjuanV1. Additionally, our proposed quality classifier $classifier_{CCI3.0\text{-}HQ}$[5] achieved superior performance, surpassing the $classifier_{FineWeb\text{-}edu}$[6], $classifier_{IndustryCorpus2}$[7], and $classifier_{ChineseWebText}$[8] in terms of F1 score.

In summary, our major contributions are as follows:

- We present CCI3.0-HQ, a groundbreaking 500GB Chinese pre-training dataset that leverages a sophisticated hybrid quality filtering methodology to enhance data integrity.
- We conduct rigorous experimental evaluations, demonstrating that CCI3.0-HQ substantially outperforms the original CCI3.0 dataset and other prominent open-source Chinese corpora, thereby establishing new benchmarks for performance.
- We introduce and open-source the CCI3-HQ classifier, an advanced quality classification tool that significantly improves data selection processes in LLM training.

2   Pipeline

As illustrated in Figure 1, the data processing pipeline comprises two primary phases: **Fundamental Processing** and **High-Quality Processing**. The raw data encompasses a wide array of Chinese corpora, including news, social media, and blogs, thereby enhancing the coverage and representativeness of our dataset. Following the Fundamental Processing steps, we obtain the CCI3.0 dataset. This dataset undergoes further refinement through model-based High-Quality Processing, resulting in the CCI3.0-HQ dataset. The subsequent sections provide a detailed explanation of these two stages in the dataset construction workflow.

2.1   Fundamental Processing

This section outlines the four key processes involved in the Fundamental Processing phase, which is critical for preparing the CCI3.0 dataset and supports subsequent data preparation stages.

- **Safety Filtering**: We implement filters to exclude data from websites likely to contain unsafe content, targeting domains identified as harmful by safety standards and those known for adult material, thereby ensuring compliance with stringent safety criteria.
- **Text Extraction and Cleaning**: Given the dataset's diverse sources, we design specialized parsers for each source to effectively extract and clean the content.

---

[5] https://huggingface.co/BAAI/CCI3-HQ-Classifier
[6] https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier
[7] https://huggingface.co/BAAI/IndustryCorpus2_DataRater
[8] https://huggingface.co/CASIA-LM/ChineseWebText-fasttext

- **Document-Level De-Duplication**: We utilize global MinHash [12] to identify and remove near-duplicate documents, ensuring the dataset's diversity and avoiding redundancy.

- **Heuristic and Basic Quality Filtering**: A set of heuristics is employed to filter out low-quality documents, eliminate outliers, and reduce excessive repetition. Subsequently, we apply a basic quality classifier based on ChineseWebText [13], which predicts the likelihood of a text being referenced by reliable sources, such as Wikipedia, e-books, and news articles.

2.2 High-quality Processing

This stage focuses on the distillation of the quality scoring ability of Qwen2-72B-instruct into a 0.5B model to effectively scoring large amount of data.

2.2.1 Methods for High-Quality Sample Annotation

A primary focus of High-Quality Processing is the precise definition of "high quality" within the pre-training context. After exploring and comparing leading methods, we adopt the FineWeb-edu approach to define high-quality samples and develop a classifier targeting high-quality educational content in Chinese. This aims to enhance the overall quality of Chinese corpora. A detailed comparison of annotation methods and their effectiveness is presented in Section 3.3.

With quality criteria established, the next challenge is efficiently constructing billions of compliant samples. To address this, we implement a structured process for defining and annotating samples according to established benchmarks, ensuring alignment with necessary educational and informational value—critical for robust Chinese language datasets. The workflow is as follows:

We utilize Qwen2-72B-Instruct to score 145,000 random web samples from the CCI3.0 dataset on a scale from 0 (non-educational) to 5 (highly educational), employing a prompt similar to FineWeb-edu. The locally deployed API of Qwen2, integrated with vLLM [14], facilitates the annotation process. Finally, we perform manual and GPT-4 evaluations on a subset of the labeled results, achieving an agreement rate exceeding 80%.

2.2.2 Efficient Training of Quality Classifiers

Labeling all samples for quality identification with a large model like Qwen2-72B-Instruct would be prohibitively costly. Following the FineWeb-edu methodology, we accumulate hundreds of thousands of annotated samples through automated processes and subsequently train a smaller classification model for efficient labeling at scale. This approach significantly reduces costs while ensuring proper identification of high-quality samples, facilitating comprehensive dataset annotation with practical resource investment.

We enhance BGE-M3 [15] (approximately 0.5B parameters) by adding a classification head with a single regression output, training for 20 epochs at a learning rate of 3e-4. During training, the embedding and encoder layers remain frozen to concentrate on the classification head, with dropout not employed. The training script is available on GitHub[9]. The optimal learning rate and intermediate checkpoint are determined based on the F-score across both categories, with training curves documented. For the configuration regarding whether the backbone model is locked in Figure 2, we chose to lock the backbone model, as the performance improvement when it was not frozen was minimal and choosing to lock the backbone model will save a significant amount of training time. Considering the model's generalization ability, we decided to keep the backbone model in a frozen state. Additionally, we performed a grid search for the learning rate in Figure 2.

Finally, the model is converted to a binary classifier using a score threshold of 3 and we apply the classifier to about 1.5 billion samples, a process that requires 9700 A100 GPU hours.

3 Experiments

In this section, we first conduct experiments to evaluate the effectiveness of our curated corpus in pre-training from scratch. Next, we explore and compare two methods for high-quality annotation, detailing our choice of the FineWeb-edu approach for defining and annotating high-quality samples within the Chinese corpus context. Finally, we perform a comparative analysis of existing high-quality classifiers, highlighting the superior performance of our trained classifier.
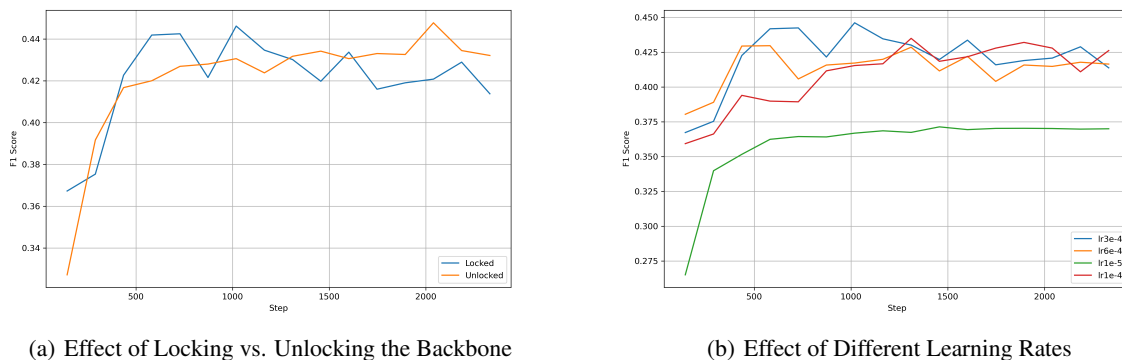
---

[9]`https://github.com/FlagAI-Open/FlagAI/tree/master/examples/CCI3-HQ`

(a) Effect of Locking vs. Unlocking the Backbone

(b) Effect of Different Learning Rates

Figure 2: Effects of Backbone Freezing and Learning Rate Adjustments on Classifier Tuning Performance

3.1 Experimental Setting

3.1.1 Model Training Configuration

For our evaluation, we utilized the Qwen2-0.5B tokenizer and model architecture, training on a bilingual dataset comprising 100 billion tokens. This configuration ensures effective handling of both Chinese and English data while maintaining experimental consistency. Key training parameters include a sequence length of 4096, a weight decay of 0.1, and gradient clipping at 1.0. The training set comprises 25 million samples with a global batch size of 1024. The learning rate starts at 3e-04, with a minimum of 3e-05 and a warmup covering 2,048,000 samples, following a cosine decay schedule.

Parameter	Value
attention_dropout	0.0
bos_token_id	151849
eos_token_id	151850
hidden_act	silu
hidden_size	896
intermediate_size	2432
max_position_embeddings	4096
num_attention_heads	14
num_hidden_layers	24
num_key_value_heads	2
pad_token_id	151643
rms_norm_eps	1e-06
rope_theta	10000
tie_word_embeddings	True
torch_dtype	bfloat16
vocab_size	151851

Table 1: Pre-training Model Configuration Parameters

3.1.2 Dataset Composition

We conducted two primary experiments to evaluate dataset performance:

- **Mixed Dataset Experiment**: This dataset includes 60% English, 10% code, and 30% Chinese content. For the English portion, we employed the FineWeb-edu[10], while the code data was sourced from StarCoder[16].

---

[10] https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

- **Chinese Dataset Experiment**: This experiment utilized a 100% Chinese content dataset, incorporating Wanjuan-v1[11], SkyPile[12], CCI3.0, and CCI3.0-HQ. The CCI3.0 dataset serves as a baseline, as it has not undergone the high-quality filtering process, allowing for a direct evaluation of the impact of quality improvements on dataset integrity.

3.1.3 Evaluation Metrics

We employed the lighteval [17] library for model evaluation, mirroring the setup used with the FineWeb dataset and all evaluation metrics are based on a zero-shot setting. Evaluation metrics include:

- $Average_{Chinese}$: Average score of Chinese metrics, including CEval[18] and CMMLU[19].
- $Average_{English}$: Average score across standard English metrics such as ARC-C[20], ARC-E[20], HellaSwag[21], Winograd[22], MMLU[23], OpenbookQA[24], PIQA[25] and SIQA[26].
- *Average*: Combined average score of all evaluation metrics above.

Additionally, we compared our classifier against existing models using the F1-score derived from the same test samples, employing the macro-average F1-score from 'sklearn.metrics.classification_report' to quantify performance differences.

3.2 Impacts of CCI3.0-HQ on Model Training

We conducted a direct comparison of different datasets through end-to-end pre-training experiments, using the final checkpoint of model training for performance evaluation. Detailed experimental results, including dataset comparisons and metric details, are provided in 2. In analyzing the dataset experiment results, three key points emerge that highlight the performance differences across the various datasets and the strengths of CCI3.0-HQ:

- **Mixed Dataset Experiment Results**: In the mixed dataset evaluations, CCI3.0-HQ consistently performs well across most metrics. Notably, it achieves the highest scores in ARC-E (0.542), Winograd (0.523), MMLU (0.292), and SIQA (0.394), showcasing its robust performance in these specific tasks. CCI3.0, while strong in *HellaSwag* (0.36), is overall outperformed by CCI3.0-HQ. Skypile exhibits good results in *OpenbookQA* (0.334), but lags behind in other metrics, demonstrating a less balanced performance across the board. The ARC-C metric focuses on more challenging knowledge and questions and to address the current gaps in performance for this metric, in future work we plan to increase the quota for high-education-level content to improve results.

- **Chinese Dataset Experiment Results**: In the Chinese-specific evaluations, CCI3.0-HQ stands out significantly, particularly in ARC-C (0.235), ARC-E (0.388), and CEval (0.331). These scores surpass all other datasets, solidifying its advantage in tasks involving the Chinese language. Skypile, while performing well in *Winograd* (0.49) and *OpenbookQA* (0.254), is generally less effective in other areas, while Wanjuan-v1 and CCI3.0 trail behind across most metrics.

- **Average Performance**: CCI3.0-HQ emerges as the top performer with an overall average score of 0.395, compared to 0.388 for Skypile and CCI3.0. In both English ($Average_{English}$ = 0.418) and Chinese ($Average_{Chinese}$ = 0.303) tasks, it maintains a clear lead, affirming its superior generalization across diverse benchmarks. This underlines the effectiveness of the CCI3.0-HQ dataset in enhancing language model training across multilingual tasks.

In conclusion, the results clearly demonstrate that CCI3.0-HQ consistently outperforms the other datasets, particularly in key tasks across both English and Chinese benchmarks, making it a superior dataset for comprehensive multilingual evaluation and model training. We also compare the evaluation performance on each dataset at various intermediate checkpoints during training, as detailed in Appendix 5.1.

3.3 Assessment of Quality Annotation Techniques

In the context of pre-training large-scale natural language models, we explore two leading methods for defining high-quality samples: FineWeb-edu and DataComp-LM(DCLM)[27]. These two methods are based on a new approach that has recently emerged for filtering pre-training datasets of large language models: using synthetic data to develop classifiers for identifying content.

---

[11]https://github.com/opendatalab/WanJuan1.0
[12]https://huggingface.co/datasets/Skywork/SkyPile-150B

Mixed Dataset Experiment Results			
Metrics	SkyPile	Wanjuan-v1	CCI3.0	CCI3.0-HQ
ARC-C	0.270	**0.277**	0.265	0.269
ARC-E	0.521	0.517	0.539	**0.542**
HellaSwag	0.355	0.347	**0.36**	0.357
Winograd	0.507	0.502	0.498	**0.523**
MMLU	0.286	0.287	0.289	**0.292**
OpenbookQA	**0.334**	0.312	0.326	0.318
PIQA	0.651	0.651	**0.652**	0.648
SIQA	0.38	0.387	0.375	**0.394**
CEval	0.279	0.275	0.278	**0.296**
CMMLU	0.294	0.286	0.292	**0.309**
$Average_{English}$	0.413	0.410	0.413	**0.418**
$Average_{Chinese}$	0.287	0.280	0.285	**0.303**
$Average$	0.388	0.384	0.388	**0.395**
Chinese Dataset Experiment Results			
Metrics	SkyPile	Wanjuan-v1	CCI3.0	CCI3.0-HQ
ARC-C	0.192	0.217	0.202	**0.235**
ARC-E	0.313	0.282	0.323	**0.388**
HellaSwag	0.279	0.269	0.283	**0.295**
Winograd	**0.490**	0.487	0.485	0.481
MMLU	0.244	0.254	0.245	**0.259**
OpenbookQA	**0.254**	0.232	0.232	0.242
PIQA	0.528	0.539	0.53	**0.556**
SIQA	**0.387**	0.377	0.372	0.382
CEval	0.305	0.279	0.294	**0.331**
CMMLU	0.304	0.298	0.296	**0.328**
$Average_{English}$	0.336	0.332	0.334	**0.355**
$Average_{Chinese}$	0.304	0.289	0.295	**0.329**
$Average$	0.330	0.324	0.326	**0.350**

Table 2: Comparison of Dataset Impacts on Model Performance in Mixed and Chinese Dataset Experiments

- **FineWeb-edu**: FineWeb-edu evaluates the educational quality of web pages on a scale from 0 to 5, focusing primarily on content aimed at grade-school and middle-school levels. To ensure a balance between educational content of various complexities, a threshold of 3 is used during the filtering process. This allows the retention of not only mid-level educational pages but also some high-level content, ensuring that both foundational and advanced knowledge is included in the dataset for further analysis or training purposes.

- **DCLM**: In the DCLM paper, multiple datasets are compared to identify high-quality samples, ultimately selecting the OpenHermes 2.5[28] dataset as the positive data. This dataset is then used to train a binary classifier. The trained classifier is subsequently applied to the pre-training corpus to identify high-quality content. Since a Chinese version of the OpenHermes 2.5 dataset is not available, the original dataset is translated into Chinese for use in further processing[13].

Through experimental comparisons 3, we find that for Chinese corpora, the FineWeb-edu approach outperformed DCLM especially in $Average_{Chinese}$. The table highlights the performance comparison between two quality annotation methods: *DCLM* and *FineWeb-edu*. Based on the metrics shown, two key points emerge:

- **Performance in Chinese-Specific Metrics**: In the Chinese-specific metrics, *FineWeb-edu* consistently outperforms *DCLM*. Specifically, for the *CEval* and *CMMLU* benchmarks, *FineWeb-edu* scores 0.331 and 0.328 respectively, whereas *DCLM* trails behind with scores of 0.298 and 0.311. Additionally, the $Average_{Chinese}$ score shows that *FineWeb-edu* performs significantly better, with a score of 0.329 compared to *DCLM*'s 0.305.

- **Overall Performance**: In terms of the overall average, *FineWeb-edu* maintains a slight edge with a score of 0.350 over *DCLM*'s 0.348. While the differences in English-focused metrics are minor, *FineWeb-edu*'s stronger performance in Chinese benchmarks, particularly *CEval* and *CMMLU*, demonstrates its superior adaptability for multilingual evaluation.

---

[13]https://huggingface.co/datasets/ldwang/OpenHermes-2.5-zh

In summary, the *FineWeb-edu* method shows stronger results in Chinese-specific evaluations, and its overall performance demonstrates its effectiveness, particularly in tasks requiring higher precision in Chinese language datasets. As a result, we decide to adopt the FineWeb-edu method for identifying high-quality samples in the subsequent steps.

Metrics	DCLM	FineWeb-edu
ARC-C	0.211	**0.235**
ARC-E	0.378	**0.388**
HellaSwag	**0.310**	0.295
Winograd	**0.485**	0.481
MMLU	**0.259**	0.259
OpenbookQA	**0.262**	0.242
PIQA	**0.571**	0.556
SIQA	**0.389**	0.382
CEval	0.298	**0.331**
CMMLU	0.311	**0.328**
$Average_{English}$	**0.358**	0.355
$Average_{Chinese}$	0.305	**0.329**
$Average$	0.348	**0.350**

Table 3: Comparison of Two Quality Annotation Methods

3.4 Evaluation of Quality Classifiers for Chinese Web data

All models are converted into binary classifiers using a score threshold of 3.0 and are evaluated on the same test dataset of about 14k samples. These 14k samples were randomly extracted from a large corpus of Chinese texts, containing both the original text and corresponding labels. They can form a benchmark[14] for subsequent evaluation of text quality. Our high-quality classifier achieves the best performance and detailed results are in Table 4. The table provides a comparison of four classifiers: $classifier_{FineWeb-edu}$, $classifier_{IndustryCorpus2}$ and $classifier_{ChineseWebText}$, and our $classifier_{CCI3.0-HQ}$, evaluated across precision, recall, and F1-score for both positive and negative classes, along with macro averages. Three key observations can be made:

- **Performance on Positive Samples**: The $classifier_{CCI3.0-HQ}$ classifier demonstrates a notable advantage in classifying positive samples, achieving a precision of 0.86 and an F1-score of 0.53. In contrast, $classifier_{FineWeb-edu}$, despite its high precision (0.91), has a recall of only 0.02, leading to a very low F1-score of 0.03 for positive samples. This highlights $classifier_{CCI3.0-HQ}$'s balanced ability to handle positive classes effectively.

- **Macro Average Performance**: The macro average F1-score for $classifier_{CCI3.0-HQ}$ is 0.73, the highest among all classifiers, showing that it maintains strong overall performance across both positive and negative classes. In comparison, $classifier_{FineWeb-edu}$ has a macro average F1-score of 0.47, largely due to its poor performance on positive samples.

- **Balanced Classification**: While $classifier_{IndustryCorpus2}$ performs well in terms of recall for positive samples (0.86), its lower precision (0.32) results in a lower F1-score for positive classification (0.47). On the other hand, $classifier_{CCI3.0-HQ}$ balances both precision and recall across classes, achieving a more consistent and reliable performance overall, especially in terms of the macro F1-score (0.73), which suggests that it is the most robust classifier across all datasets.

- **Importance of Quality Classification**: When compared with the $classifierChineseWebText$, the new added $classifier_{CCI3.0-HQ}$ showcases significantly improved performance in handling diverse data and distinguishing high-quality content. This advancement underscores the essential role that proper quality filtering plays in pre-training, which is also a critical factor contributing to the CCI3.0-HQ dataset's superior performance over the original CCI3.0 dataset.

In summary, $classifier_{CCI3.0-HQ}$ demonstrates superior classification performance, particularly in handling positive samples more effectively and maintaining a strong macro average across all metrics. Compared to the $classifier_{FineWeb-edu}$ mainly trained on English corpora, $classifier_{IndustryCorpus2}$ and $classifier_{ChineseWebText}$ trained on Chinese corpora, our classifier shows significant improvements in both precision and recall for distinguishing positive and negative samples. We attribute this to better suitability for Chinese and data distribution, as well as the amount of training data and model tuning.

---

[14]https://huggingface.co/datasets/BAAI/CCI3-HQ-Annotation-Benchmark

Classifier	Precision	Recall	F1-score
classifier$_{FineWeb\text{-}edu}$			
Positive	0.91	0.02	0.03
Negative	0.82	1.00	0.90
Macro F1	0.87	0.51	0.47
classifier$_{ChineseWebText}$			
Positive	0.18	0.58	0.27
Negative	0.80	0.38	0.52
Macro F1	0.49	0.48	0.39
classifier$_{IndustryCorpus2}$			
Positive	0.32	0.86	0.47
Negative	0.95	0.59	0.73
Macro F1	0.64	0.73	0.60
classifier$_{CCI3.0\text{-}HQ}$			
Positive	0.86	0.38	**0.53**
Negative	0.88	0.99	**0.93**
Macro F1	0.87	0.68	**0.73**

Table 4: Evaluation of Different Quality Classifiers

4   Conclusion and Limitation

We have released and open-sourced the CCI3.0-HQ dataset, which has undergone sophisticated hybrid quality filtering methodology to enhance data integrity. Through comparison after pre-training small-scale models from scratch and rigorous experimental evaluations, CCI3.0-HQ significantly outperforms existing well-known Chinese open-source datasets. Also, We introduce and open-source the CCI3-HQ classifier, which demonstrates superior performance compared to existing open-source Chinese and English quality classifiers and the CCI3.0-HQ dataset demonstrates the importance of high-quality filtering in the pre-training of Chinese large language models. As the largest high-quality Chinese pre-training corpus currently available, CCI3.0-HQ is poised to contribute to the advancement of large language models, especially those focused on Chinese. The dataset includes data collected up until early 2024, which means it might lack information about more recent events or trends. Despite the data cleaning efforts to enhance the dataset's quality, there may still be some lower-quality samples present. We will continue data processing and quality filtering efforts to further support the development of high-quality large language models. As future work, the Infinity Instruct[15] dataset can be leveraged to further optimize the quality classifier, which will lead to additional improvements in the performance of the Aquila series of large language models[29][30].

References

[1] Abhimanyu Dubey etc. The llama 3 herd of models, 2024.

[2] An Yang and Baosong Yang etc. Qwen2 technical report, 2024.

[3] 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, 2020.

[4] Common Crawl. Common Crawl Corpus. `https://commoncrawl.org`, 2024.

[5] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[6] Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei

---

[15]`https://huggingface.co/datasets/BAAI/Infinity-Instruct`

Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. Nemotron-4 340b technical report, 2024.

[7] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024.

[8] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023.

[9] 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.

[10] Tianwen Wei and Liang Zhao etc. Skywork: A more open bilingual foundation model, 2023.

[11] 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.

[12] A. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, page 21, USA, 1997. IEEE Computer Society.

[13] Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. Chinesewebtext: Large-scale high-quality chinese web text extracted with effective evaluation model, 2023.

[14] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

[15] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[16] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023.

[17] Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023.

[18] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, 2023.

[19] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024.

[20] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[21] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

[22] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press, 2012.

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

[24] Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. Careful selection of knowledge to solve open book question answering. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy, July 2019. Association for Computational Linguistics.

[25] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

[26] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019.

[27] Jeffrey Li and Alex Fang etc. Datacomp-lm: In search of the next generation of training sets for language models, 2024.

[28] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023.

[29] Bo-Wen Zhang, Liangdong Wang, Jijie Li, Shuhao Gu, Xinya Wu, Zhengduo Zhang, Boyan Gao, Yulong Ao, and Guang Liu. Aquila2 technical report, 2024.

[30] Bo-Wen Zhang, Liangdong Wang, Ye Yuan, Jijie Li, Shuhao Gu, Mengdi Zhao, Xinya Wu, Guang Liu, Chengwei Wu, Hanyu Zhao, Li Du, Yiming Ju, Quanyue Ma, Yulong Ao, Yingli Zhao, Songhe Zhu, Zhou Cao, Dong Liang, Yonghua Lin, Ming Zhang, Shunfei Wang, Yanxin Zhou, Min Ye, Xuekai Chen, Xinyang Yu, Xiangjun Huang, and Jian Yang. Aquilamoe: Efficient training for moe models with scale-up and scale-out strategies, 2024.

5 Appendix

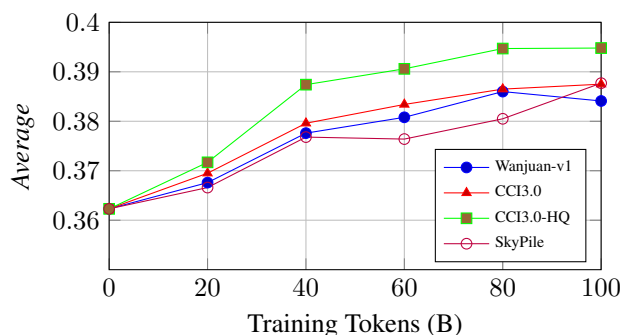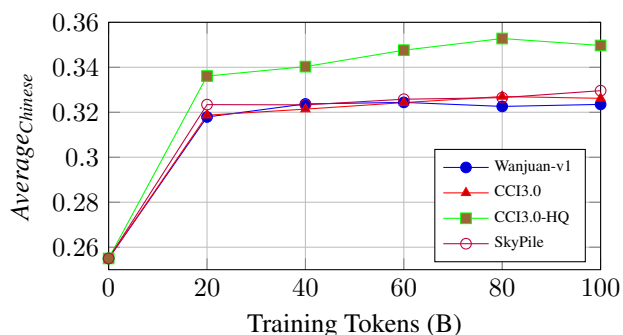
Figure 3: Mixed Dataset Experiment

Figure 4: Chinese Dataset Experiment

5.1 Evaluation of Training Dynamics

We conduct evaluations at every 20 billion tokens of training to closely monitor and compare the performance of various datasets throughout the training process. All results across training tokens of **Mixed Dataset Experiment** and **Chinese Dataset Experiment** are depicted in Figures 3 and 4. The figures compare the performance of different datasets across training tokens in both mixed and Chinese-specific dataset experiments. Three key points emerge from the analysis:

- **Superior Performance of CCI3.0-HQ**: In both the mixed dataset experiment and the Chinese dataset experiment, *CCI3.0-HQ* consistently demonstrates superior performance compared to the other datasets. In the mixed experiment, *CCI3.0-HQ* achieves the highest average score of 0.395 when trained with 100B tokens, significantly outperforming *Wanjuan-v1*, *CCI3.0*, and *SkyPile*. Similarly, in the Chinese-specific experiment, *CCI3.0-HQ* leads with a score of 0.355, showcasing its robustness in both multilingual and Chinese-centric tasks.

- **Gradual Improvement Across Tokens**: All datasets show an increase in performance as the number of training tokens increases. However, *CCI3.0-HQ* demonstrates a steeper improvement curve in both experiments, indicating that it scales more efficiently with larger amounts of training data. In contrast, *Wanjuan-v1* and *SkyPile* show slower growth, particularly in the Chinese dataset experiment.

- **Mixed Dataset vs. Chinese Dataset Performance**: While all datasets perform better in the mixed dataset experiment compared to the Chinese-specific one, *CCI3.0-HQ* maintains a notable gap over the others in both scenarios. This suggests that *CCI3.0-HQ* has been effectively optimized for both general and Chinese-specific data, making it the most balanced and high-performing dataset overall.

The same conclusion confirms that the results highlight the significant advantage of the *CCI3.0-HQ* dataset in both general and Chinese-specific tasks, demonstrating superior scalability and adaptability throughout the training process. The intermediate checkpoints of the models trained in all comparison experiments will be open-sourced [16].

---

[16] https://huggingface.co/BAAI/CCI3-HQ-Intermediate-Checkpoints