Oasis: Data Curation and Assessment System for Pretraining of Large Language Models

Tong Zhou[1], **Yubo Chen**[1,2*], **Pengfei Cao**[1,2], **Kang Liu**[1,2,3], **Shengping Liu**[4], **Jun Zhao**[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Shanghai Artificial Intelligence Laboratory
[4]Beijing Unisound Information Technology Co., Ltd
tong.zhou@ia.ac.cn, {yubo.chen,pengfei.cao,kliu,jzhao}@nlpr.ia.ac.cn, liushengping@unisound.com

Abstract

Data is one of the most critical elements in building a large language model. However, existing systems either fail to customize a corpus curation pipeline or neglect to leverage comprehensive corpus assessment for iterative optimization of the curation. To this end, we present a pretraining corpus curation and assessment platform called Oasis — a one-stop system for data quality improvement and quantification with user-friendly interactive interfaces. Specifically, the interactive modular rule filter module can devise customized rules according to explicit feedback. The debiased neural filter module builds the quality classification dataset in a negative-centric manner to remove the undesired bias. The adaptive document deduplication module could execute large-scale deduplication with limited memory resources. These three parts constitute the customized data curation module. And in the holistic data assessment module, a corpus can be assessed in local and global views, with three evaluation means including human, GPT-4, and heuristic metrics. In addition, an 800GB bilingual corpus curated by Oasis is publicly released.

1 Introduction

Building large language models (LLMs) for proficiency in versatility tasks has been spotlighted recently [OpenAI, 2023; Touvron *et al.*, 2023; Anil *et al.*, 2023]. The power of LLMs only emerges when their parameter size exceeds a certain threshold [Wei *et al.*, 2022], propelling the models to evolve in parameter scale. Recent studies [Kaplan *et al.*, 2020; Rae *et al.*, 2021; Rosset, 2020] have demonstrated that larger models crave a massive, high-quality, and diverse pretraining corpus. The importance of data curation and assessment is increasingly evident.

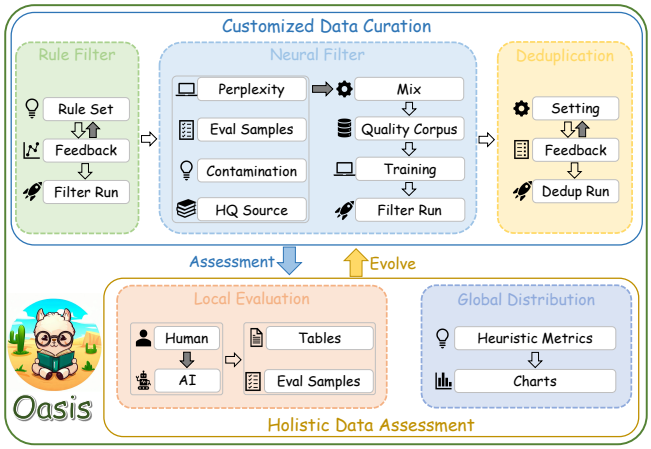

Figure 1: Overview of Oasis functionality.

Data Curation: Some work details preprocessing pipelines for specific sources like Common Crawl [Wenzek *et al.*, 2019; Abadji *et al.*, 2022; Penedo *et al.*, 2023] or Reddit [Gao *et al.*, 2020]. However, these pipelines cannot be directly applied elsewhere because different curation pipelines should be built for various data sources by native speakers of target languages to ensure better quality control [Laurençon *et al.*, 2022]. Unfortunately, an open-source system for customized pretraining data curation is still absent in the community.

Data Assessment: The assessment of the pretraining corpus [Kreutzer *et al.*, 2022; Dodge *et al.*, 2021] aids in the development of LLMs in a data-centric fashion [Fries *et al.*, 2022] more efficiently. It avoids optimizing the data curation by comparing the final model's performance after resource-consuming training. Although there is no conclusion on quantifying the corpus's value, the consensus is that various aspects of pretraining data affect LLM performance, such as fluency, coherence, diversity, and bias [Longpre *et al.*, 2023; Gunasekar *et al.*, 2023]. However, there is still a lack of a holistic data assessment system for the progressive improvement of the data curation pipeline.

In this paper, we present a system for customized pretraining

*Corresponding author
[1]Project: https://github.com/tongzhou21/Oasis
[2]Video: https://youtu.be/YLfMlnrUZPk
[3]Corpus: https://huggingface.co/datasets/Oasis-Team/Oasis-Corpus

data curation and holistic corpus assessment called **Oasis**. The functionality of this system covers three types of filters used to curate high-quality corpora and two perspectives for the holistic assessment of these corpora.

Specifically, in the **Customized Data Curation** part, the first step in our pipeline is an *Interactive Modular Rule Filter* module, which enables users to construct the customized heuristic rule set with hit rate and bad cases as a reference. Then, we debias the neural filter for text quality estimation by paying attention to the process of constructing source-specific quality classification datasets for training, constituting a *Debiased Neural Filter* module. Finally, in the *Adaptive Document Deduplication* module, we optimize the widely used LSH deduplication method in memory requirement and exhibit the effect of different configurations for customized settings. In the **Holistic Data Assessment** part, we provide options to inspect the corpus in sentence fluency and document coherence by humans or GPT-4 in the *Local Quality Evaluation* module. The evaluated cases with quality labels could be further used to evolve the filtering pipeline. Additionally, the *Global Distribution Assessment* module displays the distribution information of the corpus in terms of diversity and richness by multiple heuristic metrics.

Aside from introducing Oasis, we demonstrate a complete case that utilizes this platform to build a high-quality and high-diversity Common Crawl corpus. Meanwhile, we holistic assess the corpus in the different development stages. The assessments also prove the effectiveness of the customized data curation process. In addition, we publicly release an 800GB English-Chinese bilingual corpus Oasis-Corpus cultivated from web pages by Oasis to promote LLM development.

2 System Design and Algorithms

In this section, we will introduce the system design of Oasis and detail the internal algorithms that differ from previous paradigms.

2.1 Customized Data Curation

Interactive Modular Rule Filter. Building a rule filter for the pretraining corpus is a routine in state-of-the-art LLMs. A heuristic rule filter could preliminarily filter undesirable content efficiently. The heuristic ideas for building rules range from text length, punctuation, special tokens, blocklist, and language model perplexity. However, no rule sets can always be valid on various data sources and languages. Corpora from different sources could vary in quality, style, format, template, and meta information. Filter rules in the book field may emphasize removing structural information among high-quality content. On the contrary, when handling documents from the massive web, rules would pay more attention to inspecting the content quality. The essential processes in building and improving the rules involve manually concluding patterns to distinguish high- and low-quality texts and adjusting a single heuristic by examining the hit samples.

We design functions in the Interactive Modular Rule Filter module according to the above intuitions. A user builds a rule pipeline by interactively editing and connecting rule cells, referring to the patterns heuristic summarized from randomly displayed samples. A rule cell could be initiated with the predefined heuristic, and the user could also customize a heuristic function and add it to the predefined pool by typing Python code. Each rule cell's configuration, like thresholds and string patterns, can be freely adjusted according to the inspection of the hit rate and bad cases. After building a customized rule filter pipeline, Oasis can automatically generate a corresponding script according to settings and run the rule filter in the background.

Debiased Model Filter. The original intention of the neural filter is to select high-quality content from massive web pages, similar to high-quality sources like Wikipedia. The model can filter out content with non-summarizable patterns in quality aspects. However, treating another well-known high-quality source as positive and current sources as negative samples could lead the model to bias toward the high-quality source, affecting the quantity and diversity of the filtered data. [Penedo *et al.*, 2023] even abandoned this process due to scruples about the adverse effects of undesirable biases.

To address the bias issue, we propose a negative-centric dataset-building method for neural filter training. This method gathers the majority of positive samples from rule-filtered texts in the current source and obtains most negative samples through heuristic contamination of positive samples. The predefined text contamination rule focuses on coherence and readability, involving shuffling, replacing, inserting, and deleting at the word, span, and sentence levels. The perplexities from the statistical language model may detect these undesirable low-quality contents. However, the perplexity metric is susceptible to low-frequency special tokens and biased towards the training corpus (usually Wikipedia). We use perplexity solely to identify extremely low-quality content, which constitutes a part of the negative samples. These quality patterns are modeled using a neural filter with strong generalization capabilities, such as BERT. The finetuned BERT predicts scores for the text quality of every rule-filtered document. We then drop documents according to the quality score below the threshold.

The Debiased Model Filter module provides a management panel for the quality classification dataset. Users can adjust the composition of positive and negative samples, customize text contamination rules based on editing feedback, and set perplexity quantiles to identify extremely low-quality content through case inspection. Moreover, the dataset for neural classifier training could be further enhanced by incorporating evaluated texts from humans or GPT-4. After building a quality classification dataset, Oasis can generate corresponding scripts through parameter settings on the interface and run in the background with one click for neural filter training and the running process.

Adaptive Document Deduplication. Repetitive documents in the pretraining corpus would harm the LLM's generalization ability in various downstream tasks. Massive deduplication among documents has a theoretical time complexity of $O(n^2)$. The Locally Sensitive Hash algorithm approximates document similarity and reduces the time complexity, but it comes at the cost of increasing memory requirements to store hash collisions. Large-scale fuzzy deduplication becomes infeasible with limited resources.

Corpus	Size	Human Rating	Knowledge Density	PPL in Wikipedia
WuDaoCorpus2.0-200G	193 GB	75%	7.11%	875.41
Oasis-Corpus-zh (with Debias Neural Filter)	370 GB	90%	7.20%	922.97
Oasis-Corpus-zh (with Wiki-vs-CC Neural Filter)	$\sim$ 50 GB	90%	7.99%	192.27

Table 1: Comparison of evaluation metrics for different processing approaches on Chinese corpora. We obtain WuDaoCorpus2.0 from [Yuan *et al.*, 2021]. Oasis-Corpus-zh (with Wiki-vs-CC Neural Filter), has a data scale estimated based on the filter ratio.

$$Pr(d_i, d_j | Jaccard(d_i, d_j) = s_(i, j)) = 1 - (1 - s_{i,j}^b)^r \quad (1)$$

To achieve this goal, we reduce the memory requirement of the LSH deduplication algorithm to adapt to customized hardware by adjusting r in the conditional probability formula. The system predicts the maximum r according to the user's configuration in corpus size and memory size. Since a smaller r will lead to a lower collision probability, the system also suggests the running times based on the Jaccard threshold and the expected duplication recall.

Although document-level deduplication could improve the diversity of the cultivated dataset, it could also significantly decrease the quantity. Our Adaptive Document Deduplication module also provides an interface to visualize the duplicated documents in a graph, offering options for users to make trade-offs between the removal rate and quantity.

Holistic Data Assessment. Evaluating LLMs pre-trained on different curated corpora using downstream tasks' performance serves as an oracle for assessing the data value. This post-hoc method is resource-consuming and ineffective. It is urgent to establish a holistic data assessment system to quantify the data quality and support the optimization process of data curation. We achieve this goal through two views: local quality and global distribution, employing three evaluation methods: human assessment, heuristic metrics, and GPT-4.

Local Quality Evaluation. In this module, we focus on a document's fluency, readability, and coherence as assessed by humans or GPT-4. Due to the high consumption of the human inspection process, we only provide two quality options, *High* and *Low*, in the user-friendly human evaluation interface. It displays real-time statistics of manually labeled quality conditions. State-of-the-art LLMs like GPT-4 have demonstrated sufficient ability to score a document in multiple aspects, reflecting overall quality [Chen *et al.*, 2023]. We provide predefined prompts for quality assessment, achieving more than 95% consistency with human opinions. The system also supports customized prompts for diverse demands. Moreover, the local quality evaluation samples can be incorporated into quality classification datasets to evolve the neural filter.

Global Distribution Assessment. Apart from the local document perspective, the global view of the corpus in statistical distribution can also reflect the broadly defined quality.

Oasis adopts six metrics to assess the corpus in heuristics from a randomly sampled subset of data: (1) **Lexical Diversity Distribution** [McCarthy and Jarvis, 2010]: We calculate each document's Measure of Textual Lexical Diversity (MTLD) score to reflect lexical diversity and plot the frequency histogram to obtain an overall perspective. (2) **Task2Vec Diversity Coefficient** [Lee *et al.*, 2023]: The task2vec diversity coefficient is proven to have a high correlation with humans' intuitive diversity of the corpus. We sample batches of text and display the calculated overall score. (3) **Semantic Diversity Distribution**: We obtain all sampled documents' global semantic vectors using BERT and calculate the cosine similarity of each pair of documents to plot the frequency histogram. (4) **Topic Diversity Distribution**: We cluster the sampled documents by global vector and calculate the similarity of centroid vectors among clusters to reflect overall topic diversity. (5) **Knowledge Density and Diversity**: We inspect the knowledge view of the corpus by counting the different entities that occur. The density means the entities count normalized by word count, and diversity means the semantic similarity of all emerged entities. (6) **Similarity to Wikipedia Distribution**: [Jansen *et al.*, 2022] shows that the Kenlm model's perplexity on the target source could reflect the approximation of the Kenlm model's training source. We train a Kenlm model on Wikipedia and plot the perplexity distribution to inspect the extent of corpus bias in Wikipedia.

These metrics can be displayed on a single page and overlay multiple corpora for convenient visual comparison.

2.2 Comparative Analysis

As shown in Table 1, the human-evaluated quality of the Chinese portion in the Oasis Corpus constructed by the Oasis system surpasses that of Wudao. Additionally, it exhibits a larger scale and greater knowledge diversity, demonstrating the advantage of Oasis, a comprehensive construction and evaluation system, over traditional data construction pipelines in pretraining data construction.

Compared to the corpora obtained by traditional positive-centric neural filters, the debias neural filter can produce comparable quality in human evaluation and a larger quantity. The perplexities in the Wikipedia source also indicate that our neural filter could alleviate the bias toward high-quality sources in the corpus, ensuring diversity.

3 Conclusion

We propose Oasis, a one-stop system for LLM's pretraining data curation and assessment. In customized data curation, users can tailor their pipeline according to specific corpus requirements and limited hardware resources in rule filter, neural filter, and document deduplication. In holistic data assessment, a corpus can be evaluated from two perspectives: local document and global distribution; and in three ways: human assessment, GPT-4 evaluation, and heuristic metrics. These two components collaborate to enhance the value of the LLM's pretraining corpus. The comparative analysis of the constructed corpora demonstrates the effectiveness of Oasis.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2022ZD0160503), the National Natural Science Foundation of China (No. 62176257). This work is also supported by the Youth Innovation Promotion Association CAS.

References

[Abadji *et al.*, 2022] Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*, 2022.

[Anil *et al.*, 2023] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[Chen *et al.*, 2023] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

[Dodge *et al.*, 2021] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

[Fries *et al.*, 2022] Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. Bigbio: a framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806, 2022.

[Gao *et al.*, 2020] 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.

[Gunasekar *et al.*, 2023] 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.

[Jansen *et al.*, 2022] Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*, 2022.

[Kaplan *et al.*, 2020] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[Kreutzer *et al.*, 2022] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

[Laurençon *et al.*, 2022] 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, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826, 2022.

[Lee *et al.*, 2023] Alycia Lee, Brando Miranda, and Sanmi Koyejo. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data. *arXiv preprint arXiv:2306.13840*, 2023.

[Longpre *et al.*, 2023] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

[McCarthy and Jarvis, 2010] Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

[OpenAI, 2023] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[Penedo *et al.*, 2023] 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.

[Rae *et al.*, 2021] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[Rosset, 2020] Corby Rosset. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog*, 1(2), 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wei *et al.*, 2022] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[Wenzek *et al.*, 2019] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet:

Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

[Yuan *et al.*, 2021] 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.