

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

Tong Zhou¹, Yubo Chen^{1,2}, Pengfei Cao^{1,2}, Kang Liu^{1,2}, Jun Zhao^{1,2}, Shengping Liu³

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 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^{1,2} – 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. We exhibit a complete process to use Oasis for the curation and assessment of pre-training data. In addition, an 800GB bilingual corpus curated by Oasis is publicly released².

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,

¹Project:<https://github.com/tongzhou21/Oasis>

²Video:<https://youtu.be/YLfMlnrUZPk>

³Corpus: <https://huggingface.co/datasets/Oasis-Team/Oasis-Corpus>

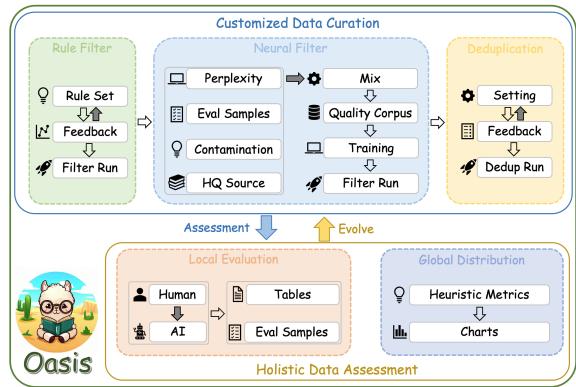


Figure 1: Overview of Oasis functionality.

high-quality, and diverse pretraining corpus. The importance of data curation and assessment is increasingly evident.

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 pre-training 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 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 Related Work

2.1 Data Cultivation

The quantity (Hoffmann et al., 2022) and quality (Gunasekar et al., 2023) of the pretraining corpus guaranteed LLM’s performance in downstream tasks. State-of-the-art data cultivation methods can be classified into rule filter, neural filter, and deduplication.

Rule filter: (Penedo et al., 2023; Laurençon et al., 2022; Abadji et al., 2022) treat too low lan-

guage identification confidence as a first criterion to drop the document. Moreover, some heuristic rules (Sun et al., 2021; Rae et al., 2021; Penedo et al., 2023) focus on document length, punctuation ratio, word length, and stop words, which are widely used in deciding a document’s quality. (Laurençon et al., 2022) also consider the closed class words ratio to distinguish machine-generated text. Statistical language models like Kenlm (Heafield, 2011) are useful tools for efficiently estimating the coherence and fluency of sentences (Wenzek et al., 2019; Laurençon et al., 2022; Wei et al., 2023). In removing undesirable information, (Gu et al., 2023; Wu et al., 2021) build a word list to match and drop documents. (Rae et al., 2021; Wei et al., 2023) utilize a URL block list to discard target web pages. While these methods could lead to bias, (Penedo et al., 2023) optimize the block list by carefully reweighting these URLs. However, the process of rule pipeline construction in a diverse customized corpus lacks attention.

Neural filter: Although the neural filter is more time-consuming than the rule filter, it can explore patterns between high- and low-quality data that cannot be literally concluded (Brown et al., 2020). The training dataset utilizes well-known high-quality sources like Wikipedia, WebText (Radford et al., 2019), and Books as positive samples, meanwhile extensive various web pages as negative samples. A neural model like fastText or BERT trained on this dataset is responsible for scoring documents in quality (Touvron et al., 2023; Brown et al., 2020; Gao et al., 2020). (Wu et al., 2021) also consider utilizing a model to classify advertisements. However, the neural filter could bias the filtered corpus due to the positive source of the training set (Dodge et al., 2021; Welbl et al., 2021). Some works (Du et al., 2022; Wei et al., 2023) organize the positive sample in a mixture of various sources of high-quality texts to decrease the bias from the positive source. (Penedo et al., 2023) abandoned the neural filter on account of worrying about undesirable biases.

Deduplication: Repetition contents in pretraining corpus are proven to hurt the LLM’s performance (Lee et al., 2021). Corpus cultivation pipelines focus on fuzzy deduplication at the document (Zhang et al., 2022; Biderman et al., 2023; Rae et al., 2021) or line (Touvron et al., 2023) levels. These large-scale deduplication processes are mainly based on the locally sensitive hash algo-

rithm (Rajaraman and Ullman, 2011) by means of collision to calculate similarity. (Sun et al., 2021) calculate the MD5 of the three longest sentences to match the redundancy documents. (Penedo et al., 2023) further construct a huge prefix array to drop duplicate substrings. These methods significantly improve the efficiency of the deduplication process, but the memory requirements become a barrier to deployment on a larger scale.

2.2 Data Assessment

Researchers have no consensus about the approaches in pretraining corpus assessment. (Gao et al., 2020) visualize the various components of The Pile and utilize GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) to explore perplexity distribution and topic diversity. They also show the score distribution from the neural filter and the inspection of equality problems. (Kreutzer et al., 2022) horizontally compare multiple corpora in linguistic correctness by human labeling. (Lucioni and Viviano, 2021) focus on offensive content in the high- and low-quality scopes. (Dodge et al., 2021) explored the topic distribution of documents filtered by bad word lists and discovered harmless clusters like medicine and religion. (Marone and Van Durme, 2023) propose a more efficient algorithm for assessing data contamination in downstream tasks. (Laurençon et al., 2022) emphasize the difference among various languages, including the filtered ratio frequency distribution by different methods. (Piktus et al., 2023a,b) build a tool to search strings in the entire corpus efficiently, providing the foundation for various exploration, like detecting personal identity information, inspecting undesired content, and fact verification. There is still a lack of a holistic, multi-dimension, easy-to-use data assessment system.

3 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.

3.1 Customized Data Curation

3.1.1 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.

3.1.2 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, replac-

ing, 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.

3.1.3 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.

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

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.

3.2 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.

3.2.1 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 (SOTA) 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.

3.2.2 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)

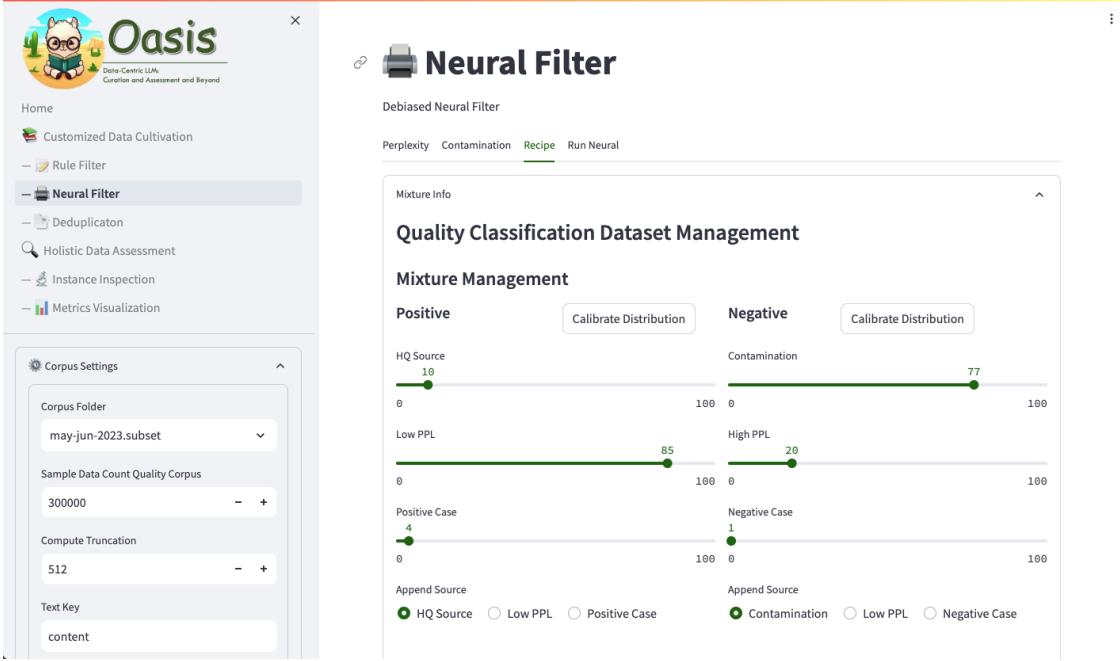


Figure 2: Screenshot of the recipe management interface of Oasis.

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.

4 Usage Examples and Experiments

In this section, we provide an example of how to interact with Oasis in data curation and assessment. We use the newest dump of Common Crawl (May/Jun 2023) as an illustration, focusing on English content.

4.1 Customized Data Cultivation

After the language identification and target language extraction pipeline (Abadji et al., 2021) from WET files, we obtained a 2.4TB raw English dataset with meta information.

Rule Filter: Select the raw dataset in the Interactive Modular Rule Filter module and load the predefined rule pipeline. We can observe the hit rate for each rule cell and a random hit case after clicking the rule cell in the "Build Pipeline" panel. Based on the case shown in the "Case Study" panel, an undesired advertising span is observed, inserted in a coherent sentence. Add a rule cell by setting arguments with the target span and clicking the "remove span" button in the left sidebar. Move up this cell before the last "min word count" cell. After saving the customized pipeline, you can find and load this pipeline in the configuration of the "Run Pipeline" panel and generate a runnable Python script for background multiprocessing running. After applying this rule filter pipeline, we obtain 112GB of data.

Neural Filter: In the "Perplexity" panel, select

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)	~ 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.

a Kenlm model trained on Wikipedia and calculate perplexity to determine a quantile split between normal quality and extremely low-quality content. Drag the slider to change the quantile, inspect the cases, and finally decide on 0.85 as the boundary. Then, adjust the contamination set in the "Contamination" panel, following a logic similar to building a rule filter pipeline. In the "Recipe" section, manage the constitution of the quality classification dataset, both in positive and negative, build the dataset, and train a finetuned BERT model. Select the best checkpoint to run the neural filter in the "Run Neural" panel. Both the training and running processes occur in the background and do not affect other operations. After applying the neural filter, 100GB of high-quality data is obtained.

Document Deduplication: Utilize the duplication cluster graph to visualize the repeated pairs with different Jaccard thresholds in the "Dedup Case" panel. After a few trials, determine the Jaccard threshold as 0.8. In the "Run Dedup" panel, based on the corpus size and available memory, the system generates recommendations for parameter settings and can run document deduplication, utilizing multiple CPU cores in the background. The deduplication process finally removed 5% of the documents.

4.2 Holistic Data Assessment

Instance Inspection: We aim to compare the quality of the corpus in its raw, rule-filtered, and neural-filtered states. First, select the four corpora and manually inspect each with 50 samples in the "Human Rating" panel. Set the default quality to high and click "low" only when a sample is not qualified for LLM training. The quality statistics are displayed in real-time in the sidebar. Then, in the "LLM Evaluation" panel, enter the API key for OpenAI and set 200 samples for each corpus to be evaluated, considering the cost. After receiving feedback for all the requests, check the average score of GPT-4. Conclusively, the document quality improves gradually as the build progresses.

Heuristic Metrics: In the "Heuristic Calculation" panel, select all heuristic metrics and choose multi-corpus for calculation. After obtaining the results for these metrics, pick the files to visualize in the "Report" panel. These charts demonstrate that our filter pipeline loses some diversity but substantially increases the quality. Our negative-centric dataset-building method introduced fewer biases than the previous wiki-vs-cc neural filters, achieving better lexical and topic diversity.

4.3 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 Wu-dao. 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.

5 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 pre-training corpus. The comparative analysis of the constructed corpora demonstrates the effectiveness of Oasis.

6 Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science Foundation of China (No. 61976211, 62176257). This work is also supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA27020100), the Youth Innovation Promotion Association CAS, and Yunnan Provincial Major Science and Technology Special Plan Projects (No.202202AD080004)

References

- Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *CMLC 2021-9th Workshop on Challenges in the Management of Large Corpora*.
- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.
- Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.
- Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio: a framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Lei Liu, Xiaoyan Zhu, et al. 2023. Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *Machine Intelligence Research*, 20(2):207–219.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojgan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

- 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. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.
- Alycia Lee, Brando Miranda, and Sanmi Koyejo. 2023. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data. *arXiv preprint arXiv:2306.13840*.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.
- Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023. A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.
- Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What’s in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*.
- Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data. *arXiv preprint arXiv:2303.03919*.
- Philip M McCarthy and Scott Jarvis. 2010. Mtld, vcd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Alexandra Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023a. The roots search tool: Data transparency for llms. *arXiv preprint arXiv:2302.14035*.
- Aleksandra Piktus, Odunayo Ogundepo, Christopher Akiki, Akintunde Oladipo, Xinyu Zhang, Hailey Schoelkopf, Stella Biderman, Martin Potthast, and Jimmy Lin. 2023b. Gaia search: Hugging face and pyserini interoperability for nlp training data exploration. *arXiv preprint arXiv:2306.01481*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susanah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.
- Corby Rosset. 2020. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog*, 1(2).
- Yu Sun, Shuhuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.
- Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.
- Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jianguang Luo, Liang Xu, et al. 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arXiv preprint arXiv:2110.04725*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher De-wan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.