Perplexed by Perplexity: Perplexity-Based Pruning With Small Reference Models

Zachary Ankner [1,2] **Cody Blakeney**[1] **Kartik Sreenivasan**[1]
Max Marion[1] **Matthew L. Leavitt**[3] **Mansheej Paul**[1]
[1]MosaicML [2]MIT [3]DatologyAI

Abstract

In this work, we consider whether pretraining on a pruned high-quality subset of a large-scale text dataset can improve LLM performance. While existing work has shown that pruning based on the perplexity of a larger model can yield high-quality data, we investigate whether smaller models can be used for perplexity-based pruning and how pruning is affected by the domain composition of the data being pruned. We demonstrate that for multiple dataset compositions, perplexity-based pruning of pretraining data can *significantly* improve downstream task performance: pruning based on perplexities computed with a 125 million parameter model improves the average accuracy on downstream tasks of a 3 billion parameter model by up to 1.35% and achieves up to a $1.36\times$ reduction in pretraining steps to reach commensurate baseline performance.

1 Introduction

Recently, the machine learning community has focused on improving the performance of large language models (LLMs) while reducing their training costs. We consider how to improve the quality of an LLM by improving the quality of the pretraining data. Although there are many techniques to improve data quality, such as augmenting training samples with additional information (Li et al., 2024; Korbak et al., 2023), in this work we focus on the predominant method of *data pruning*: intelligently selecting a high-quality subset of a larger dataset to train on.

Data pruning is commonly employed as a tool for quality filtering noisy text data. Simple approaches include using symbolic rules (Bane et al., 2022; Raffel et al., 2020) or using simple classifiers to determine high-quality samples (Wenzek et al., 2020). In addition to basic quality filtering, more complex data pruning techniques are also applied to datasets to *further* improve their quality. Xie et al. (2023b) perform importance resampling where importance scores are calculated based on feature similarity to a target text to sample high-quality texts. Tirumala et al. (2023) deduplicate data based on a pretrained language model's embeddings. Xie et al. (2023a) re-weight domain proportions based on learnability as determined by a smaller proxy model. Marion et al. (2023) investigate data pruning based on multiple neural heuristics of sample difficulty, ultimately concluding that the perplexity of a sample under a reference language model is the best pruning metric.

In this work, we empirically investigate the impact that data pruning based on sample perplexity (Marion et al., 2023) has on LLM pretraining. In particular, we focus on the interplay between pretraining dataset composition and pruning methodology. We also investigate the impact of evaluating based on test-set perplexity vs. downstream task performance. To perform perplexity-based data pruning, we train a language model on a random subset of the given pretraining dataset and then evaluate its perplexity on each sample in the dataset. We then prune the dataset to only include samples within some range of perplexities (i.e., sub-sample to the highest or lowest perplexity samples). We demonstrate that for two vastly different pretraining data compositions, a small language model can be used to effectively prune the pretraining dataset of a significantly larger model leading to significant gains in the final model's downstream performance.

Our work differs from previous work on perplexity-based data pruning for LLM pretraining in two key ways: (i) our emphasis on downstream model quality evaluation and (ii) our exploration of different pretraining dataset domain compositions. While previous works evaluate the resulting LLM's quality based on upstream metrics such as perplexity on the test split of the pretraining dataset, we

Correspondence to ankner@mit.edu.

evaluate data pruning's impact based on downstream evaluation benchmarks (e.g. *mmlu* (Hendrycks et al., 2021), *hellaswag*(Zellers et al., 2019), etc.). Evaluating on more meaningful benchmarks enables us to make stronger, more rigorous conclusions about the impact of perplexity-based data pruning, as some techniques that significantly improve downstream performance have no, or even adverse, effects on upstream performance. Additionally, while previous works only investigate pruning on datasets composed of a singular domain (CommonCrawl[0]), we consider two datasets with different domain compositions: (i) the Pile (Gao et al., 2020), which is composed of many diverse curated domains and only 15.61% of the data is derived from CommonCrawl and (ii) a custom web-scrape heavy dataset where 87.7% of the data is derived from CommonCrawl. We find that successful pruning techniques vary greatly for different dataset compositions to the point that the best technique for one dataset composition may degrade performance for a different composition.

Contributions Our work makes the following contributions:

- We demonstrate that a smaller reference model can successfully prune the pretraining dataset of a significantly larger language model ($30\times$ greater parameters), providing both a significant increase in downstream performance and decrease in pretraining steps, and find that this result holds in two datasets of vastly different domain composition: one heavily curated, one primarily web-scraped (Table 1 and Figure 1).
- We show that data pruning techniques can be highly sensitive to the domain composition of the dataset, suggesting the need to evaluate multiple distinct dataset compositions when conducting data pruning research (Table 1 and table 3).
- We find that test set perplexity can be a misleading metric for evaluating the efficacy of data pruning techniques, as interventions that result in significantly higher perplexity can nevertheless achieve better performance on downstream tasks (Table 2).

2 PERPLEXITY-BASED DATA PRUNING

We start by training a *reference model* that will be used to calculate the perplexity of all samples in our dataset. First, we partition the original dataset into two splits: one for training the reference model and one for training the *final model*. After training the reference model on the standard next-token prediction objective, we compute the reference model's perplexity on each of the samples in the final model's training split. We then prune the final model's dataset split to a fraction of its original size, referred to as the *selection rate* (r_s), by selecting samples according to a *selection criteria* which can be one of low, medium, or high. In low selection, samples with the lowest perplexity are selected. In medium selection, we select samples whose perplexity is close to the median perplexity, i.e., samples with perplexity in the $[50-\frac{r_s}{2}, 50+\frac{r_s}{2}]$ percentiles of all perplexities. In high selection, the samples with the highest perplexity are selected. After pruning our dataset, we train a final model using the standard next token prediction objective on the pruned version of the final model training split. We present a pseudo-code for pruning based on perplexity in algorithm 1.

We consider the setting where the reference model is significantly smaller than the final model. While this assumption is not strictly necessary, we believe that it is the most practically relevant setup, as it best reflects a data pruning paradigm that would be used for the next generation of LLMs where the models being trained are larger than any existing models.

3 EXPERIMENTS

3.1 SETUP

Models. All models are based on the MPT family of transformer models (Vaswani et al., 2017; MosaicML, 2023c). All reference models have 125 million parameters, and we consider final models with 1 billion and 3 billion parameters.

Data. We consider two datasets in this work. The Pile (Gao et al., 2020) is composed of 22 different domains that range from general web scrapes to legal text. For the second dataset, we propose a custom domain mix that we name the Web Dense Dataset (WDD). WDD represents a realistic web-scrape skewed dataset composition for training LLMs; while there are code and math datasets present in the mix, over 87.7% of the dataset is from web scrapes. We provide more details on WDD in section 6. We tokenize all datasets using the GPT-4 tokenizer (OpenAI, 2022).

[0]https://data.commoncrawl.org/

Table 1: Average task accuracy grouped by task category for both datasets and both final model sizes. For all datasets and model sizes we find that training on perplexity pruned data outperforms the baseline. Bold results are within one standard error of the highest accuracy.

Pruning Method	World Knowl-edge	Common Sense Rea-soning	Language Under-stand-ing	Symbolic Prob-lem Solving	Reading Com-prehen-sion	Average
1B Parameters Trained on Pile						
No Pruning (Baseline)	26.32	48.13	52.07	**8.56**	**29.94**	33.00
High Perplexity Selected	**28.87**	**49.71**	**55.45**	**8.41**	**29.61**	**34.41**
3B Parameters Trained on Pile						
No Pruning (Baseline)	32.17	49.88	59.23	**9.93**	**33.06**	36.86
High Perplexity Selected	**35.84**	**51.47**	**62.07**	8.06	**33.63**	**38.21**
1B Parameters Trained on WDD						
No Pruning (Baseline)	29.8	**49.68**	55.97	**8.59**	32.02	35.21
Medium Perplexity Selected	**30.88**	**50.82**	**57.83**	**8.86**	31.83	**36.04**
3B Parameters Trained on WDD						
No Pruning (Baseline)	36.5	51.75	64.15	**9.4**	**34.17**	39.19
Medium Perplexity Selected	**37.25**	**52.83**	**65.37**	**9.43**	**34.78**	**39.93**

Training and evaluation. All reference models are trained for a fixed duration of 26 billion tokens, and all final models are trained to Chinchilla optimal (Hoffmann et al., 2022), meaning that each final model's training duration in tokens is 20 times its parameter count. Training is conducted using llm-foundry (MosaicML, 2023b) and using both Nvidia `A100s` and `H100s`. We evaluate models on 33 different downstream question-answering tasks using the MosaicML evaluation gauntlet (MosaicML, 2023a). We report the average accuracy for each task category as well as the average accuracy across all task categories. More details on tasks and task categories are in section 11.

3.2 PERPLEXITY-BASED DATA PRUNING IMPROVES DOWNSTREAM PERFORMANCE

If a certain range of perplexities is a good heuristic for data quality, training on that perplexity-pruned subset should improve downstream performance. We sweep across pruning selection criteria and selection rates (section 8) and find that the best settings are to select high-perplexity samples at a 50% rate for the Pile and to select medium-perplexity samples at a 50% rate for WDD. The difference in optimal selection criteria (i.e. high vs. medium) for different datasets demonstrates the need to evaluate data pruning techniques across differing dataset compositions. We compare the most performant pruning settings to baseline models trained on the original datasets without pruning in table 1. Across all datasets and model sizes, models pretrained on the perplexity pruned version of the dataset significantly outperform the baseline model on average. This result suggests that the perplexity of a small model provides a strong signal of data quality for a much larger model.

3.3 PERPLEXITY-BASED DATA PRUNING IMPROVES TRAINING EFFICIENCY

We investigate how perplexity-based data pruning affects the training duration required to reach a certain downstream accuracy. We plot the average downstream performance of partially trained checkpoints from the baseline and perplexity pruned models in fig. 1. Perplexity pruning outperforms the baseline model on all intermediate pretraining durations evaluated. Furthermore, perplexity pruned models reach the same average accuracy as baseline models in $1.35\times$, $1.36\times$, $1.27\times$, and $1.15\times$ fewer steps for Pile 1B, Pile 3B, WDD 1B, and WDD 3B, respectively. This result demonstrates that the resulting high-quality data from perplexity-based pruning enables faster learning that can be leveraged to improve the pretraining efficiency of LLMs.

3.4 UPSTREAM PERPLEXITY IS NOT A RELIABLE EVALUATION METRIC FOR DATA PRUNING

As previous works have used the perplexity of the model on a test split of the pretraining dataset as an approximation to downstream performance, we wanted to explore whether such perplexity-based evaluations should be used to quantify the performance of data pruning techniques. Pruning per-

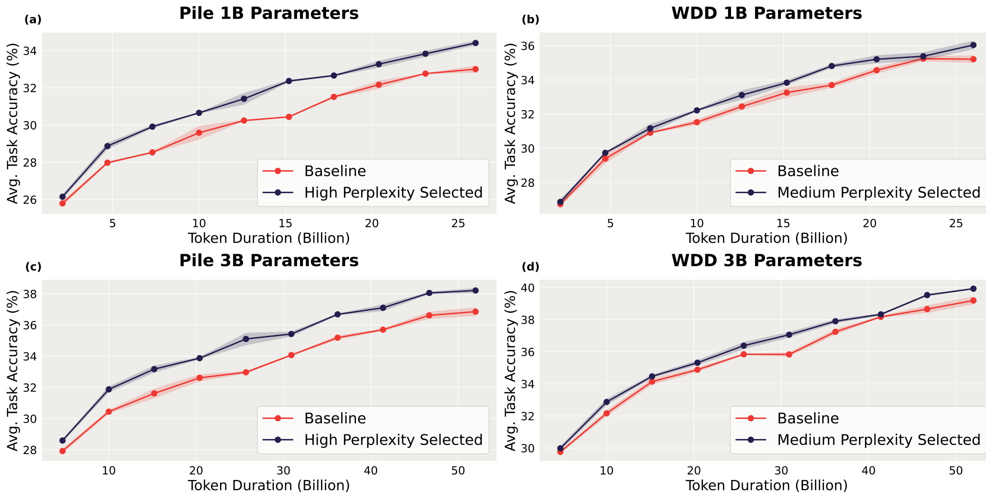

Figure 1: Average task accuracy evaluated intermittently throughout pretraining for each dataset and model size investigated. Model based perplexity filtering leads to a significant improvement in performance and speed of learning across model scales and pretraining datasets.

Table 2: Performance as evaluated by perplexity on a test split of the original dataset as well as average task accuracy for 1 billion parameter final models trained on the Pile. The model trained on pruned data has worse pretraining test split perplexity even though it significantly improves average downstream accuracy.

Pruning Method	Test Set Pplx. (↓)	Downstream Task Avg. (↑)
No Pruning (Baseline)	**7.83**	33.00
High Perplexity Selected	8.51	**34.41**

forms an intervention on the dataset, making models trained on the pruned dataset biased estimators of the original data distribution. Therefore, it is unlikely that the performance on the original data distribution is a fair evaluation of model quality. We compare the test set perplexity and average downstream task accuracy for 1 billion parameter models trained on the original and pruned version of the Pile in table 2. We find that while high-perplexity selection significantly improves average downstream accuracy, it significantly degrades test set perplexity. This result suggests that test set perplexity may not always be a sound metric for data pruning work and that researchers should instead directly use downstream benchmarks.

4 Discussion

In this work, we conduct an empirical investigation of the impact that perplexity-based data pruning has on model performance. We demonstrate that small reference models can be used to prune the data of models with up to *30×* more parameters, leading to both significant downstream performance increases and increased training efficiency. We investigate upstream metrics for evaluating data pruning techniques and provide an example where evaluating the performance of data pruning techniques based on test set perplexity does not align with downstream model performance. Additionally, we demonstrate that optimal pruning techniques can vary greatly for differening dataset compositons. Although we do not present a predictive theory for how pruning parameters should be selected for different datasets, we demonstrate that the optimal pruning parameters for a 1 billion parameter model transfer to 3 billion parameter models, potentially suggesting that empirically determining the optimal pruning parameters can be done cheaply. Our work takes a step towards establishing perplexity-based data pruning as a key technique in the modern data researcher's toolkit.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL `https://aclanthology.org/N19-1245`.

Jean-michel Attendu and Jean-philippe Corbeil. NLU on data diets: Dynamic data subset selection for NLP classification tasks. In Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal (eds.), *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 129–146, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sustainlp-1.9. URL `https://aclanthology.org/2023.sustainlp-1.9`.

Amittai Axelrod. Cynical selection of language model training data. *arXiv preprint arXiv:1709.02279*, 2017.

Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. A comparison of data filtering methods for neural machine translation. In Janice Campbell, Stephen Larocca, Jay Marciano, Konstantin Savenkov, and Alex Yanishevsky (eds.), *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pp. 313–325, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2022.amta-upg.22`.

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. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Rui Castro and Robert Nowak. Active learning and sampling. In *Foundations and Applications of Sensor Management*, pp. 177–200. Springer, 2008.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL `https://aclanthology.org/N19-1300`.

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. *arXiv:1803.05457v1*, 2018.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJg2b0VYDr`.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen

Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/du22c.html.

Yukun Feng, Patrick Xia, Benjamin Van Durme, and João Sedoc. Automatic document selection for efficient encoder pretraining. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9522–9530, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.647. URL https://aclanthology.org/2022.emnlp-main.647.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024.

Leo Gao. An empirical exploration in quality filtering of text data, 2021.

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.

Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. *arXiv preprint arXiv:2404.07177*, 2024.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pp. 181–195. Springer, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Aditi Jha, Sam Havens, Jeremy Dohmann, Alexander Trott, and Jacob Portes. LIMIT: Less is more for instruction tuning across evaluation paradigms. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=QxtL4Q1enz.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17506–17533. PMLR, 2023. URL `https://proceedings.mlr.press/v202/korbak23a.html`.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=UoEw6KigkUn`.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer, 2012.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=1oijHJBRsT`.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 3622–3628. ijcai.org, 2020. doi: 10.24963/IJCAI.2020/501. URL `https://doi.org/10.24963/ijcai.2020/501`.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining LLMs at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL `https://openreview.net/forum?id=XUIYn3jo5T`.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. Smaller language models are capable of selecting instruction-tuning training data for larger language models. *arXiv preprint arXiv:2402.10430*, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018. URL `https://api.semanticscholar.org/CorpusID:52183757`.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15630–15649. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/mindermann22a.html`.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre (eds.), *Proceedings of the ACL 2010 Conference Short Papers*, pp. 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://aclanthology.org/P10-2041`.

MosaicML. Llm evaluation scores, 2023a. URL `https://www.mosaicml.com/llm-evaluation`.

MosaicML. Llm foundry. `https://github.com/mosaicml/llm-foundry`, 2023b.

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

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=j5BuTrEj35`.

OpenAI. Tiktoken: A fast bpe tokeniser for use with openai's models. `https://github.com/openai/tiktoken`, 2022.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL `https://aclanthology.org/P16-1144`.

Dongmin Park, Dimitris Papailiopoulos, and Kangwook Lee. Active learning is a strong baseline for data subset selection. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022.

Dongmin Park, Seola Choi, Doyoung Kim, Hwanjun Song, and Jae-Gil Lee. Robust data pruning under label noise via maximizing re-labeling accuracy. *arXiv preprint arXiv:2311.01002*, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=kM5eGcdCzq`.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol

Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis and insights from training gopher, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI, 2011. URL `http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2418`.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, pp. 8732–8740, 2020. URL `https://aaai.org/ojs/index.php/AAAI/article/view/6399`.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=CG0L2PFrb1.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Thom Wolfe, Lewis Tunstall, and Patrick von Platen. Jeopardy dataset on hugging face hub. https://huggingface.co/datasets/jeopardy, 2022.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=lXuByUeHhd.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=uPSQv0leAu.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation. *arXiv preprint arXiv:2104.06245*, 2021.

5 Related Work

Classical methods for pruning text data. In order to improve the quality of raw web scrapes, which often contain very noisy samples, pruning via quality filtering has become a common practice. Simple rules-based methods have been employed to prune datasets by filtering out low-quality samples according to some hand-crafted heuristic such as whether the text contains prohibited words, is predominantly English, etc. (Bane et al., 2022; Raffel et al., 2020; Rae et al., 2022; Penedo et al., 2023). N-gram perplexity-based methods, in which an n-gram model is first trained on a high quality, curated corpus and then used to score another corpus, have also been applied to filter text data (Moore & Lewis, 2010; Axelrod, 2017; Gao, 2021; Laurençon et al., 2022; Muennighoff et al., 2023). Although our method also uses perplexity to prune data, it does so in a very different manner. In n-gram perplexity pruning, perplexity is used to estimate whether new text is in distribution as compared to the currated text the n-gram was trained on, while in our model-based perplexity pruning, the reference model is trained on the same distribution of text and the perplexity is more akin to an estimate of the difficulty of an example.

In this work, the datasets we leverage already have some basic rules-based pruning applied, and as such, the method we investigate is largely complementary to these existing techniques.

Neural network based methods for pruning text data. Recently, there has been much interest in using neural networks to compute metrics that can be used to intelligently prune datasets. A common technique in this family of methods is using a model to sample high quality data from large datasets based on the sample's similarity to a curated high quality corpus that serves as a target distribution (Feng et al., 2022; Xie et al., 2023b). Xie et al. (2023a) also consider how to use a small reference model to prune pretraining data for a much larger model. They do so by using a small reference model to learn the optimal weighting of domain proportions to maximize the "learnability" of the resulting dataset. While their work prunes data at the granularity of the dataset's domain composition, our work explores a sample-wise method for pruning text datasets. Pruning based on the difficulty or loss of a sample has previously been explored for text data, but the majority of such work focuses on curating data for finetuning (Swayamdipta et al., 2020; Attendu & Corbeil, 2023; Coleman et al., 2020; Mindermann et al., 2022; Mekala et al., 2024). Marion et al. (2023), however, investigate multiple model-based sample difficulty heuristics for pruning pretraining text datasets. Their work provides important analysis into different metrics for text pruning determining that perplexity is the best scoring function. Although we use the same method for pruning text pretraining datasets, our analysis differs substantially as we evaluate model quality based on downstream metrics and extend our analysis to multiple different dataset compositions. The difference in analysis enables us to determine that a smaller reference model can prune the dataset of a substantially larger model, while Marion et al. (2023) conclude that a reference model substantially larger than the final model is required to see performance improvements.

Data pruning on vision tasks. While data pruning is becoming more and more relevant with large amounts of text data, it has also been extensively applied in the vision domain (Paul et al., 2021; Toneva et al., 2018; Park et al., 2023). These works often prune data points based on their loss or gradients during training (Killamsetty et al., 2021; Mirzasoleiman et al., 2020). Some methods instead rely on the geometry of the decision boundary in the feature space (Chen et al., 2012; Sener & Savarese, 2017). Note that in the literature, data pruning is also sometimes referred to as coreset selection (Guo et al., 2022). More recently, Park et al. (2022) show that, somewhat surprisingly, active learning (Castro & Nowak, 2008) based algorithms tend to outperform most data subset selection algorithms. In the context of contrastive learning, hard-negative mining has been effective as a data pruning method (Kalantidis et al., 2020; Robinson et al., 2020; Zhang & Stratos, 2021). Recently, Goyal et al. (2024) investigated scaling laws for training on filtered data in the context of vision models.

6 *Web Dense Dataset*

The Pile (Gao et al., 2020) dataset is one of the largest and most diverse open pretraining datasets. Therefore, it has become one of the cornerstones of research in pretraining LLMs (Biderman et al., 2023) and pretraining data selection (Xie et al., 2023b;a). While the Pile does contain general

Algorithm 1: Psuedo code for performing perplexity-based data pruning.

Input: Raw dataset $D = \{x^{(i)}\}_{i=1}^{M}$, where each $x^{(i)}$ is a tokenized text sample; `selection_criteria` $\in$ {low, medium, high}; selection rate $r_s \in (0, 1)$; reference training split size R.
Output: Parameters of model trained on the perplexity pruned dataset θ^*_{final}.

$\theta_{\text{ref}} \leftarrow$ random parameter initialization
$D_{\text{ref}}, D_{\text{train}} \leftarrow \texttt{random\_split}(D, R)$
$\theta^*_{\text{ref}} \leftarrow \texttt{train}(\theta_{\text{ref}}, D_{\text{ref}})$
$\texttt{P} \leftarrow \{\}$
for $x^{(i)} \in D_{\textit{train}}$ **do**
  $\texttt{PPLX}_{x^{(i)}} = \frac{1}{|x^{(i)}|} \sum_{t_j \in x^{(i)}} -\log P(t_j | t_{<j}; \theta_{\text{ref}})$
  $\texttt{P}[x^{(i)}] = \texttt{PPLX}_{x^{(i)}}$
end
$\hat{F}_P \leftarrow$ empirical CDF of `P.values()`
if *`selection_criteria`* == *"low"* **then**
  `min_percentile` $\leftarrow 0.0$
  `max_percentile` $\leftarrow r_s$
end
else if *`selection_criteria`* == *"medium"* **then**
  `min_percentile` $\leftarrow 0.5 - \frac{r_s}{2}$
  `max_percentile` $\leftarrow 0.5 + \frac{r_s}{2}$
end
else if *`selection_criteria`* == *"high"* **then**
  `min_percentile` $\leftarrow 1 - r_s$
  `max_percentile` $\leftarrow 1.0$
end
$D_{\text{pruned}} \leftarrow []$
for $x^{(i)}, \mathit{PPLX}_{x^{(i)}} \in P$ **do**
  if *`min_percentile`* $< \hat{F}_P(\mathit{PPLX}_{x^{(i)}}) <$ *`max_percentile`* **then**
    D_{pruned}`.append(`$x^{(i)}$`)`
  end
end
$\theta_{\text{final}} \leftarrow$ random parameter initialization
$\theta^*_{\text{final}} \leftarrow \texttt{train}(\theta_{\text{final}}, D_{\text{pruned}})$
return $\theta^*_{\textit{final}}$

webscrapes such as the Pile-CommonCrawl and OpenWebText2, a relatively large fraction of its data comes from high-quality but domain-specific datasets such as PubMed Central. This is in contrast to other recent pretrianing data mixes (Du et al., 2022; Touvron et al., 2023; Soldaini et al., 2024) that skew towards large CommonCrawl webscrapes.

One of the central goals of this work is to rigorously investigate how the impact of perplexity-based data pruning and related optimal hyperparameters depends on the distribution of the underlying corpus. Toward this end, we construct a CommonCrawl heavy dataset—*Web Dense Dataset (WDD)*—to compare how our pruning method's performance changes on a dataset with a different domain distribution. WDD has a large fraction of CommonCrawl derived data (87.7%) and is inspired by Dolma (Soldaini et al., 2024) with the addition of a few currated, high-quality domains.

7 Perplexity-Based Data Pruning Methodology

We provide a pseudo code for how perplexity-based data pruning works in algorithm 1.

Table 3: Results from sweeping different selection criteria. We report the average accuracy for each task grouping as well as across all tasks. While high perplexity selection is optimal for the Pile, medium perplexity selection is optimal for WDD. Bold results are within one standard error of the highest accuracy.

Pruning Method	World Knowledge	Common Sense Reasoning	Language Understanding	Symbolic Problem Solving	Reading Comprehension	Average
1B Parameters Trained on Pile						
No Pruning (Baseline)	26.32	48.13	52.07	**8.56**	**29.94**	33.0
Low Perplexity Selected	22.49	45.64	45.44	**8.67**	27.96	30.04
Medium Perplexity Selected	26.86	47.46	52.14	8.41	**29.76**	32.92
High Perplexity Selected	**28.87**	**49.71**	**55.45**	**8.41**	**29.61**	**34.41**
1B Parameters Trained on WDD						
No Pruning (Baseline)	29.8	**49.68**	55.97	**8.59**	32.02	35.21
Low Perplexity Selected	29.17	**49.99**	55.84	**9.01**	**32.88**	35.38
Medium Perplexity Selected	**30.88**	**50.82**	**57.83**	**8.86**	31.83	**36.04**
High Perplexity Selected	29.67	49.36	56.56	**8.79**	30.32	34.94

8 Full Data Pruning Settings Sweep

In this section, we report the results of sweeping over different perplexity-based pruning setting configurations. In particular, for each dataset, we first sweep over the selection criteria to determine where from the distribution of perplexities samples should be selected. Then, using the best selection criteria, we sweep the selection rate to determine how much we should prune.

Setup. We use the same training and evaluation setup as detailed in section 3.1. We only perform the sweep over pruning parameters for 1 billion parameter final models for computational budget reasons; however, we find that the best selection criteria at the 1 billion parameter scale also lead to a performance improvement at the 3 billion parameter scale, as detailed in 3.2.

8.1 Finding the Best Selection Criteria

For each dataset, we first sweep the selection criteria while keeping the selection rate fixed at 50%. We report the performance of each selection criteria in table 3. We find that on the Pile high perplexity selection works the best and on WDD medium perplexity selection works the best, improving the average downstream performance by 1.45% and 0.83% respectively. An important observation from the sweep is that the best selection criteria from one dataset does not transfer to another dataset and may actually degrade performance compared to the baseline. Although high-perplexity selection is the best method on the Pile, selecting high-perplexity samples on WDD leads to a decrease in the average downstream accuracy of 0.27%. Similarly, medium-perplexity selection is the best method on WDD, but selecting medium-perplexity samples on the Pile does not have a statistically significant impact compared to the baseline without pruning. These results inform us that high and medium perplexity selection are the optimal selection criteria for the Pile and WDD respectively, and that the optimal selection criteria does not necessarily transfer between datasets with different domain compositions.

8.2 Finding the Best Selection Rate

Using the optimal selection criteria that we found for each dataset, we next investigate the best selection rate for each dataset. We investigate three different selection rates: 25%, 50%, and 75%. We present the results for each selection rate in table 4. On the Pile, we find that selection rates of 25% and 50% achieve exactly the same average downstream accuracy; on WDD we find that while a selection rate of 50% achieves the best absolute average downstream accuracy, its improvement is not statistically significant compared to a selection rate of 75%. For simplicity, we chose to conduct

Table 4: Results from sweeping different selection rates. We report the average accuracy for each task grouping as well as across all tasks. Bold results are within one standard error of the highest accuracy.

Pruning Method	World Knowl-edge	Common Sense Rea-soning	Language Under-stand-ing	Symbolic Prob-lem Solving	Reading Com-prehen-sion	Average
1B Parameters Trained on Pile						
25% Selection Rate	**28.99**	49.47	**56.11**	**8.63**	**28.83**	**34.41**
50% Selection Rate	**28.87**	**49.71**	55.45	**8.41**	**29.61**	**34.41**
75% Selection Rate	27.86	47.99	54.13	**8.84**	28.57	33.48
1B Parameters Trained on WDD						
25% Selection Rate	30.56	49.6	**57.5**	**9.26**	30.88	35.56
50% Selection Rate	**30.88**	**50.82**	**57.83**	8.86	**31.83**	**36.04**
75% Selection Rate	29.94	**50.81**	**57.52**	8.6	**31.98**	**35.77**

Table 5: Downstream task performance for Chinchilla Optimal and $5\times$ over-trained data budgets. Perplexity-based data pruning outperforms training on the original dataset in both settings, but the improvement over the baseline decreases in the over-trained regime.

Pruning Method	Average	Improvement Over Baseline
1B Parameters Trained on High Perplexity Pile		
Chinchilla Optimal	34.41	1.41
$5\times$ Over-Trained	37.02	1.44

the rest of the experiments in the paper using a selection rate of 50% on both datasets. Furthermore, we find that all the selection rates tested outperform the baseline of no data pruning as measured by average downstream accuracy. This suggests that the selection criteria has a greater impact on the performance of a pruning configuration than the selection rate.

9 Extended Results

In this section we perform an extended analysis of perplexity-based data pruning and evaluate perplexity-based data pruning in two non-standard settings: the over-training and data-constrained regimes.

9.1 perplexity-based Data Pruning for Over-Trained Models

A recent trend with LLMs has been to over-train models by training them on more tokens than the chinchilla optimal number of tokens (Touvron et al., 2023; Gadre et al., 2024). As our work targets the data component of LLM pretraining, we investigate the hypothesis that over-training would be more beneficial for models trained on perplexity filtered datasets as the data is of higher quality. We test this hypothesis by training a 1B parameter model for 130B tokens, which is $5\times$ the chinchilla optimal number of tokens. We evaluate the downstream performance of each over-trained model in table 5. On the Pile, we find that the gain from perplexity pruned data is similar in the compute optimal regime and the over-trained regime: we see a gain in average performance of 1.41% when training compute optimal and a gain of 1.44% when over-training. These results suggests that while the higher quality data resulting from perplexity-based data pruning does still lead to an improvement in downstream performance in the over-trained regime, there is not a relative increase in downstream improvement over the baseline when over-training.

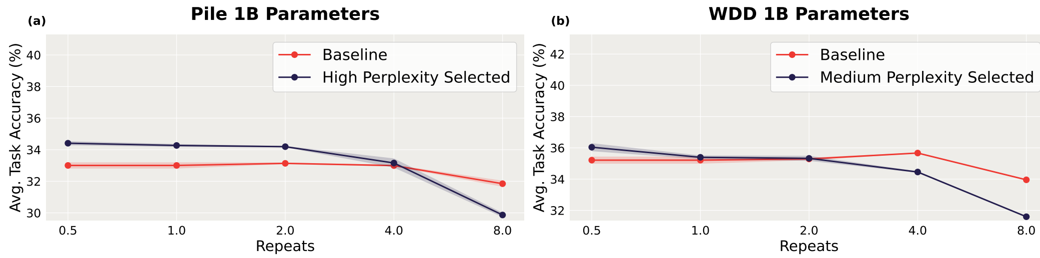

Figure 2: Downstream task performance as a function of available dataset size. The number of repeats denotes the number of repeats over the raw dataset necessary to achieve the Chinchilla optimal number of tokens. Training on perplexity pruned data leads to an improvement for up to four repeats on the Pile and one repeat on WDD.

9.2 PERPLEXITY-BASED DATA PRUNING FOR THE DATA CONSTRAINED REGIME

Our experiments so far were conducted in the setting where there exists a sufficient abundance of data such that after pruning with the desired selection rate there are enough data points such that no data has to be repeated to fill the desired token budget However, there are many training settings which don't fall under this data-abundant regime. Accordingly, we evaluate how perplexity-based data pruning performs when the number of tokens is constrained, and pruning induces a greater number of repetitions of the data. For each dataset we vary the available data such that training for a Chinchilla optimal number of tokens requires a different number of repeats. Specifically, we investigate data budgets requiring $\{0.5, 1, 2, 4, 8\}$ repeats to reach chinchilla optimal[1]. As each number of repeats refers to the total number of available tokens, for all filtering experiments the number of repeats after pruning is actually greater by a factor of $\frac{1}{r_s}$ since we filter the available tokens according to r_s, the selection rate. As all models use a selection rate of $\frac{1}{2}$, the models trained on the pruned data see the data for $2\times$ more repetitions.

We plot the average downstream performance as a function of the number of repetitions in fig. 2. On the Pile, we find that training on perplexity pruned data yields an improvement for up to four repeats. On WDD, training on pruned data yields an improvement for only up to one repeat. These results suggests that while it is possible to observe benefits from perplexity-based data pruning in the data-constrained regime, the the maximum number of repeats for which training on pruned data is optimal can vary greatly from dataset to dataset.

10 UNDERSTANDING THE EFFECTS OF PERPLEXITY-BASED PRUNING

In this section we show the distribution of reference model perplexities on each dataset, as well as detailing how pruning affects the domain composition of the datasets.

10.1 HOW ARE REFERENCE PERPLEXITIES DISTRIBUTED

In order to better understand how perplexity-based data pruning works, we investigate the distribution of the computed reference model perplexities for each dataset. For each dataset, we randomly sample 10% of the calculated perplexities and perform kernel density estimation to estimate the distribution of perplexities for a given dataset. We repeat this procedure for the optimal pruned version of the dataset. We plot the resulting estimates of the perplexity distribution in fig. 3. We find that the perplexity distribution for the Pile is multi-modal and asymmetric, while for WDD it is unimodal and symmetric.

[1] Repeat=0.5 means that the available number of tokens is twice the training budget, i.e. the data-abundant setting

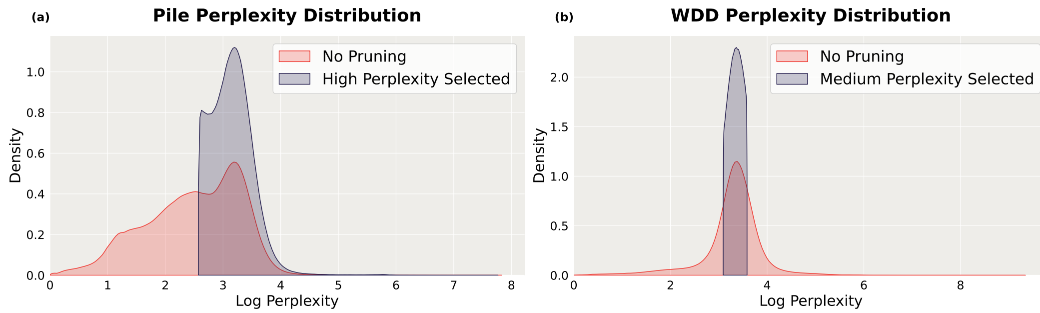

Figure 3: Distribution of sample perplexities as evaluated by the reference model for the Pile and WDD. We show both the original distribution over the full dataset without pruning as well as the distribution after applying the optimal perplexity-based data pruning technique for a given dataset.

11 Detailed Evaluation Setup

Jha et al. (2023) also use the MosaicML evaluation gauntlet to perform evaluations in their work. As such, with explicit permission from the authors, we exactly reproduce their text describing the tasks and tasks categories in the evaluation gauntlet. The following is from Section D of their paper:

The **World Knowledge** category includes the following datasets:

- Jeopardy (2,117 questions that are a custom subset of the dataset originally obtained from Wolfe et al. (2022))
- MMLU (14,042 four-choice multiple choice questions distributed across 57 categories Hendrycks et al. (2021)
- BIG-bench wikidata (20,321 questions regarding factual information pulled from wikipedia) Srivastava et al. (2023)
- ARC easy (2,376 easy multiple choice middle school science questions) Clark et al. (2018)
- ARC challenge (1,172 hard multiple choice science questions) Clark et al. (2018)
- BIG-bench: misconceptions (219 true or false questions regarding common misconceptions) Srivastava et al. (2023)

The **Commonsense Reasoning** category loosely assesses a model's ability to do basic reasoning tasks that require commonsense knowledge of objects, their properties, and their behavior. It includes the following datasets:

- BIG-bench Strategy QA (2,289 very eclectic yes/no questions on a wide range of commonsense subjects e.g "Can fish get Tonsilitis?")Srivastava et al. (2023)
- BIG-bench Strange Stories (174 short stories followed by questions about the characters)Srivastava et al. (2023)
- BIG-bench Novel Concepts (32 find-the-common-concept problems)Srivastava et al. (2023)
- COPA (100 cause/effect multiple choice questions) Roemmele et al. (2011)
- PIQA (1,838 commonsense physical intuition 2-choice questions) Bisk et al. (2020)
- OpenBook QA (500 questions that rely on basic physical and scientific intuition about common objects and entities) Mihaylov et al. (2018).

Language Understanding tasks evaluate the model's ability to understand the structure and properties of languages, and include the following datasets:

- LAMBADA (6,153 passages take from books - we use the formatting adopted by OpenAI's version)Paperno et al. (2016)

- HellaSwag (10,042 multiple choice scenarios in which the model is prompted with a scenario and choose the most likely conclusion to the scenario from four possible options)Zellers et al. (2019)
- Winograd Schema Challenge (273 scenarios in which the model must use semantics to correctly resolve the anaphora in a sentence. The Eval Gauntlet uses the partial evaluation technique technique introduced in Trinh & Le (2019)) Levesque et al. (2012)
- Winogrande (1,267 scenarios in which two possible beginnings of a sentence are presented along with a single ending) Sakaguchi et al. (2020)
- BIG-bench language identification (10,000 questions on multiple choice language identification) Srivastava et al. (2023)
- BIG-bench conceptual combinations (103 questions using made up words) Srivastava et al. (2023)
- BIG-bench conlang translation (164 example problems in which the model is given translations of simple sentences between English and some fake constructed language) Srivastava et al. (2023)

Symbolic problem solving tasks test the model's ability to solve a diverse range of symbolic tasks including arithmetic, logical reasoning, algorithms, and algebra. These datasets include:

- BIG-bench elementary math QA (38,160 four-choice multiple choice arithmetic word problems) Srivastava et al. (2023)
- BIG-bench dyck languages (1000 complete-the-sequence questions) Srivastava et al. (2023)
- BIG-bench algorithms (1,320 questions) Srivastava et al. (2023)
- BIG-bench logical deduction (1500 four-choice multiple choice questions relating to relative ordering of objects) Srivastava et al. (2023)
- BIG-bench operators (210 questions involving mathematical operators) Srivastava et al. (2023)
- BIG-bench repeat copy logic (32 samples in which the model is required to follow some instructions for copying words/symbols)
- Simple arithmetic with spaces (1000 arithmetic problems consisting of up to 3 operations and using numbers of up to 3 digits, developed by MosaicML)
- Simple arithmetic without spaces (1000 arithmetic problems consisting of up to 3 operations and using numbers of up to 3 digits, developed by MosaicML)
- Math QA (2,983 four-choice multiple choice math word problems) Amini et al. (2019)
- LogiQA (651 four-logical word problems) Liu et al. (2020)

The **Reading comprehension** benchmarks test a model's ability to answer questions based on the information in a passage of text. The datasets include:

- BIG-bench Understanding fables (189 short stories) Srivastava et al. (2023)
- Pubmed QA Labeled (1000 hand-labeled medical documents followed by a related question for which the model must respond yes/no/maybe) Jin et al. (2019)
- SQuAD (10,570 short documents followed by a related question. The model is expected to output the exact correct answer) Rajpurkar et al. (2016)
- BoolQ (3,270 short passages on a diverse range of subjects followed by a yes/no questions) Clark et al. (2019)

11.1 Evaluation Procedure

To compute model performance on the above datasets, the Eval Gauntlet uses one of the following three ICL metrics for each dataset (from MosaicML's composer library).

1. InContextLearningQAAccuracy: This metric uses the query, the corresponding correct answer and a list of alternative answers to measure a model's prediction. If the model's response conditioned on the query starts with either the correct answer or with one of the alternative answers, it is considered correct. This is used for question-answering tasks such as TriviaQA.
2. InContextLearningLMAccuracy: This metric tests a model's ability to output a precise set of tokens. A model's output conditioned on a given query is judged to be correct only if the model's highest probability tokens match the correct sequence of tokens. This is used for language modeling tasks such as LAMBADA.
3. InContextLearningMultipleChoiceAccuracy: This metric is used for testing a model's ability to answer multiple choice questions accurately. It compares the respective perplexity of the query prepended to each of the possible choices, according to the model. If the query-choice pair with the lowest per token perplexity is indeed the correct choice, then the model's output is judged to be correct. This is used for multiple choice tasks such as HellaSwag, Winograd etc.