When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale

Max Marion
Cohere for AI
mmarion538@gmail.com

Ahmet Üstün
Cohere for AI
ahmet@cohere.com

Luiza Pozzobon
Cohere for AI
luiza@cohere.com

Alex Wang
Cohere
alexwang@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Large volumes of text data have contributed significantly to the development of large language models (LLMs) in recent years. To date, efforts to prune these datasets to higher quality subsets have relied on hand-crafted heuristics encoded as rule-based filters. In this work, we explore scalable estimates of data quality that can be used to systematically measure the quality of pretraining data, namely perplexity, the Error L2-Norm, and memorization. These metrics are used to rank and prune pretraining corpora, and we subsequently compare LLMs trained on these pruned datasets. We find that perplexity outperforms other scoring methods and improves over our no-pruning baseline while training on as little as 30% of the original training dataset without degradation in downstream finetuned tasks. Our work explores strategies in automatically curating high quality corpora and suggests that large amounts of pretraining data can be removed while retaining performance.

1 Introduction

Recent years of progress in scaling large language models (LLMs) have shown strong evidence that more data leads to better performance with remarkable gains in language understanding and generation capabilities [9, 37, 19, 3]. Common practice is to use massive datasets such as C4 [32], RefinedWeb [29], and The Pile [17]. These datasets are typically compiled by scraping raw web pages from the internet, leading to a substantial portion of the text being noisy and of low quality [14, 21, 24]. Practitioners have established a number of rule-based techniques to remove low-quality examples from these datasets [43, 32, 31, 18, 29, 40, 14, 31]. While hand-curated filters can eliminate noisy examples, they are not a substitute for a measure of "quality" for individual training examples, for which there are currently no established best practices [26].

We aim to find a rigorous estimator of data quality through *data pruning*, defined as isolating a subset of a larger training dataset such that a model trained on said subset preserves or improves performance over a model trained on the full dataset. Previous work on data pruning for language has either studied the fine-tuning setting, which typically has an order of magnitude less data [15, 4, 10] or based their method on hand picking high-quality corpora [16, 40, 9]. Our contributions are the following:

1. We benchmark data pruning based on perplexity, EL2N [28], and memorization in the LLM pretraining setting. **We find the simple technique of ranking examples based on their perplexity outperforms EL2N and memorization.**

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

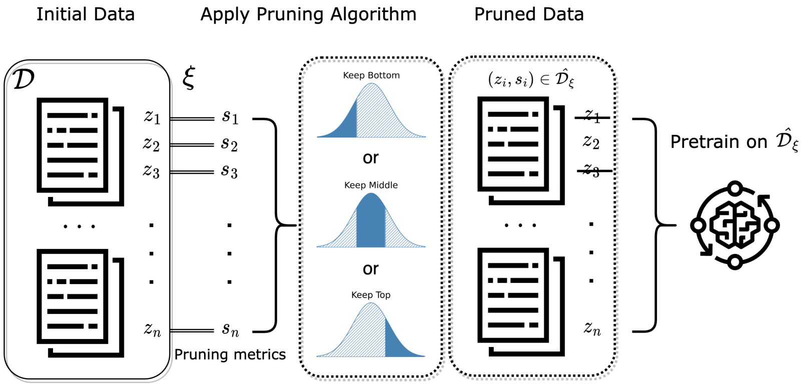


Figure 1: Demonstration of our pruning methodology. For each sequence z_i, a pruning algorithm ξ generates score s_i. We then choose which subset of the distribution of scores to keep and a new model is trained with the pruned data $\hat{\mathcal{D}}_\xi$.

2. We run a series of ablations to better understand a variety of facets of data pruning, more fully defined Table 1 We also finetune a selection of our models on six tasks from the GLUE benchmark [39] to evaluate the effect of pruning on downstream generalization, found in Table 2.
3. We test our pruning methods at both 124M and 1.5B parameters, achieving a 1% improvement in test set perplexity using half of the dataset over a baseline model trained on the entire dataset at 124M parameters. This scales to 1.5B parameter models, achieving 1.5% improvement in test set perplexity over a no-pruning baseline of the same size.

2 Methodology

Given a dataset $\mathcal{D}$, we tokenize all documents and append a special <eod> token to their end. We then concatenate and split them into n sequences z_i of fixed length t equal to the model's context length: $\mathcal{D} = \{z_1, \ldots, z_n\}$. Consider the subset of training instances $\mathcal{P}_\xi$ where ξ refers to the algorithm used to select the subset. We build this subset by computing the pruning score $Score_\xi(z_i)$ for each data point z_i. We then populate $\mathcal{P}_\xi$ with instances that fit our selection criteria: $\mathcal{P}_\xi = \{z_i \in \mathcal{D} \mid Criteria(Score_\xi(z_i))\}$. By removing $\mathcal{P}_\xi$ from $\mathcal{D}$, the remaining instances are described as: $\hat{\mathcal{D}}_\xi = \mathcal{D} \setminus \mathcal{P}_\xi$. Our goal is to choose a pruning algorithm ξ such that when training a language model on the remaining training instances, $\hat{\mathcal{D}}_\xi$, the model's performance is not diminished: $\mathbb{P}_\tau(\mathcal{M}_{\hat{\mathcal{D}}_\xi}) \geq \mathbb{P}_\tau(\mathcal{M}_\mathcal{D})$ where $\mathcal{M}_{\hat{\mathcal{D}}_\xi}$ is the model trained on $\hat{\mathcal{D}}_\xi$ and $\mathbb{P}_\tau$ is the performance on task τ.

We evaluate different reference models $\tilde{\mathcal{M}}$ that are used to calculate pruning scores. For each metric, we consider three different selection criteria to determine $\mathcal{P}_\xi$: isolating the `top`, `middle`, or `bottom` percentiles of $\mathcal{D}$ as the data to be kept. We pretrain separate models using these criteria with different percentages of the dataset to understand the dynamics and impact of each pruning metric. Figure 1 demonstrates our experimental setup.

2.1 Pruning Metrics

2.1.1 Selection via Perplexity

PERPLEXITY measures how probable a given piece of text is based on a particular language model. A lower perplexity score indicates that the model assigns a high probability to the text. For each instance z_i in $\mathcal{D}$, we compute the perplexity metric as:

$$PPL(z_i) = \exp\big(\frac{1}{|z_i|} \sum_{t_j \in z_i} NLL(t_j)\big) \tag{1}$$

Experimental axes	Choices
Pruning Metric	Perplexity, EL2N, Memorization
Pct. Data Remaining	10, 30, 50, 70
Pruning Subset	Bottom, Middle, Top
Reference Model Size	124M, 6B, 13B, 52B
Reference Model Epoch Perc.	14%, 55%, 440%, Full
Reference Model Tr. Data	CC, Wiki, Web-scale
Pruned Model Size	124M, 1.5B

Table 1: Pruning choices explored in the experiments. Under “Reference Model Training Steps”, “Full” refers to the fully trained Cohere LLMs. Under “Reference Model Training Data”, “Web-scale” refers to the significantly larger training datasets used by the Cohere reference models.

where $NLL(t_j)$ is the negative log likelihood of token t_j in sequence z_i:

$$NLL(t_j) = -\log P(t_j|t_{<j};\theta) \tag{2}$$

2.1.2 Selection via EL2N

The Error L2-Norm (EL2N) score was originally proposed in a computer vision setting to identify which samples are important for learning [28]. The authors suggest that exhibiting a low EL2N score are typically those the model learns in its early stages of training, likely because they are relatively easier. We define the EL2N score on text sequences as the average L_2 norm of the error vector, where $\hat{y}_i$ is the reference model's predicted probability distribution over the vocabulary and y_t is the one-hot encoded representation of the ground truth:

$$\text{EL2N}(z_i) = \frac{1}{t}\sum_{i}^{t} \|\hat{y}_t - y_t\|_2 \tag{3}$$

Additional specifics of EL2N calculations can be found in Appendix Section B

2.1.3 Memorization Ranking

Memorization in language models is a well-studied phenomenon [11, 12, 6]. We explore memorization scores applied as a data pruning ranking using memorization score as defined by [6]:

$$score(M, N) = \frac{1}{N}\sum_{i}^{N} 1(z_{M+i} = \hat{z}_{M+i}) \tag{4}$$

where z is a data point, $\hat{z}$ is a sequence of tokens predicted by the reference model, and $1(\cdot)$ is an indicator function. A reference model guaranteed to have seen the full training set is prompted with the first M tokens of a data point z to calculate the memorization score. We then greedily generate N additional tokens, $\hat{z}$. The memorization score is the fraction of the N greedily generated tokens ($\hat{z}_{M:M+N}$) that match exactly with the original data point ($z_{M:M+N}$). For our experiments, $M = N = 32$. A high memorization score indicates the model reproduces more of the text verbatim.

2.1.4 Random Pruning

We also evaluate a lower bound of expected performance: pruning a random selection of samples.

3 Experiments

3.1 Model

We train autoregressive decoder-only Transformer models [38] with a standard language modeling objective. Given an input sequence of $z_i = [r_1, \cdots, r_t]$ from training data $\mathcal{D}$, a language model with parameters θ is trained to minimize the negative log-likelihood loss as defined in Equation 2. Our language models follow the traditional GPT-style architecture [30].

While training our models, we use AdamW [23] with linear cosine scaling and a batch size of 2048. The 124M parameter models are trained for 8000 steps, which amounts to a total of 33B tokens with a learning rate that linearly increases from 0 to 1.5e-4 over the course of training. This is approximately 4.4 epochs over the unpruned dataset. We tokenize the data with Byte Pair Encoding [33] with a vocabulary of 51200. Due to the memory and computational costs of training 1.5B parameter models, our experiments at this size are trained with a batch size of 512 for 14568 steps. As such, the models see only 7.6B tokens, equivalent to a single epoch of our unpruned dataset. The learning rate for 1.5B parameter models linearly increases from 0 to 1.2e-4 over the course of training. All models use a context window length of 2048.

3.2 Data

We use a random sample of the May 2022 snapshot of CommonCrawl[1] in our experiments. After downsampling the unpruned dataset has 7.6B tokens, about 20% of the full snapshot. This downsampling is required due to the computational cost of our various ablation experiments. This dataset is *prefiltered* using a combination of automatic and hand-crafted filters, as we aim to further improve data quality beyond common rule-based filters. The filters exclude repetitive documents, documents with percentages of special characters, and documents that contain explicit words and toxic text, similar to deduplication steps seen in [36, 20].

3.3 Ablations

For all techniques, we compare performance when 10%, 30%, 50%, and 70% of the full dataset is preserved. We retain the `top`, `middle`, and `bottom` subsets according to the pruning ranking, e.g., when retaining 30% of the bottom of the pruning metric's distribution over the training set, we calculate the 30th percentile of the pruning metric's distribution and remove all data points with a perplexity above it. When retaining the `middle` 30%, we calculate the 35th and 65th percentile and remove all data points above and below those numbers respectively. Each ablation study (pruning method, percent data remaining, section of distribution preserved) **requires training a new model from random initialization**. Table 1 summarizes the perplexity pruning variations we explore in this paper. We call models used to compute the perplexity ranking *reference models* and the models trained on the pruned datasets *pruned models*.

3.4 Evaluation

We report perplexity on a test set from the same CommonCrawl snapshot with identical prefiltering as the training data. This test set contains 266M tokens, equivalent to about 3.5% of the training set. We also finetune a subset of our models on six different classification tasks from GLUE, details for which can be found in Appendix Section D.

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

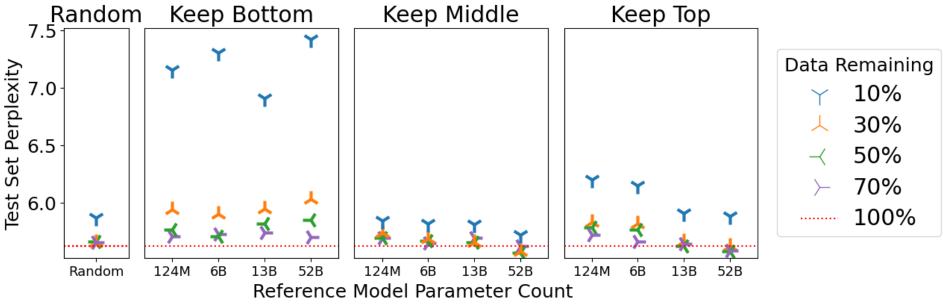
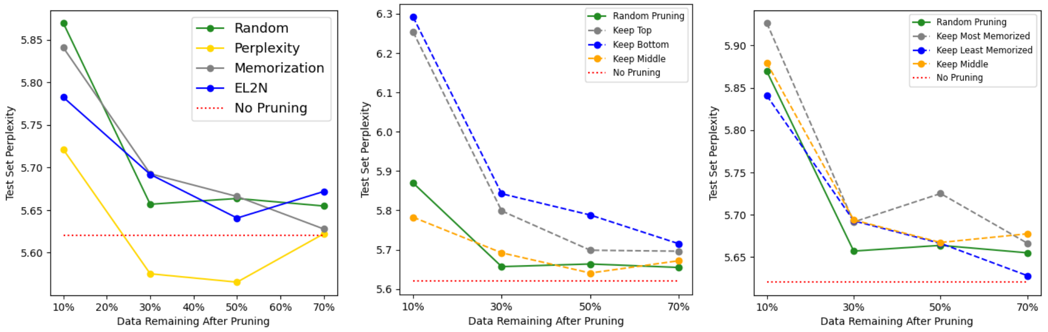


Figure 2: The effect of reference models of different sizes on test set perplexity. The three subset selection approaches (keep `bottom`, `middle`, or `top`) for each set of experiments are showcased separately.



(a) Top variant of each metric

(b) EL2N

(c) Memorization

Figure 3: Top performing variants across dataset sizes 3a. Evaluation of different subset selection criteria for 3b EL2N and 3c Memorization. Perplexity-based pruning consistently surpasses other metrics and the no pruning experiments. See Section 4.2 for details on the featured variants.

4 Results and Discussion

4.1 Removing Easy Instances Improves Performance

Though the most competitive variant for each pruning method varies based on the subset of the scoring distribution retained, we observe a consistent pattern: the most performant variants are *not* the subsets that correspond to the "easiest" data. The interpretation of the term "easy" varies according to the measurement employed. When employing the PERPLEXITY metric, it refers to the `bottom` samples with the lowest perplexity. With the EL2N metric, it also pertains to the `bottom` samples exhibiting the lowest initial loss, while for MEMORIZATION, it relates to the `top` samples that have been most thoroughly memorized.

Figure 2 demonstrates this pattern when using PERPLEXITY. In contrast to the `middle` or `top` subsets, the `bottom` subset has much less variance in results between reference models of varying sizes, indicating the `bottom` subset is not suitable for training. The `middle` experiments achieve consistently lower test set perplexities for various reference model sizes and pruning ratios. Figure 3c shows the results for the EL2N metric, where the `middle` subset is also the best variant. While the best performing run does not outperform the no pruning baseline, it is achieved when retaining 50% of the `middle` subset, outperforming the model trained on 70% of the dataset, similar to the results when using PERPLEXITY. Finally, when using MEMORIZATION FACTOR as a pruning metric, keeping the least memorized samples (`bottom` subset) generally performs best. Figure 3c shows model performances for this metric. The most competitive variant of the memorization metric is the `bottom` 70% of the distribution. Memorization never outperforms the no-pruning baseline.

We posit that the `middle` subset performs best because it removes easy data that no long contributes to training, as well as hard data that might not be useful for general purpose modeling of the test set. Additionally, a lower predicted probability for the correct token has a more pronounced impact on PERPLEXITY as compared to EL2N. This effect is particularly notable in challenging instances, as it tends to push them towards the extremes of the PERPLEXITY distribution more significantly than the distribution observed with EL2N, as observed in Fig 7, likely leading to PERPLEXITY'S improved performance over EL2N. MEMORIZATION has significantly less variance in its scores, likely leading to its less pronounced effects as a pruning metric.

4.2 Simple Pruning Metrics Outperform More Sophisticated Approaches

In Figure 3a we present results comparing the performance of the best variant of each pruning metric: (1) retaining the `middle` of the distribution of PERPLEXITY scores by the fully trained 52B reference model, (2) retaining the `bottom` of the distribution of the MEMORIZATION FACTOR (least memorized samples), and (3) retaining the `middle` of the distribution of EL2N scores from the 1000 step checkpoint. Our results show that training on the `middle` subset using PERPLEXITY outperforms other pruning metrics across all dataset sizes. For some variants, it also outperforms training on the entire dataset. Compared with the no-pruning baseline, pruning to the `middle` 50% of the perplexity distribution leads to a 0.97% improvement in perplexity. Using only the `middle` 30% of the data achieves nearly the same performance, with a 0.80% improvement over the no-pruning baseline.

4.3 Pruning Benefits from Using Larger Reference Models

Our results in Figure 2 show pruned model performance after pruning with PERPLEXITY calculated with reference models ranging from 124M to 52B parameters. We find that increasing reference model size improves pruned model performance over the no-pruning baseline when either the `middle` or `top` subsets are used. Pruning using 52B parameter reference model perplexity scores achieves a 2.2% improvement in perplexity over the best-performing pruned model from the 124M parameter reference model experiments. Furthermore, for 13B and 52B reference models, we observe better performances with less training data when keeping the middle and top subsets. For both of these larger models, retaining the `middle` 30% and 50% of the training data produces pruned models that outperform the pruned models trained on the `middle` 70% of the training set.

We note that the effects of subset selection, such as the `bottom` subset performing worse and approximately scale with the size of the reference models. The larger reference models' `bottom` subset training runs perform even worse than their smaller counterparts when retaining the same percentage of the training set. This overall points to a consistent finding that larger models are better calibrated at computing a useful data pruning ranking.

4.4 Perplexity-based Pruning Improvements Generalize to Larger Scale Models

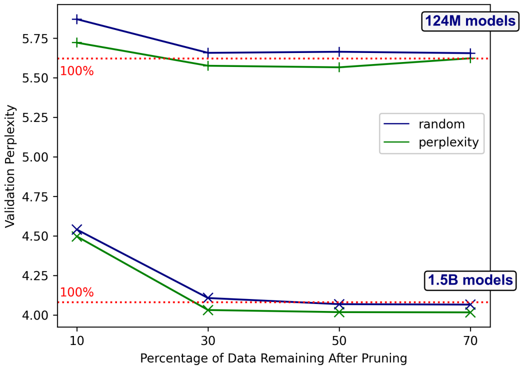


Figure 4: Comparing the most performant pruning variant (keeping the `middle` subset with 52B parameter reference model) with random pruning at 124M and 1.5B parameters. The improvement in performance of a perplexity-based pruning approach carries from 124M to 1.5B parameter models.

We explore our strongest pruning variant – PERPLEXITY computed using a 52B parameter reference model while retaining the `middle` subset – at a larger scale by training 1.5B parameter models. Training a 1.5B model, seen in Figure 4, we observe that perplexity-based pruning achieves better results than random pruning across all pruning percentages. The improvement observed with perplexity-based pruning over random pruning follows a consistent pattern for both the 124M and 1.5B models, demonstrating the scalability of our approach to a large-scale pretraining setting.

5 Conclusion

In this study, we showed that data pruning cam improve model performance in a pretraining setting. We find that training on the "easiest" examples in a dataset usually degrades performance, where "easiest" varies depending on the pruning metric used. Models trained on as little as half of the data selected by perplexity achieve up to 1.5% improvement over models trained on the full dataset. Additionally, we establish the consistency of our findings as we scale the model sizes. While scaling up the amount of data LLMs are trained on remains a popular avenue for improving models, our work demonstrates that carefully pruning these large training corpora is also a fruitful direction for making models better.

References

[1] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

[2] C. Agarwal, D. D'souza, and S. Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10368–10378, June 2022.

[3] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu. Palm 2 technical report, 2023.

[4] J.-M. Attendu and J.-P. Corbeil. Nlu on data diets: Dynamic data subset selection for nlp classification tasks, 2023.

[5] F. Bane, C. S. Uguet, W. Stribiżew, and A. Zaretskaya. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325, Orlando, USA, Sept. 2022. Association for Machine Translation in the Americas.

[6] S. Biderman, U. S. Prashanth, L. Sutawika, H. Schoelkopf, Q. Anthony, S. Purohit, and E. Raff. Emergent and predictable memorization in large language models, 2023.

[7] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

[8] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[10] Y. Cao, Y. Kang, and L. Sun. Instruction mining: High-quality instruction data selection for large language models, 2023.

[11] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying memorization across neural language models, 2023.

[12] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models, 2021.

[13] W. Chen. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[14] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus, 2021.

[15] M. Fayyaz, E. Aghazadeh, A. Modarressi, M. T. Pilehvar, Y. Yaghoobzadeh, and S. E. Kahou. Bert on a data diet: Finding important examples by gradient-based pruning, 2022.

[16] L. Gao. An empirical exploration in quality filtering of text data, 2021.

[17] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021.

[18] D. Hernandez, T. Brown, T. Conerly, N. DasSarma, D. Drain, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume, S. Johnston, B. Mann, C. Olah, C. Olsson, D. Amodei, N. Joseph, J. Kaplan, and S. McCandlish. Scaling laws and interpretability of learning from repeated data, 2022.

[19] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

[20] D. Kocetkov, R. Li, L. B. Allal, J. Li, C. Mou, C. M. Ferrandis, Y. Jernite, M. Mitchell, S. Hughes, T. Wolf, D. Bahdanau, L. von Werra, and H. de Vries. The stack: 3 tb of permissively licensed source code, 2022.

[21] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, A. N. Rubungo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. Çabuk Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 01 2022.

[22] H. Laurençon, L. Saulnier, T. Wang, C. Akiki, A. V. del Moral, T. L. Scao, L. V. Werra, C. Mou, E. G. Ponferrada, H. Nguyen, J. Frohberg, M. Šaško, Q. Lhoest, A. McMillan-Major, G. Dupont, S. Biderman, A. Rogers, L. B. allal, F. D. Toni, G. Pistilli, O. Nguyen, S. Nikpoor, M. Masoud, P. Colombo, J. de la Rosa, P. Villegas, T. Thrush, S. Longpre, S. Nagel, L. Weber, M. Muñoz, J. Zhu, D. V. Strien, Z. Alyafeai, K. Almubarak, M. C. Vu, I. Gonzalez-Dios, A. Soroa, K. Lo, M. Dey, P. O. Suarez, A. Gokaslan, S. Bose, D. Adelani, L. Phan, H. Tran, I. Yu, S. Pai, J. Chim, V. Lepercq, S. Ilic, M. Mitchell, S. A. Luccioni, and Y. Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[24] A. Luccioni and J. Viviano. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online, Aug. 2021. Association for Computational Linguistics.

[25] S. Mindermann, J. Brauner, M. Razzak, M. Sharma, A. Kirsch, W. Xu, B. Höltgen, A. N. Gomez, A. Morisot, S. Farquhar, and Y. Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022.

[26] M. Mitchell, A. S. Luccioni, N. Lambert, M. Gerchick, A. McMillan-Major, E. Ozoani, N. Rajani, T. Thrush, Y. Jernite, and D. Kiela. Measuring data, 2023.

[27] N. Muennighoff, A. M. Rush, B. Barak, T. L. Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf, and C. Raffel. Scaling data-constrained language models, 2023.

[28] M. Paul, S. Ganguli, and G. K. Dziugaite. Deep learning on a data diet: Finding important examples early in training, 2023.

[29] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023.

[30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[31] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis and insights from training gopher, 2022.

[32] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[33] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units, 2016.

[34] S. A. Siddiqui, N. Rajkumar, T. Maharaj, D. Krueger, and S. Hooker. Metadata archaeology: Unearthing data subsets by leveraging training dynamics, 2022.

[35] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023.

[36] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science, 2022.

[37] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[39] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[40] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[41] S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023.

[42] S. M. Xie, S. Santurkar, T. Ma, and P. Liang. Data selection for language models via importance resampling, 2023.

[43] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[44] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. Lima: Less is more for alignment, 2023.

A Related Work

A.1 Rule-Based Data Pruning in NLP

Significant portions of web-scraped data used for language model pretraining have been shown to be of low quality, machine-generated spam, pornographic content [21]. Selection processes to determine what should be included in large-scale datasets have centered on rule-based filters and heuristics [5], such as keeping only text written in English [32, 31] or removing sequences containing blocklisted words [32]. There are also quality-based rules such as removing duplicated samples [43] or filtering sentences that do not fit a certain amount of words [32, 31]. Rule-based approaches for data filtering have shown controversial effects on model performance, with some works advertising improvements on language modeling capabilities [29, 32], while others do not [8, 7]. Also, heuristics are prone to undesired outcomes due to their simplicity. For instance **(author?)** [14] show how removing blocklisted words disproportionately removes text from and about minority individuals.

A.2 Metric-Based Data Pruning in NLP

Recent work on metric-based pruning has mainly focused on pruning data from the fine-tuning stage of LLMs [4, 42] most probably due to the prohibitive cost of pruning at the pretraining scale. **(author?)** [4] perform dynamic pruning during the fine-tuning stage by establishing a curriculum of samples based on their EL2N scores [28]. Similarly, we benchmark EL2N as a static data-pruning metric for language datasets. Our work joins the few others that aim to reduce pretraining dataset sizes [41, 13, 1]. **(author?)** [1] apply their deduplication method based on embeddings to further improve the performance of a previously filtered dataset. We also perform pruning on previously filtered datasets, aiming to enhance performance further. Previously, perplexity has been used to filter datasets [27, 40, 22], but its pruning capabilities have been underexplored. **(author?)** [22] and **(author?)** [27] filter out high-perplexity samples from their corpus as those are framed as unnatural language and harmful for performance according to their reference domain, which is Wikipedia. In contrast, we benchmark pruning to low perplexity values and high and medium-valued subsets of a dataset's distribution to understand which is the most valuable section for pretraining at scale. We also explore different reference model sizes and training sets.

A.3 Data pruning in Computer Vision

The majority of work to date on data pruning [35] and isolating data subsets [34, 25] using model signal has centered on computer vision. These are typically structured in a supervised setting. In contrast, our focus is on a large-scale NLP pretraining where the objective is unsupervised pretraining. Most relevant to our method is work by **(author?)** [35] which empirically studies reducing datasets in a teacher/trained regime, using a teacher model's margin as a pruning metric. They find that, with abundant data, training only on the hardest examples yields better performance, while conversely when data is scarce, training on only the easiest example yields better performance.

B Early Reference Model Checkpoints Serve as Effective Scoring Models

Motivated by several works that have found that there is a signal in early training checkpoints [28, 2, 34], we investigate whether early checkpoint of a reference model during training offers adequate signal for calculating discriminative pruning scores. We study PERPLEXITY and EL2N scores obtained from two early checkpoints: after training on approximately 14% and 55% of the full training dataset (250 and 1000 training steps respectively). We train ten different reference models with different random initializations and average the EL2N score from all ten models to obtain our final EL2N score. Figure 5 showcases the results of these experiments. Examining the

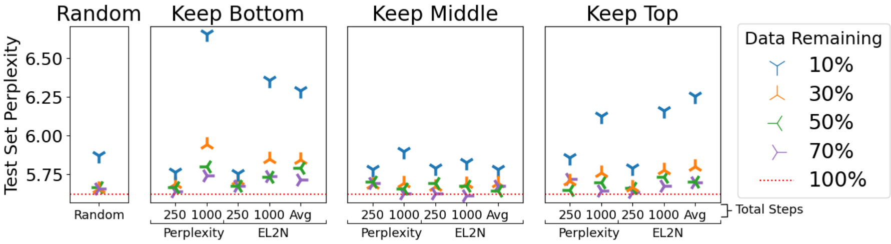


Figure 5: The impact of using an early checkpoint of the reference model in pruning based on Perplexity and EL2N metrics.

14% checkpoint for both perplexity and EL2N, we notice minimal variance across percentages and subset selection criteria. Performance across subsets changes considerably less than either the 55% checkpoint or the fully trained models.

Given this, we deduce that training on only 14% of the data is inadequate for our reference model to offer precise pruning scores. In contrast, the 55% reference models perform in a similar manner to the fully trained models, performing best with the `middle` subset, worst with the `bottom` subset, and comparably with the `top` subset. Fully training the reference model is shown not to be necessary to uphold comparable performance. Halving the reference model training steps proves effective, enabling the utilization of early checkpoints. In practice, we expect many practitioners to use off the shelf models for computing perplexity and may not need to carry the cost of pretraining a reference model from random initialization.

We also show performance for EL2N scores averaged across 10 reference models, initialized with different random seeds. We selected the 55% reference models given our previous result.

While the best pruned models using the averaged EL2N score did not outperform the best pruned models trained on only one reference model's EL2N score, the pattern of performance more similarly mirrors what we see with the larger, fully trained reference models. Specifically, in the `middle` subset, using 50% of the dataset outperforms using 70%. When constrained to the `bottom` subset, performance more clearly monotonically degrades when using less data than when using the 55% reference model, whereas the earlier checkpoint has comparable performance when retaining 30, 50, and 70% of the data. This implies that averaging scores across reference models helps hone the pruning signal, identifying subsets "easy" or "hard" subsets in more similar ways to larger models.

C Improved Pruning Signals Result from Reference Models Trained on Cleaner Data

In this section we ask: *does the data the reference model is trained on impact the quality of the ranking?* We compare the perplexity rankings generated by reference models trained on two different corpora: Wikipedia and CommonCrawl. We investigate whether a model trained on Wikipedia, a dataset frequently hand-picked as a high-quality dataset [42, 40], generates more effective pruning signals for perplexity rankings. In Figure 6, we compare the performance of the two variants across different pruning percentages and subset selections. We observe that in the two optimal selection variants from the general reference models (`middle` and `top`) a model trained on Wikipedia consistently yields lower validation perplexity compared to a model trained on CommonCrawl. Wikipedia's best variant, pruning to the middle 70%, outperforms CommonCrawl's best variant, also pruning to the middle 70%, by 0.69%. This finding overall suggests that investing in a high quality reference model to generate rankings results in more effective data pruning. Reference models trained on higher quality data are better at identifying a subset of data points most conducive to model performance.

D Downstream Evaluation on GLUE

Previously, we demonstrated various ways of pruning the pretraining data and training models with different data sizes. Considering that the pretraining stage primarily focuses on knowledge acquisi-

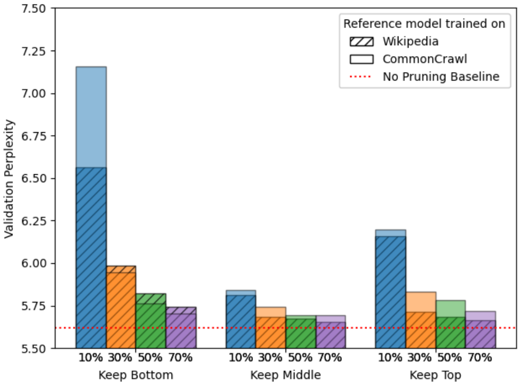


Figure 6: Performance of different pruning strategies using two different reference models: one trained on Wikipedia and one trained on CommonCrawl. A reference model trained on Wikipedia (an example of a clean noise-free corpus) achieves consistently lower validation perplexity compared to a reference model trained on a noisier CommonCrawl in our two robust settings (`middle` and `top`).

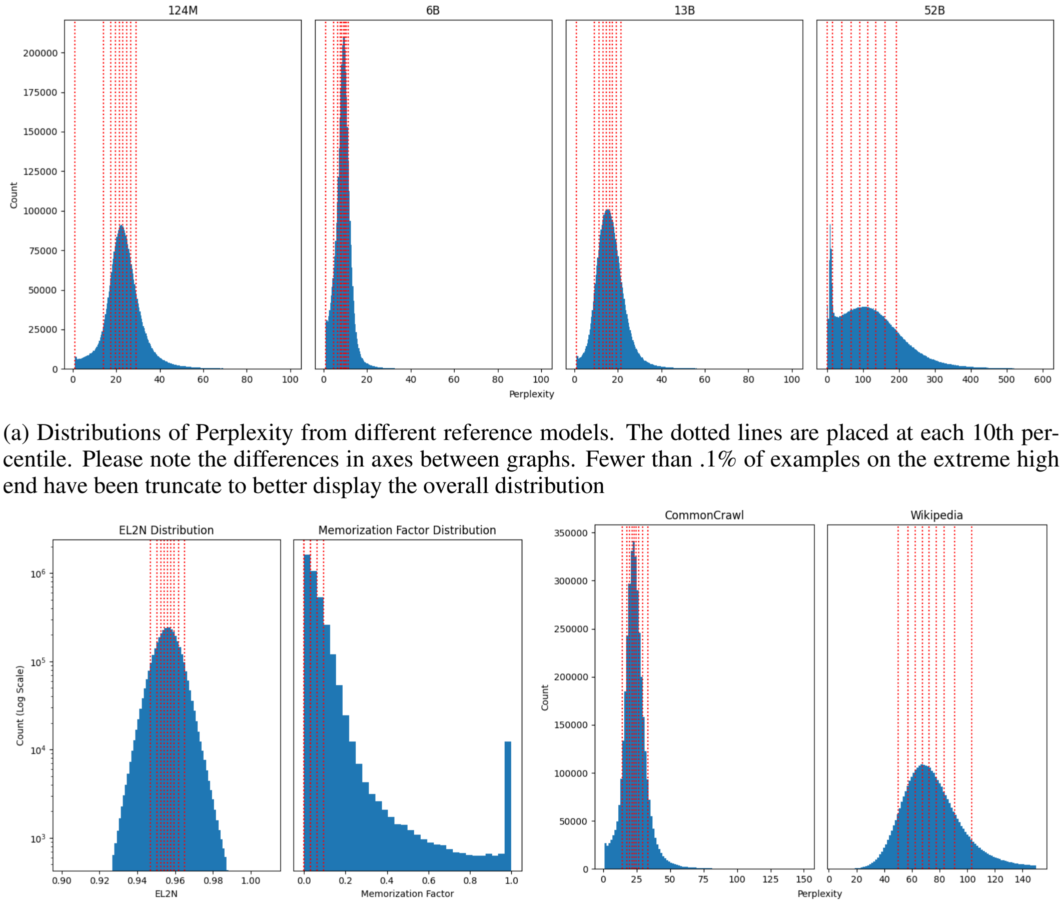
Table 2: Mean accuracy and standard deviation of the best variants of each pruning algorithm for GLUE classification tasks. Underlined results surpass the baseline performance with no pruning. The best results for each task are marked in bold. Results are reported for 5 runs of each model, trained for 3 epochs with a learning rate of $1e-5$.

	Data Remaining	SST2	MRPC	QQP	QNLI	RTE	WNLI
No Pruning	100%	$78.15_{0.002}$	$64.32_{0.021}$	$76.55_{0.001}$	$65.40_{0.006}$	$49.69_{0.024}$	$51.56_{0.040}$
Random Pruning	70%	$77.92_{0.002}$	$\underline{65.21}_{0.017}$	$\underline{76.58}_{0.002}$	$65.11_{0.006}$	$49.69_{0.013}$	$48.44_{0.038}$
	50%	$\underline{78.19}_{0.003}$	$\underline{65.16}_{0.020}$	$76.40_{0.001}$	$\underline{65.44}_{0.006}$	$\underline{49.92}_{0.009}$	$49.69_{0.062}$
	30%	$77.29_{0.007}$	$\mathbf{66.04}_{0.017}$	$76.36_{0.001}$	$65.22_{0.005}$	$\underline{51.33}_{0.024}$	$50.31_{0.057}$
	10%	$76.44_{0.006}$	$\underline{65.83}_{0.021}$	$75.91_{0.001}$	$64.40_{0.007}$	$\underline{50.70}_{0.007}$	$50.62_{0.016}$
Memorization Bottom subset	70%	$77.29_{0.006}$	$\underline{64.38}_{0.016}$	$76.42_{0.001}$	$\underline{66.03}_{0.007}$	$49.06_{0.021}$	$49.06_{0.042}$
	50%	$77.89_{0.006}$	$\underline{65.47}_{0.017}$	$76.51_{0.001}$	$\underline{65.99}_{0.005}$	$\underline{49.77}_{0.013}$	$50.31_{0.048}$
	30%	$\mathbf{78.52}_{0.004}$	$\underline{65.89}_{0.016}$	$76.48_{0.001}$	$\underline{65.91}_{0.006}$	$\underline{50.31}_{0.009}$	$\mathbf{54.38}_{0.061}$
	10%	$76.64_{0.004}$	$\underline{65.16}_{0.015}$	$76.11_{0.001}$	$64.61_{0.006}$	$\underline{50.39}_{0.016}$	$\underline{51.88}_{0.059}$
EL2N Middle subset	70%	$\underline{78.61}_{0.008}$	$\underline{66.46}_{0.018}$	$\underline{76.93}_{0.001}$	$\underline{67.00}_{0.005}$	$48.67_{0.017}$	$50.00_{0.058}$
	50%	$\underline{79.17}_{0.007}$	$\underline{65.42}_{0.016}$	$76.35_{0.001}$	$62.43_{0.007}$	$\underline{51.41}_{0.028}$	$51.56_{0.049}$
	30%	$\underline{78.98}_{0.005}$	$\underline{65.41}_{0.012}$	$\mathbf{77.47}_{0.001}$	$\underline{68.63}_{0.005}$	$49.69_{0.022}$	$\underline{55.31}_{0.067}$
	10%	$\underline{78.31}_{0.006}$	$63.38_{0.016}$	$\underline{76.93}_{0.001}$	$65.34_{0.006}$	$\mathbf{51.95}_{0.021}$	$51.25_{0.064}$
Perplexity (52B) Middle subset	70%	$\underline{78.40}_{0.004}$	$\underline{64.43}_{0.020}$	$\underline{76.68}_{0.001}$	$\mathbf{66.74}_{0.007}$	$\underline{50.16}_{0.023}$	$49.06_{0.012}$
	50%	$78.01_{0.006}$	$\underline{64.37}_{0.021}$	$76.82_{0.001}$	$\underline{66.00}_{0.004}$	$\underline{50.62}_{0.023}$	$50.31_{0.021}$
	30%	$77.34_{0.005}$	$\underline{64.84}_{0.023}$	$\underline{76.76}_{0.001}$	$\underline{65.89}_{0.002}$	$\underline{50.86}_{0.009}$	$50.94_{0.031}$
	10%	$77.66_{0.006}$	$\underline{65.36}_{0.017}$	$76.40_{0.001}$	$\underline{66.52}_{0.007}$	$\underline{51.17}_{0.012}$	$\underline{53.44}_{0.040}$

tion [44], we inquire about the potential ripple effects of pruning data during pretraining when these models are subsequently finetuned on downstream tasks. To analyze the impact of different pruning strategies on LLM capabilities, we finetune and evaluate models on a subset of the GLUE tasks [39]. Results are presented in Table 2. No single pruning strategy (combining both pruning metric and percentage of remaining data) stands out as superior across all tasks, the absence of a universally dominant approach is consistent with earlier findings in the literature [16]. Even random pruning shows improvements in certain tasks, underscoring the significance of downsampling when handling noisy data during the pretraining stage to mitigate potential learning degradation. While the scale of model size and training time cannot be expected to achieve state of the art results, our results do not show that directed pruning drastically degrades downstream performance. As such, our pruning methods have not excluded a significant language modeling capability required in GLUE, suggesting that future work on larger trained models are unlikely to lose general performance capabilities either.

E Metric Distributions

We present the total distributions of the pruning metrics used in our analysis in Figure 7.



(a) Distributions of Perplexity from different reference models. The dotted lines are placed at each 10th percentile. Please note the differences in axes between graphs. Fewer than .1% of examples on the extreme high end have been truncate to better display the overall distribution

(b) Distributions of the EL2N and Memorization Factor metrics. The dotted lines are placed at each 10th percentile and omitted from Memorization Factor due to overlap. Please note the log-scaled y-axis.

(c) Distributions of Perplexity from reference models trained on Wikipedia and CommonCrawl. The CommonCrawl model is the same as the 124M parameter model in Figure 7a. The dotted lines are placed at each 10th percentile.

Figure 7: Distributions of different pruning metrics and reference models.

F Examples from different selection criteria

Examples from the pretraining data, drawn from distinct subsets (keep bottom, keep middle, keep top), are presented in Tables 3, 4, 5, 6, and 7, with rankings based on perplexity.

Table 3: Samples from different distribution subsets using perplexity of a 52B reference model trained on CommonCrawl.

Bottom 10%	Middle 10%	Top 10%
Submissions, you hereby grant Company a license to translate, modify (for technical purposes, for example making sure your content is viewable on an iPhone as well as a computer) and reproduce and otherwise act with respect to such User Submissions, in each case to enable us to operate the Services, as described in more detail below. This is a license only – your ownership in User Submissions is [...]	House Municipal Heritage Building is a two-storey, wooden, vernacular building with a low-hipped roof, and is located at the Norris Point Lookout, 104 Main Road, Norris Point, Newfoundland and Labrador. The former family dwelling now operates as a heritage museum with a view of the Tablelands of Gros Morne National Park located on the great Northern Peninsula. The municipal heritage designation [...]	and a nice book as a nice price. Postage is via Royal Mail 1st Class in the UK. If you are buying from overseas then please contact me before completing your purchase for a quote. I will always combine P&P so if ordering multiple books, please wait for the invoice so that discounts can be applied. We are slowly populating our store with post war Wisden's so if there is anything you need that [...]
provided on the Site is not intended for distribution to or use by any person or entity in any jurisdiction or country where such distribution or use would be contrary to law or regulation or which would subject us to any registration requirement within such jurisdiction or country. Accordingly, those persons who choose to access the Site from other locations do so on their own initiative and are [...]	selection of fuel type and input of soot index, coefficient of fuel, selection of measurement units, input of date and time with keyboard and via RS232 or RS485 Procedure of industrial emissions monitoring with the use of AHKAT-410 has been agreed in FSUE "SRI Atmosphere" AHKAT-410-16 is approved for diesel locomotive and diesel train emission monitoring at environment monitoring stations in [...]	can be returned up to 28 days after the date of purchase. Please note, we cannot offer refunds on beauty, pierced jewellery or on swimwear if the hygiene seal is not in place or has been broken. We now offer FREE label-free returns with InPost Lockers (available 24/7), FREE Doddle Returns to all UK customers as well as a FREE UK Collect+ returns service via over 5,900 local stores nationwide.[...]
license only – your ownership in User Submissions is not affected. You agree that the licenses you grant are royalty-free, perpetual, sublicensable, irrevocable, and worldwide. Any information or content publicly posted or privately transmitted through the Services is the sole responsibility of the person from whom such content originated, and you access all such information and content at your [...]	1 1/2 " steel plate, all weld construction Hammer mill machine manufacturers, suppliers, exporters, dealers and traders in India and worldwide hammer mill machines from Gujarat and Mumbai since 1960 as per the ISO standards with required industrial features and specifications Replaceable bar type grate is available for specific applications SPECIFICATIONS : Hammer stone crusher is a kind of equip [...]	several turns. Nearly a month after a foreclosure lawsuit was filed against Freestyle Music Park and its parent company, more than a dozen former department heads have sued seeking more than $232,000 in unpaid wages and bonuses, according to court papers filed late Friday. Seventeen employees are listed as plaintiffs. Backpay I can understand, but can you honestly expect any kind of bonuses [...]

Table 4: Samples from different distribution subsets using perplexity of a 124M reference model trained on CommonCrawl.

Bottom 10%	Middle 10%	Top 10%
risk your food going bad in a lukewarm fridge when you can lease kitchen appliances in West Hollywood through Acima! Are you a budding DJ? A bit of a high-fidelity audio snub? Love to level up with the latest video game system? Level up your entertainment at home and on the road with sound systems for lease in West Hollywood. You can make flexible lease renewal payments on the best in-home sound [...]	gratitude exercise. Before you get out of bed, think of five things you are most grateful for. If your Life Path number is 2, you have a duality fit for any earthly experience. You are deeply rooted in balance and harmony when dealing with the other numbers. In order to stay connected to your community, start your day by connecting with your friends and family. Instead of hopping on social [...]	keepers" definitely won't help! Then there are those whose idea of a school librarian is based on one they remember from their childhood, who perhaps didn't let them borrow from the adult shelves or maybe told them to be quiet. You know - the cliched woman with glasses and a bun? I wear glasses myself and ended up haing to get a haircut to avoid the cliche. In summer, of course I had to put my [...]
the-art mixed-use development that features a wide variety of shops, services, and restaurants, along with over 950 luxury apartments. The sprawling urban village is pedestrian-friendly and is the perfect place if you want to indulge in a shopping spree or treat your taste buds to a hearty meal. If you're thinking about looking for the perfect home in Brookhaven, I'm ready to help! Get in touch [...]	it as a stand-alone piece but later experimented performing it as my written prediction, confabulation style, Closing Effect. It's still a work in progress but I did receive some "Standing Ovations!" ALAN ARITA "I received a copy of GAME NIGHT and IT IS EXCELLENT! First, the quality of the book is outstanding; everything from the artwork, layout, hidden gems, and of course the precision cut [...]	and view the supernal beauty that lies beyond. (I wish I'd have said that first; actually I stole it from a guy who wrote it a hundred years ago!*) But if I couldn't see into the future for a few years, there wouldn't be a Christmas story today. I've a whole lot of notes still in my jeans. One's about Rabbi Frankel of the Synagogue across West Street from old Reno High School. He was a pretty [...]
toilet drains are overwhelmed with toilet paper or clogged by non-flushable things that find their way into the drain. If that's the case, it may be time to call a plumbing technician. Unexpected toilet issues interrupt your daily routine, turning what you expected to be a good day right into a stressful one. You need help ASAP! Best quality Plumbing is ready to solve your toilet troubles no [...]	who offer 3D printing services these days. Try searching for someone who offers them in your area.Last week, Apple announced the new A15 processor in a peculiar way: by comparing its new chip to the Android competition, rather than the A14 that powered last year's generation of iPhones. We were all left to try to infer the speed of the A15 based on Apple's claims, and wondering if the company was [...]	floor study, family room, kitchen, unfinished basement for future expansion & 2 car garage. Lennar seamlessly blended & showcased the unparalleled beauty of Colorado with the most innovative homes, energy efficient technologies & modern conveniences, bringing the best of both worlds together. Beautiful finishes and upgrades throughout. Lennar provides the latest in energy efficiency and state of [...]

Table 5: Samples from different distribution subsets using perplexity of a 124M reference model trained on Wikipedia.

Bottom 10%	Middle 10%	Top 10%
of our kids, demonstrated ability to create meaningful change, a strong commitment to learning, and an ability to work in partnership with others." Individuals accepted to this program agree to a two-year teaching commitment. If you become a core member you are required to attend an intensive summer training program to prepare for your two-year commitment. Each region has different requirements b [...]	HST single cylinder hydraulic cone crusher. HST single cylinder hydraulic cone crusher integrates mechanical, hydraulic, electrical, automation, intelligent control and other technologies, which can be widely used in medium, fine and ultra-fine crushing operations in metal and non-metal mines, cement, sandstone, metallurgy and other industries... 1,214 roller cone crusher products are offered [...]	active play outdoor. Users without a subscription are not able to see the full content on this page. Please subscribe or login.On the net betting houses include was able to offer followers a fabulous best range of luring optimistic aspects. A style of online casino money provides consistently continually really been ornamented and acquired in reaction to make sure you basic safety issues. Insi [...]
to be that way. Weight loss surgery in Hanover is a great option for those who are at least fifty pounds overweight and have struggled with weight loss over the years. There are a number of surgical weight loss procedures available to those seeking treatment, and Nusbaum Weight Loss Centers of New Jersey, with offices and bariatric surgeons in Morristown, Morris County, Morris County, and surrou [...]	sperm whales. Learn firsthand about Sri Lanka's amazing biodiversity on this private tour to the Kanneliya Rainforest. With a dedicated guide leading you, explore the UNESCO-listed biosphere reserve, home to monkeys, snakes, chameleons, and a wide range of bird life. Learn about the flora and fauna through commentary tailored to your interests and enjoy plenty of chances to ask questions. Explo [...]	row for spotting this Sabal Trail posting within minutes.The skin has become delicate. I just received the goods and I didn't know how to use it. I consulted the customer service. I didn't expect the customer service person to be super good and the introduction was super careful. I have been so successful and happy trading with you every time.. I hope we have more transactions in the future... Ha [...]
to which coverage is thereby to be granted; and (2) Shall insure the person named therein and any other person, as insured, using any such motor vehicle or motor vehicles with the express or implied permission of such named insured against loss from the liability imposed by law for damages arising out of the ownership, maintenance, or use of such motor vehicle or motor vehicles within the United [...]	Also, I have attached a brief presentation of our work for better understanding.A two-year solar energy project at the University of Sheffield has shown almost all of the 2,000 systems in the scheme are still performing better than expected. Researchers running Sheffield Solar Farm, which was launched in August 2010, say 98 per cent of more than 2,000 systems involved in the scheme are working [...]	It exposes a design and construction system for horizontal plates to work as slabs in regular concrete buildings. Based to an evolutionary finite-element analysis of the topological configuration to get a curved design with a 50% reduction of traditional volume, that provide lower cost, less carbon foot-print, better performance and innovative ceiling. A library of profiles is elaborated according [...]

Table 6: Samples from different distribution subsets using EL2N from a 124M reference model trained on CommonCrawl.

Bottom 10%	Middle 10%	Top 10%
a handle on how many elevators they are supposed to oversee. Those officials have repeatedly deflected requests from reporters to detail the count of elevators in Chicago requiring inspection. Frydland, during her interview, said she doesn't know how many elevators her office is responsible for inspecting because city records lump elevators into the same class of devices as escalators, [...]	there's a possibility that you may come across a property that's sharing a driveway with the home next door. That means that one driveway needs to be shared between the two adjoining neighbors. Many real estate investors rent out their properties in order to reap the benefits of passive monthly income while increasing their equity and building wealth over time. Not only are they benefiting [...]	We have all spent happy hours listening to and sharing music we love with those closest to us. Many of the people we serve in ubu are incredibly gifted and play a wide range of musical instruments and enjoy singing and performing for other people. Judith is enabled by ubu to live more independently in Knaresborough, North Yorkshire, and has started taking singing lessons in order to 'grow' her [...]
ians 4:3? Jesus addressed this very issue with his disciples on the night of his betrayal. He would be leaving them soon, but he promised the Holy Spirit would come to comfort and aide them, "I will not leave you as orphans; I will come to you."-John 14:18. Jesus refers to the Holy Spirit as himself because, "the Helper, the Holy Spirit, whom the Father will send in my name, he will teach you all [...]	the standard as far as cement manufacturing goes several cement manufacturers still prefer ball mills for cement production when they want to design new grinding plants or a new integrated 3D design and analysis of the crushing roller of The crushing roller is one of the main parts of a highpressure grinding roller which is a type of highly efficient ore crushing equipment In the work reported [...]	range (Table 1). Active-Controlled Study: CRESTOR was compared with the HMG-CoA reductase inhibitors atorvastatin, simvastatin, and pravastatin in a multicenter, open-label, dose-ranging study of 2,240 patients with Type IIa and IIb hypercholesterolemia. After randomization, patients were treated for 6 weeks with a single daily dose of either CRESTOR, atorvastatin, simvastatin, or pravastatin [...]
Most past attemptsto define socioeconomics as a science in its own right may have been motivated tocounter such a simplistic understanding of socioeconomics.In this chapter, we review past attempts to define socioeconomics before theapproach is chosen that we applied in this book. This book, by a leading expert in urban agriculture, offers a genuine solution to today's global food crisis. By [...]	which adopted our buttons such that when we went to Boston.com (part of NY times) branding was not part of our discussions. Of course, we had matured in our thinking and offered them a co-branded offer hosted by Coola. When Switchboard did not work for us, we went to their competition Infospace.com, which was much larger than them. They accepted a branded Coola button but offered a complex deal [...]	Trend.com: I had no idea this was coming. There'd been talk over the years about setting up a sort of business portal that integrated all of Trend's regular and annual publications, but there was never enough momentum to actually get it going. Trend had a regular spot on the Times' online Business section, but it was a pretty low-impact thing (even though quite a bit of traffic would come to the [...]

Table 7: Samples from different distribution subsets using memorization of a 124M reference model trained on CommonCrawl.

Mem. Factor = 0	Mem. Factor = 0.5	Mem. Factor = 1.0
doesn't prevent you from clearly seeing the road. Hi, thank you so much for your words, appreciate it! Moreover, we noted your comments, we'll think what can be done, for sharing more ideas, feel free to contact us at support@hudwayapp.com any time. Happy to help you always! I do a lot of mudding. And it's got a pitch and roll gauge, which I like when I'm in the hole, do I don't flip my truck. [...]	160 countries. There are abundant hot-selling projects accessible to you. Cheap and environmentally friendly: Factory-direct sale, fast delivery with guaranteed quality at factory price, in line with the concept of environmental development. Feb 19 2021 should pelletisation of sulfide solidelectrolytesafterball millinghas to be done in argon atmosphere question 7 answers i am using a spex 8000b [...]	reference. My company's NACHI 230/600E bearing price concessions, adequate inventory, and other similar products are available for recommendation 1 . Less than 45 KGS, we will send by express. (Door to Door, Convenient) 2 . 45 - 200 KGS , we will send by air transport . (Fastest and safest, but expensive) 3 . More than 200 KGS, we will send by sea . (Cheapest and common use) The bearing 240/8 [...]
disposal and processing of contaminated suspensions such as drilling mud, road sweepings and similar. The rising demand on the international market to meet current as well as future environmental regulations is the main driver for the development in this area of our work," explains Managing Director Ing. Mag. Erich Trunkenpolz. "The plants are currently developed for stationary and semi-mobile du [...]	$97 monthly subscription package. If you decide to make an annual payment of $997, you get two free months. I started with this basic package but I later decided to upgrade to Etison Suite since this one has some limitations. As a marketer, I was only allowed to use 3 custom domains, get a limit of 20,000 visitors, and make a maximum of 100 web pages. I discovered that some advanced features are [...]	takes your bank to process our refund request (5 to 10 business days). If you need to return an item, simply login to your account, view the order using the 'Complete Orders' link under the My Account menu and click the Return Item(s) button. We'll notify you via e-mail of your refund once we've received and processed the returned item. We can ship to virtually any address in the world. Note the [...]
time:If you're looking into faster-than-light fiber internet, there's a Verizon Fios deal for you in Silver Spring, MD. Want more than a Verizon Fios internet-only plan? Open your home up to more entertainment choices with Verizon Fios packages. Ready to improve your home with the best internet available? Get lightspeed internet with Verizon plans that suit every lifestyle. Whether you only need [...]	Select options that apply then copy and paste the RDF/HTML data fragment to include in your application Note: Adjust the width and height settings defined in the RDF/HTML code fragment to best match your requirementsCause.—Upon the ascension of William and Mary to the throne of England, the Protestants of Maryland demanded the Colonial management of the Territory. The Roman Catholics, after rep [...]	to assess the success of our marketing and advertising campaigns). Finally, we may also share your Personal Information to comply with applicable laws and regulations, to respond to a subpoena, search warrant or other lawful request for information we receive, or to otherwise protect our rights. Additionally, you can opt out of some of these services by visiting the Digital Advertising Alliance [...]