Irreducible Curriculum for Language Model Pretraining

Simin Fan
EPFL
simin.fan@epfl.ch

Martin Jaggi
EPFL
martin.jaggi@epfl.ch

Abstract

Automatic data selection and curriculum design for training large language models is challenging, with only a few existing methods showing improvements over standard training. Furthermore, current schemes focus on domain-level selection, overlooking the more fine-grained contributions of each individual training point. It is difficult to apply traditional datapoint selection methods on large language models: most online batch selection methods perform two-times forward or backward passes, which introduces considerable extra costs with large-scale models. To mitigate these obstacles, we propose IRREDUCIBLE CURRICULUM as a curriculum learning algorithm for language model pretraining, which prioritizes samples with higher learnability. Specifically, to avoid prohibitive extra computation overhead, we simulate the sample loss along the main model's training trajectory using a small-scale proxy model. Our experiments on the RedPajama-1B dataset demonstrate a consistent improvement on validation perplexity across all 7 domains compared to random uniform baseline and the anti-curriculum strategy. Our method also reduces the sharpness of the network and illustrates a better 5-shot accuracy on MMLU benchmarks.

1 Introduction

The evolution of language models demonstrates impressive generation and reasoning abilities from a continuous scaling-up of model size and training corpus [Brown et al., 2020, Chowdhery et al., 2022, Kaplan et al., 2020, Hoffmann et al., 2022], along with very significant growth in computation cost. Besides the quantity [Kaplan et al., 2020] and quality [Lee et al., 2022, Longpre et al., 2023] of training data, several recent works show that the ordering [Chen et al., 2023] and composition [Xie et al., 2023] of data during the training can highly impact performance and efficiency. Xie et al. [2023] proposes DOREMI, which determines the optimal domain mixture to construct the pretraining corpus using auxiliary models; Chen et al. [2023] introduces an online selection scheme, which seeks to dynamically update the mixture of data from each skill-set at each training step. While DOREMI and SKILL-IT demonstrates great downstream performance, their curriculum are built upon the group-level where all the instances within one group (domain/skill) shares the same sampling probability, without looking inside each domains on the sample-specific attributions.

Nevertheless, few research have succeeded applying traditional sample-level selection schemes for large langauge model pretraining. Online batch selection methods [Loshchilov and Hutter, 2015, Katharopoulos and Fleuret, 2018, Jiang et al., 2019, Schaul et al., 2015] select hard samples with high loss or high gradient norm, which require a second forward/backward pass. That introduces large extra computation costs when the model size is large, which hurts the scalability. On the other side, Campos [2021] shows that the linguistic-based curriculum learning failed to improve on causal language model pretraining. Schaul et al. [2015] introduces RHO-LOSS, which is an online batch selection scheme based on the gap between current training loss and an irreducible loss term measured using a proxy model. Inspired by the idea of proxy model approximation, we propose IRREDUCIBLE CURRICULUM as a curriculum learning algorithm, construct an ordered data stream by simulating the learning trajectory of large model through the lens of a small-scale proxy model.

ATTRIB: Workshop on Attributing Model Behavior at Scale at NeurIPS 2023.

Specifically, we train the small-scale proxy model on a holdout set, and measure the loss on each data samples in the training set using the proxy model at different training stages. We estimate the *learnability score* of the training samples by the loss gap from the proxy model at early- and late-stage. The IRREDUCIBLE CURRICULUM would prioritize the samples with highest *learnability score* and proceed to the samples with lower *learnability score*. We present the details of our methodology in Section 3.

2 Background

2.1 Reducible Holdout Loss Selection (RHO-LOSS)

We recapitulate RHO-LOSS [Mindermann et al., 2022] with the emphasis on their data selection criterion and approximation using the proxy model. Considering a model $\mathcal{M}(\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, and training corpus $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{n}$, online batch selection pre-samples a larger batch $B_t \in \mathcal{D}_{train}$ uniformly, from which it then selects a smaller subset $b_t \in B_t, |B_t| \gg |b_t|$.

At training step t, given the model has seen part of training corpus $\mathcal{D}_{\text{t}}$, the output distribution from $\mathcal{M}(\boldsymbol{\theta})$ for an input x is denoted as $p(y|x; \mathcal{D}_{\text{t}})$. Therefore, the log-likelihood objective on the holdout evaluation set after adding a new training point (x, y) can be written as $\log \mathrm{p}(y^{\text{ho}}|x^{\text{ho}}; \mathcal{D}_{\text{t}} \cup (x, y))$. Applying Bayesian rule and conditional independence, the notion of the RHO-LOSS can be derived as,

$$
\begin{aligned}
&\log \mathrm{p}(y^{\text{ho}}|x^{\text{ho}}; \mathcal{D}_{\text{t}} \cup (x, y)) \\
&= \log \frac{\mathrm{p}(y|x; x^{\text{ho}}, y^{\text{ho}}, \mathcal{D}_{\text{t}})\, \mathrm{p}(y^{\text{ho}}|x^{\text{ho}}, x; \mathcal{D}_{\text{t}})}{\mathrm{p}(y|x, x^{\text{ho}}; \mathcal{D}_{\text{t}})} \quad \text{Bayes rule} \\
&= \log \frac{\mathrm{p}(y|x; y^{\text{ho}}, x^{\text{ho}}, \mathcal{D}_{\text{t}})\, \mathrm{p}(y^{\text{ho}}|x^{\text{ho}}; \mathcal{D}_{\text{t}})}{\mathrm{p}(y|x; \mathcal{D}_{\text{t}})} \quad \begin{array}{c}\text{conditional}\\ \text{independence}\end{array} \\
&\propto L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{t}}\right] - L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{ho}}, \mathcal{D}_{\text{t}}\right] \\
&\approx L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{t}}\right] - L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{ho}}\right] \qquad (1) \\
&\approx \overbrace{\underbrace{L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{t}}\right]}_{\text{training loss}} - \underbrace{L\left[y|x; \boldsymbol{\theta}', \mathcal{D}_{\text{ho}}\right]}_{\text{irreducible holdout loss (IL)}}}^{\text{reducible holdout loss}} \quad \begin{array}{c}\text{proxy-model}\\ \text{approximation}\end{array} \qquad (2)
\end{aligned}
$$

, where $\boldsymbol{\theta}'$ denotes a smaller scale proxy model pretrained on holdout set $\mathcal{D}_{\text{ho}}$. To avoid retraining the model on $(\mathcal{D}_{\text{t}} \cup \mathcal{D}_{\text{ho}})$ at each step, the second term in Equ. (1) is approximated with the model trained on the holdout set $L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{ho}}, \mathcal{D}_{\text{t}}\right] \approx L\left[y|x; \boldsymbol{\theta}[t], \mathcal{D}_{\text{ho}}\right]$ *(irreducible holdout loss)*, such that the second term could be computed by a pretrained, reusable model. To further improve computational efficiency, the *small-scale proxy model* $\mathcal{M}_{proxy}(\boldsymbol{\theta}')$ is applied to compute the irreducible holdout loss. Mindermann et al. [2022] empirically proved that the approximation can preserve the desired properties with significantly lower computational costs.

Qualitatively, RHO-LOSS down-weighs the *redundant*, *noisy* samples and *outliers*: redundant points have low training loss and low irreducible holdout loss, while noisy samples and outliers have high training loss and high irreducible holdout loss, both lead to a low RHO-LOSS score. While the approximation alleviates the computational burden, the first term in Equ. (2) still requires one forward pass through the proxy model, and two forward passes through the large model at every step: On a pre-sampled batch B_t, the scoring takes one pass through the proxy model and the large model. Then the large model would be trained with a second pass on the top-ranked $|b_t|$ with highest scores. Given that it takes C_1 FLOPs for one forward pass through the large model $\mathcal{M}(\boldsymbol{\theta})$ and C_2 for proxy model $\mathcal{M}_{proxy}(\boldsymbol{\theta}')$. To train the model for T steps on $T \cdot |b_t|$ samples, RHO-LOSS takes $\mathcal{O}(TC_1 \cdot (|B_t| + |b_t|) + TC_2|B_t|)$ forward FLOPs. In case that $|B_t| \gg |b_t|$, the extra computation overhead would be significant.

3 Irreducible Curriculum

Aiming to reduce the computation cost while improving the training efficiency and robustness, we tend to design a pre-defined curriculum to simulate the online assessment of the scores $L[y|x; \mathcal{D}_{\text{t}}]$ along the training process. We refer to our algorithm as IRREDUCIBLE CURRICULUM, since it is

determined only by a proxy-model trained on a holdout dataset without resorting any large model's training dynamics.

Specifically, we train a small proxy model $\mathcal{M}_{proxy}(\boldsymbol{\theta}')$ on holdout set $\mathcal{D}_{\text{ho}}$ for T overall steps while saving an early-stage checkpoint at t_0 steps ($\mathcal{M}_{proxy}(\boldsymbol{\theta}')[t_0]$) and a late-stage checkpoint($\mathcal{M}_{proxy}(\boldsymbol{\theta}')[T]$). We define the *learnability score* for every sample $(x, y) \in \mathcal{D}_{\text{t}}$ of the training set as the gap between the loss from early-stage and the late-stage proxy model checkpoints[1]:

$$Learnability(x) := L_{early} - L_{late} = L\left[x; \boldsymbol{\theta}'[t_0]\right] - L\left[x; \boldsymbol{\theta}'[T]\right] \tag{3}$$

To differentiate better among data points, we take the average of loss from last three checkpoints as the late-stage loss.

The initial threshold of learnability is set as the top-$\lambda_0\%$ instance's score among the training set. The first training batch would be uniformly selected from these top-$\lambda_0\%$ instances whose learnability scores are above the threshold. To simulate the decreasing training loss in online data selection (Equ. 2), the threshold of learnability is decreasing along the training process following a step function. Specifically, consider that we are applying the curriculum on model $\mathcal{M}(\boldsymbol{\theta})$ during the first T_c steps (train for T steps in total), at step $t < T_c$, the training batch for $\mathcal{M}(\boldsymbol{\theta})$ is uniformly sampled from $(\lambda_0 + (1 - \lambda_0)t/T_c)$ instances from the training set with highest learnability. Thus, the sampling probability distribution across training set $\mathcal{D}_{\text{t}}$ is:

$$\begin{aligned} &\mathcal{P}(t) = Unif\left(\{learnability(x) > S(t), x \in \mathcal{D}_{\text{t}}\}\right), \\ &\textit{where } S(t) \textit{ satisfies } \frac{|\mathcal{D}_{\text{t}} > S(t)|}{|\mathcal{D}_{\text{t}}|} = \lambda_0 + (1 - \lambda_0)t/T_c. \end{aligned} \tag{4}$$

After T_c steps, the uniform random sampling would be applied on the whole dataset. Our proposed IRREDUCIBLE CURRICULUM introduces no extra forward passes through the large language model $\mathcal{M}(\boldsymbol{\theta})$ when we pre-compute the learnability score using the proxy model. Instead, the only extra overhead comes from the proxy model as $\mathcal{O}(C_2 \cdot (|D_t|)) \geq \mathcal{O}(T_c C_2 \cdot (|b_t|))$ forward FLOPs, which could be negligible comparing with $\mathcal{O}(TC_1 \cdot (|B_t| + |b_t|) + TC_2|B_t|)$ overhead from RHO-LOSS[2].

4 Experiments and Results

In this section, we present our initial inquiry experimental results on RedPajama-1B dataset, including 7 domains[3]. We randomly split each domain dataset into two subsets $\mathcal{D}_{proxy}$ and $\mathcal{D}_{train}$. $\mathcal{D}_{proxy}$ is used to train the proxy model, which is applied to score $\mathcal{D}_{train}$ and construct the curriculum. Considering a severe unbalance in domain scale, we uniformly sample from each domain, while applying our proposed IRREDUCIBLE CURRICULUM inside each domain. We also provide the results applying the curriculum on the whole dataset in Section 5.

We use a 6-layer (82M) model as the proxy model $\mathcal{M}_{proxy}(\boldsymbol{\theta}')$ to construct the curriculum, which is applied to train a 12-layer (124M) model $\mathcal{M}(\boldsymbol{\theta})$. Both the proxy model and the full size model follow the decoder-only transformer architecture [Vaswani et al., 2023] and share the same set of training hyperparameters (batch size 64, context length 512, learning rate 3e-4). We compare IRREDUCIBLE CURRICULUM (CURRICULUM) with random uniformly sampling baseline (RS) and anti-curriculum (ANTI-CURRICULUM), where we start with a fraction of λ_0 samples with lowest score, and then increase the sampling upper bound with the same rate.

4.1 Result and Analysis

IRREDUCIBLE CURRICULUM consistently improve downstream performance on all domains. We apply the irreducible curriculum across every domain within the RedPajama-1B[4] dataset, with $\lambda_0 = 50\%$, early stage checkpoint is saved at $t_0 = 2000$ steps. We train the proxy model (82M) and full model (124M) for $T = 10'000$ steps. Since the large model would learn faster than the proxy model, we set $T_c = 5000 < T$ to compensate the learning pace. According to Figure 1, IRREDUCIBLE CURRICULUM ***consistently reaches a lower validation perplexity across all 7 domains***

[1] we omit writing y here, as for language modelling the labels y refers to next-token prediction on the current sequence x.

[2] C_1, C_2, B_t, b_t notations are kept consistent with Sec. 2.1

[3] Arxiv, Wikipedia, Book, Common-Crawl, C4, Stackexchange, Github.

[4] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T-Sample

Figure 1: Domain-wise Validation Log-Perplexity on RedPajama-1B. It demonstrates IRREDUCIBLE CURRICULUM consistently outperforms BASELINE and ANTI-CURRICULUM.

Table 1: The table presents results for different methods evaluated on average 5-shot accuracy on MMLU benchmark and average validation log-perplexity across 7 domains in RedPajama-1B. The best results are in **Red** and all the results outperform the baseline are in **Bold**. GPT-3 results refer to Hendrycks et al. [2021b].

METHOD	5-shot MMLU Acc.↑ (%)	Avg. Log Perplexity↓
RS BASELINE	22.9	3.810
GPT-3 (6.7B)	24.9	/
GPT-3 (13B)	26.0	/
GPT-3 (175B)	43.9	/
CURRICULUM ($\lambda_0 = 25\%, T_c = 2000$)	**26.0**	**3.799**
CURRICULUM ($\lambda_0 = 25\%, T_c = 5000$)	**26.9**	**3.802**
CURRICULUM ($\lambda_0 = 25\%, T_c = 10000$)	**23.5**	3.840
CURRICULUM ($\lambda_0 = 50\%, T_c = 2000$)	**23.2**	**3.800**
CURRICULUM ($\lambda_0 = 50\%, T_c = 5000$)	**24.7**	**3.794**
CURRICULUM ($\lambda_0 = 50\%, T_c = 10000$)	**24.7**	**3.800**
ANTI-CURRICULUM ($\lambda_0 = 25\%, T_c = 2000$)	**26.9**	3.819
ANTI-CURRICULUM ($\lambda_0 = 25\%, T_c = 5000$)	24.7	3.838
ANTI-CURRICULUM ($\lambda_0 = 25\%, T_c = 10000$)	**26.9**	3.892
ANTI-CURRICULUM ($\lambda_0 = 50\%, T_c = 2000$)	**26.9**	3.814
ANTI-CURRICULUM ($\lambda_0 = 50\%, T_c = 5000$)	**26.9**	3.827
ANTI-CURRICULUM ($\lambda_0 = 50\%, T_c = 10000$)	**24.9**	3.860

compared with the uniformly random baseline, even trained on a relatively small dataset. Conversely, the ANTI-CURRICULUM approach leads to a deterioration in the evaluation performance across all domains.

We also perform experiments by varying the values of λ_0 and T_c (Table 1). Notably, all the settings applying IRREDUCIBLE CURRICULUM outperform the random uniform sampling baseline, except for the case of ($\lambda_0 = 25\%, T_c = 10'000$). A possible explanation for this is that the step function determined by the proxy model inadequately addresses the differing learning paces between the large and small proxy models. On the contrary, all experiment settings in the ANTI-CURRICULUM group yield worse validation perplexity results.

Additionally, the empirical results shows that both IRREDUCIBLE CURRICULUM and ANTI-CURRICULUM approaches outperform the baseline on the 5-shot inference accuracy on MMLU benchmark [Hendrycks et al., 2021a]. Specifically, ($\lambda_0 = 25\%, T_c = 5000$) works best among all the CURRICULUM settings. Besides, it is worth noting that both CURRICULUM and ANTI-CURRICULUM improve the reasoning capability compared to baseline, which indicates the subset which contributes

the most in perplexity not necessarily contains valuable factual knowledge. How to tackle with this mis-alignment towards our final target could also be a future challenge.

IRREDUCIBLE CURRICULUM reduces sharpness during training. Several recent research shows that applying sharpness aware minimization could effectively improve the generalization ability of the network [Foret et al., 2021]. Thus, we resort to inspect into how the IRREDUCIBLE CURRICULUM could impact the loss landscape by estimating its sharpness along the training trajectory. Following [Wu, 2018, Li et al., 2022b], we adopt the popular definition of sharpness as the largest eigenvalue of the Hessian matrix. We compute the Hessian of all the dense layers and record its top-eigenvalue every 200 steps along the training process. More details on sharpness computation are provided in Appendix A.1.

Figure 2: Network sharpness along training process (RedPajama-1B). The sharpness is measured as the largest eigenvalue of the hessian matrix.

According to Fig. 2, training with IRREDUCIBLE CURRICULUM results to a lower sharpness compared with random baseline. On the opposite, the anti-curriculum would lead to a sharper training trajectory. However, whether and how this lower-sharpness correlates with the generalization ability of the network need further explorations on larger scale models and datasets.

5 Analysis and Failure Mode of Global Curriculum on the Data Mixture

Assuming that we have no access to the domain information during training, we can only apply the curriculum on globally on the entire data mixture. Instead of starting with top-λ_0 fraction of samples in each domains, we rank all samples in the global dataset and start sampling from top-scored λ_0 fraction, which could mostly come from one specific domain.

Here, we compare to the anti-curriculum applied on the whole data mixture, and the baseline with uniformly sampling across all instances. The model scale, architecture, and hyperparameters are the same as in Section 4.1.

Global curriculum lowers average validation perplexity and reduces sharpness. According to Fig. 3, the blue lines denote IRREDUCIBLE CURRICULUM with various combination of λ_0 and T_c[5], while the green lines apply ANTI-CURRICULUM. Across the entire domain-mixture corpus, the proposed IRREDUCIBLE CURRICULUM could still effectively reduce the validation perplexity and sharpness compared with baseline and ANTI-CURRICULUM by a large margin.

Global curriculum is not beneficial for every domain. We further evaluate the domain-wise performance with the global curriculum applied on the entire corpus. According to Fig. 5, IRREDUCIBLE CURRICULUM reduce the validation perplexity on 4 domains (C4, CommonCrawl, Book and Wikipedia), while showing the opposite effect on the other 3 domains (Arxiv, Github and Stackexchange). This can be explained by the domain-wise differences in the learnability score. Figure 4 shows that most of the globally top-ranked samples are from Wikipedia, C4 and CommonCrawl domains, which would be prioritized by the IRREDUCIBLE CURRICULUM while instances from Github and Stackexchange are mostly within the quarter with lowest scores, which are primarily

[5]The combinations are (λ_0, T_c): $(25\%, 2000)$, $(25\%, 5000)$, $(50\%, 2000)$, $(50\%, 5000)$.

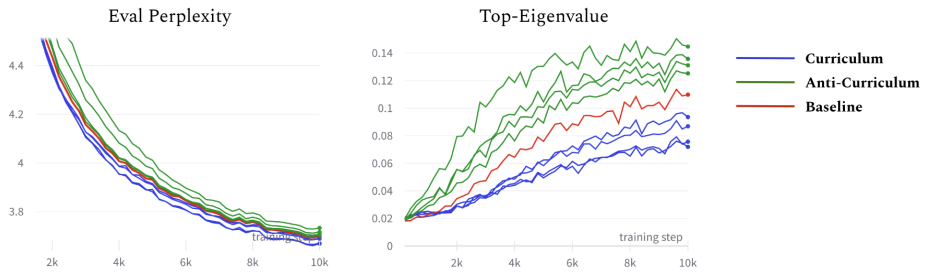

Figure 3: Validation log-perplexity on RedPajama-1B. The validation set is uniformly sampled from all the domains. The Blue lines denotes the IRREDUCIBLE CURRICULUM with different λ_0 and T_c, Green lines denotes the ANTI-CURRICULUM.

learnt by the ANTI-CURRICULUM algorithm. In addition, the in-domain score distributions are highly varying among domains. It illustrates that the an intra-domain learnability (Equ. 3) ranking would offer a more fair comparison, while the naive global curriculum could risk from introducing too much inter-domain sampling bias.

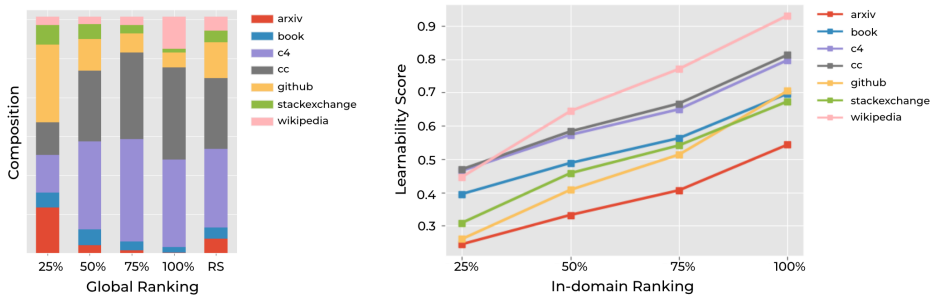

Figure 4: Learnability Scores Distribution. (Left) shows the portion of each domain in each quarter *(0-**25**%, 25%-**50**%, 50%-**75**%, 75%-**100**%)* of global ranking; (Right) shows the average learnability score of each quarter in each domain. It demonstrate that the learnability score are not necessarily comparable between different domains because of various lexical distributions and semantic knowledge.

6 Future work

Scaling up the model and training corpus. Given the emergency ability of large language models, future work is warranted to explore how this IRREDUCIBLE CURRICULUM can improve the performance of larger scale models, in particular with respect to scaling laws relating dataset size and model size.

Relation between sharpness, generalization ability and curriculum learning. In Section 4.1, we observed that the IRREDUCIBLE CURRICULUM could lead to a lower sharpness of the network along the training process while the ANTI-CURRICULUM approach would lead to the opposite. According to the recent debate on the correlation between sharpness and generalization ability of foundation models [Dinh et al., 2017, Foret et al., 2021, Andriushchenko et al., 2023], it is fascinating to uncover whether the curriculum learning algorithm could realize similar outcome as computation-heavy sharpness-aware minimization optimization.

7 Conclusion

In this paper, we propose IRREDUCIBLE CURRICULUM as a curriculum learning algorithm with the underlying hypothesis that small proxy model is capable to simulate the early stage training trajectory. For language model pretraining, the method shows a consistent improvement in validation perplexity across several domains and downstream few-shot inference accuracy on the MMLU benchmark. Besides, we also observe our IRREDUCIBLE CURRICULUM can lead to a lower sharpness compared to the random baseline. However, because of the limitation of computation resources, we leave the scaling-up experiments and exploration of further generalization patterns to future investigations.

References

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Daniel Campos. Curriculum learning for language modeling, 2021.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2021.

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

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

Angela H Jiang, Daniel L-K Wong, Giulio Zhou, David G Andersen, Jeffrey Dean, Gregory R Ganger, Gauri Joshi, Michael Kaminksy, Michael Kozuch, Zachary C Lipton, et al. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*, 2019.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pages 2525–2534. PMLR, 2018.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better, 2022.

Conglong Li, Minjia Zhang, and Yuxiong He. The stability-efficiency dilemma: Investigating sequence length warmup for training gpt models, 2022a.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability, 2022b.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, toxicity, 2023.

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

Sören Mindermann, Jan Brauner, Muhammed 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, 2022.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. Pre-training a BERT with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996, Held Online, September 2021. INCOMA Ltd. URL `https://aclanthology.org/2021.ranlp-1.112`.

Amir Gholami Michael Mahoney Joseph Gonzalez Noah Golmant, Zhewei Yao. pytorch-hessian-eigenthings: efficient pytorch hessian eigendecomposition, October 2018. URL `https://github.com/noahgolmant/pytorch-hessian-eigenthings`.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Lei Wu. How sgd selects the global minima in over-parameterized learning : A dynamical stability perspective. 2018. URL `https://api.semanticscholar.org/CorpusID:54031709`.

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, 2023.

A Appendix

A.1 Preliminary of Hessian and Sharpness

Hessian. Considering model parameter θ and input x_i, the hessian is defined as the average second-order derivative of the objective function ℓ:

$$\nabla^2_\theta \mathcal{L}(f(\theta)) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2_\theta \ell(f(\theta, \mathbf{x}_i)) \tag{5}$$

Specifically, we randomly sample 50 instances from validation set of each domain, and take the average second-order derivative as the hessian matrix. We only consider parameters from all dense layers as θ while skip the attention matrix.

Sharpness. Denote the hessian matrix as $\mathcal{H}$, we adopt a popular definition of the sharpness as the largest eigenvalue of the Hessian $\lambda_{\max}(\mathcal{H})$, following [Wu, 2018, Li et al., 2022b].

When calculating the sharpness, we apply the numerical method, power method, following [Noah Golmant, 2018] to calculate the top eigenvalue of the Hessian matrix.

A.2 Related Work

Curriculum Learning for Language Modeling. As a popular concept in education and cognitive science, there have been plenty of attempts to apply curriculum learning on training language models, by progressively expose the neural learner with samples with ascending difficulty level [Campos, 2021, Nagatsuka et al., 2021, Li et al., 2022a]. However, most of the effective methods increase the difficulty by varying the context length of sequences, without focusing on the syntactic nor semantic properties. Li et al. [2022a] empirically proves that with a incrementally increase context length, the pretraining of language model could be more stable, reducing the number of loss spikes. Similarly, Nagatsuka et al. [2021] shows that feed training samples with increasing block size could improve the performance when pretraining BERT [Devlin et al., 2019], an encoder-only model with masked language modeling objective.

Campos [2021] has conducted extensive experiments towards using linguistic-featured curriculum learning for language model pretraining. However, the conclusion shows their curriculum fails to effectively help with validation perplexity nor downstream performance.

Recently, Xie et al. [2023] has introduced a domain reweighting framework by sampling from domains with various probability. They propose to use two small-scale auxiliary models to determine the domain weights through a group distributional robust optimization game. However, all the samples within one domain would share the same sampling weights, which fails to distinguish the instance-wise attributes inside the domain barrier. How to combine the domain-level reweighting and the in-domain curriculum could also be a promising direction for future exploration.

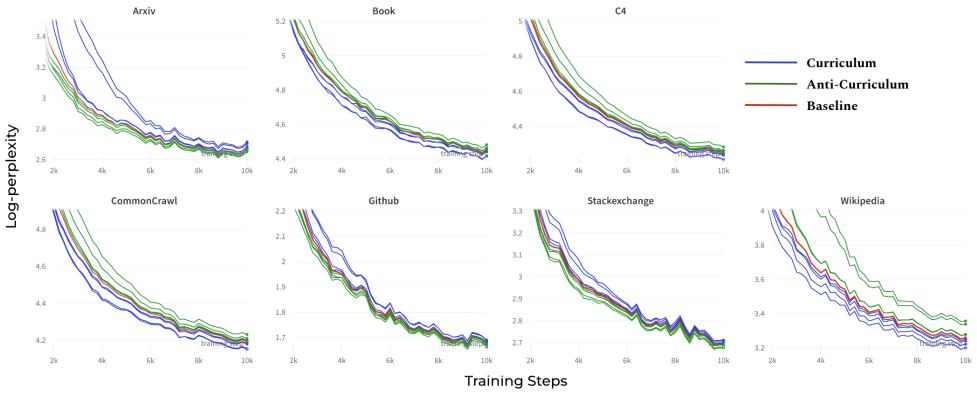

Figure 5: Domain-wise validation log-perplexity on RedPajama-1B.

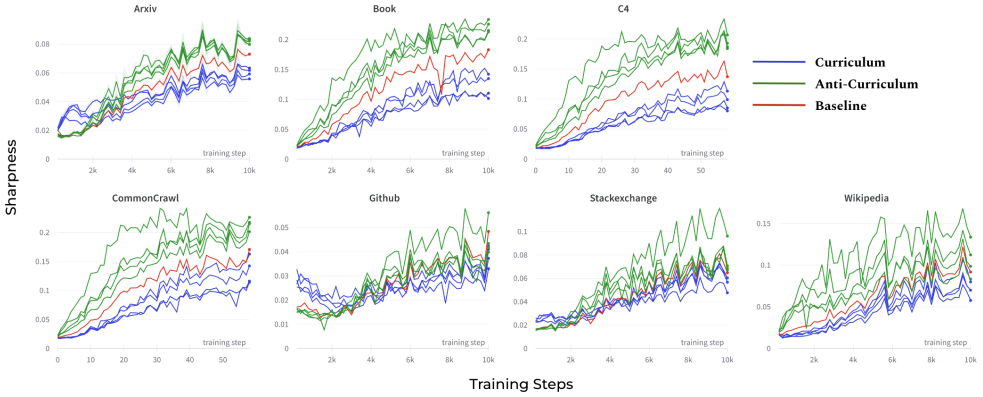

Figure 6: Network sharpness on RedPajama-1B.