Understanding Emergent Abilities of Language Models from the Loss Perspective

Zhengxiao Du[1,2], **Aohan Zeng**[1,2], **Yuxiao Dong**[2], **Jie Tang**[2]
[1]Zhipu AI [2]Tsinghua University
{zx-du20,zah22}@mails.tsinghua.edu.cn

Abstract

Recent studies have put into question the belief that emergent abilities [58] in language models are exclusive to large models. This skepticism arises from two observations: 1) smaller models can also exhibit high performance on emergent abilities and 2) there is doubt on the discontinuous metrics used to measure these abilities. In this paper, we propose to study emergent abilities in the lens of pre-training loss, instead of model size or training compute. We demonstrate that the Transformer models with the same pre-training loss, but different model and data sizes, generate the same performance on various downstream tasks, with a fixed data corpus, tokenization, and model architecture. We also discover that a model exhibits emergent abilities on certain tasks—regardless of the continuity of metrics—when its pre-training loss falls below a specific threshold. Before reaching this threshold, its performance remains at the level of random guessing. This inspires us to redefine emergent abilities as those that manifest in models with lower pre-training losses, highlighting that these abilities cannot be predicted by merely extrapolating the performance trends of models with higher pre-training losses.

1 Introduction

Scaling of language modes (LMs) on both model and data sizes has been shown to be effective for improving the performance on a wide range of tasks [42, 6, 23, 8, 65, 55, 36], leading to the widespread adoption of LM applications, e.g., ChatGPT. The success of such scaling is guided by scaling laws [22, 28, 10, 23], which study the predictability of pre-training loss given the model and data sizes.

While scaling laws focus on the pre-training loss, the scaling effect on the performance of downstream tasks has thus far less studied. Emergent abilities [58] are defined as abilities that present in larger LMs but not present in smaller one. The existence of such abilities is recently challenged for two reasons. First, small LMs trained on a sufficient amount of high-quality data can outperform large models on tasks with claimed emergent abilities [55, 56, 26]. For example, LLaMA-13B with less compute [55] can outperform GPT-3 (175B) on MMLU [21], due to more training tokens and improved data-filtering methods. Second, Schaeffer et al. [46] claim that emergent abilities appear due to the nonlinear or discontinuous metrics selected to evaluate certain datasets, rather than from a fundamental change in larger models.

The Chinchilla scaling laws [23] show that different combinations of model sizes and data sizes can lead to different pre-training losses even with the same training compute. Consequently, the pre-training loss can naturally better represent the learning status of LMs than the model or data sizes. However, the relationship between the loss of an LM and its performance on downstream tasks is not yet well understood. Existing literature has either focused on the transfer learning paradigm [33, 54] or constrained its study to single models, tasks, or prompting methods [49, 61].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1683

In this work, we propose to study emergent abilities from the perspective of pre-training loss instead of model size or training compute. To examine the relationship between the pre-training loss of LMs and their performance, we pre-train more than 30 LMs of varied model and data sizes from scratch, using a fixed data corpus, tokenization, and model architecture. Their downstream performance is evaluated on 12 diverse datasets covering different tasks, languages, prompting types, and answer forms. We demonstrate that the pre-training loss of an LM is predictive of its performance on downstream tasks, regardless of its model size or data size. The generality of this conclusion is further verified by extracting and observing the performance and loss relationship of the open LLaMA [55] and Pythia [3] models.

Over the course, we find that performance on certain downstream tasks only improves beyond the level of random guessing when the pre-training loss falls below a specific threshold, i.e., emergent abilities. Interestingly, the loss thresholds for these tasks are the same. When the loss is above this threshold, performance remains at the level of random guessing, even though performance on other tasks continues to improve from the outset. To exclude the impact of discontinuous metrics [46, 61], we evaluate the emergent performance increase using continuous metrics and show that the emergent abilities persist across both discontinuous and continuous metrics.

Based on these observations, we define the emergent abilities of LMs from the perspective of pre-training loss: an ability is emergent if it is not present in language models with higher pre-training loss, but is present in language models with lower pre-training loss. According to the loss scaling laws [22, 28], the pre-training loss is a function of model size, data size, and training compute. Therefore, the new emergent abilities can also account for the previously-observed emergent abilities in terms of model size or training compute.

The advantage of the new definition lies in its ability to better capture the tipping points in training trajectories when LMs acquire emergent abilities. Once again [58], the existence of emergent abilities suggests that we cannot predict all the abilities of LMs by simply extrapolating the performance of LMs with higher pre-training loss. Further scaling the model and data size to lower the pre-training loss may enable new abilities that were not present in previous LMs.

2 Does Pre-training Loss Predict Task Performance?

Table 1: English and Chinese datasets evaluated in the experiment, and their task types, prompting types, answer forms and metrics. For prompting type, we refer to the chain-of-thought prompting [59] as few-shot CoT and the original in-context learning prompting [6] as few-shot.

Dataset	**Task**	**Prompting Type**	**Answer Form**	**Metric**
		English datasets		
TriviaQA [27]	Closed-book QA	Few-shot	Open-formed	EM
HellaSwag [64]	Commonsense NLI	Zero-shot	Mulit-choice	Accuracy
RACE [31]	Reading Comprehension	Few-shot	Multi-choice	Accuracy
WinoGrande [44]	Coreference Resolution	Zero-shot	Multi-choice	Accuracy
MMLU [21]	Examination	Few-shot	Multi-choice	Accuracy
GSM8K [12]	Math Word Problem	Few-shot CoT	Open-formed	EM
		Chinese datasets		
NLPCC-KBQA[15]	Closed-book QA	Few-shot	Open-formed	EM
ClozeT [63]	Commonsense NLI	Zero-shot	Multi-choice	Accuracy
CLUEWSC [62]	Coreference Resolution	Zero-shot	Multi-choice	Accuracy
C3 [52]	Reading Comprehension	Few-shot	Multi-choice	Accuracy
C-Eval [25]	Examination	Few-shot	Multi-choice	Accuracy
GSM8K-Chinese	Math Word Problem	Few-shot CoT	Open-formed	EM

We study the relationship between the performance of the language models (LMs) on 12 downstream tasks and the pre-training loss. We pre-train LMs of different model sizes (300M, 540M, 1B, 1.5B, 3B, 6B, and 32B) on varied numbers of tokens with fixed data corpus, tokenization, and architecture. In addition, we leverage the open LLaMA [55] models (7B, 13B, 33B, and 65B) to validate our observations.

It is not straightforward that the loss of LMs decides the performance on downstream tasks. Generally the performance is decided by the probability to predict the ground truth y given the prompt x, i.e. $p(y|x)$. The probability can be written as a function of the cross entropy loss:

$$p(y|x) = \exp(-\ell(y|x)) \tag{1}$$

where $\ell(y|x)$ is the cross entropy loss of the LM given the context x and the target y. While $\ell(y|x)$ has the same form as the pre-training loss L, they are not equal. First, the pre-training loss is an average of all the tokens in all the documents pre-trained on. According to our empirical observation, the losses of different documents are not uniform. Second, if x and similar documents do not exist in the pre-training corpus, $\ell(y|x)$ is the generalization loss, which is often related to other factors beyond the training loss, such as the model size. For example, in computer vision, a highly over-parameterized models often improve over an under-parameterized models in test performance when both models converge on the training data [14, 7].

2.1 Pre-training Setting

All the models are pre-trained on a mixture of English and Chinese corpus. The ratio of English to Chinese is 4:1 in the pre-training corpus. The model architecture is similar to LLaMA [55] with two differences: we use grouped-query attention [1] to replace the multi-query attention and we apply rotary position embedding on half the dimensions of the query and key vectors. More details can be found in Appendix A.

2.2 Evaluation Tasks

To present a comprehensive demonstration, we evaluate the pre-trained models on 12 datasets across different tasks and prompting types in both English and Chinese. The six task types include:

Closed-book QA: Answering questions about the real world based solely on the pretrained knowledge. We use TriviaQA [31] for English. For Chinese, we build a closed-book QA dataset based on NLPCC-KBQA [15] dataset following the TriviaQA format.

Commonsense Natural Language Inference (NLI): Selecting the most likely followup given an event description. We use the HellaSwag dataset [64] for English and the ClozeT dataset in Yao et al. [63] for Chinese.

Reading comprehension: Reading a given article or paragraph and answering questions about it. We use RACE [31] for English and C3 [52] for Chinese. Both are based on multi-choice questions.

Coreference Resolution: Given a sentence with pronouns, determine which pronoun refers to which entity. We use the WinoGrande dataset [44] for English and the CLUEWSC dataset [62] for Chinese.

Examination: Multiple-choice questions in examinations. For English, we use MMLU [21], which includes mathematics, US history, computer science, law, and more. For Chinese, we use C-Eval [25] which ranges from humanities to science and engineering.

Math Word Problem: Solving real-life, situational and relevant problems using mathematical concepts. For English we use the GSM8K [12] dataset. For Chinese, we translate the questions and answers in GSM8K to Chinese, namely GSM8K-Chinese.

The prompting types cover few-shot [6], zero-shot, and few-shot chain-of-thought (CoT) [59]. The datasets are summarized in Table 1.

2.3 Pre-training Loss vs. Performance

In the first experiment, we train three models with 1.5B, 6B, and 32B parameters and observe their behaviors until trained on 3T, 3T, and 2.5T tokens, respectively. The training hyperparameters are shown in Table 4 (Appendix).

We evaluate the performance of intermediate training checkpoints. The checkpoints are saved around every 43B tokens during pre-training. We plot the points of task performance (y-axis) and training

 https://doi.org/10.52202/079017-1683

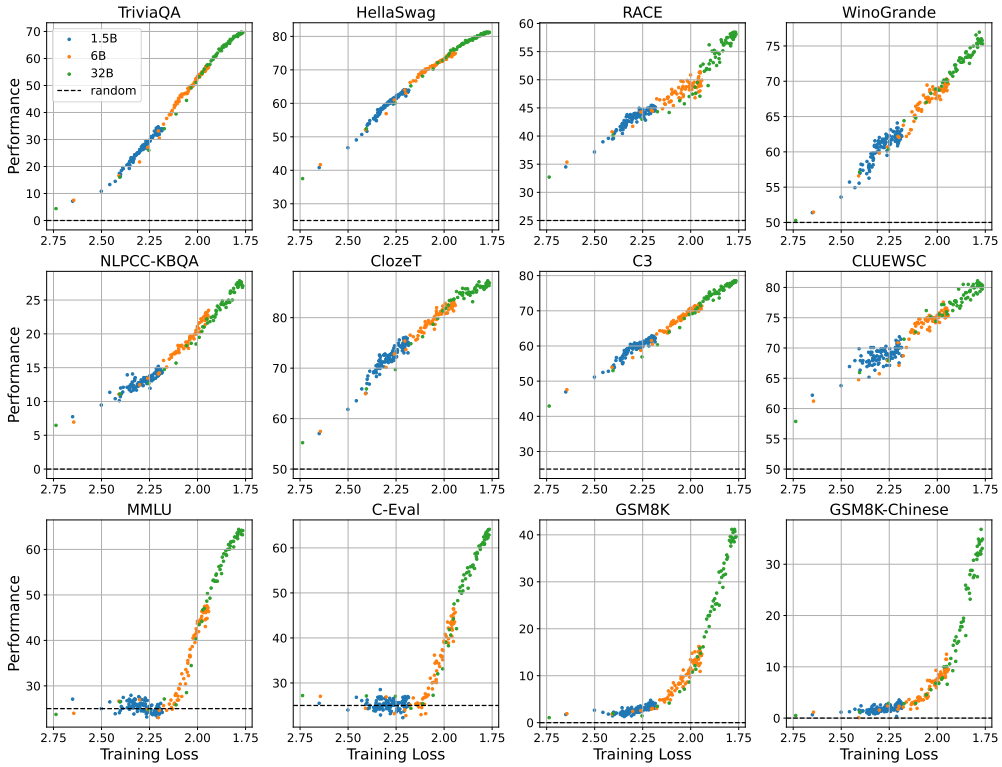

Figure 1: **The performance-vs-loss curves of 1.5B, 6B, and 32B models.** Each data point is the loss (x-axis) and performance (y-axis) of the intermediate checkpoint of one of the three models. We mark the results of random guess in black dashed lines.

Table 2: Statistical measures of the correlation between task performance and pre-training loss in Figure 1. The spearman correlation coefficient [50] measures the monotonicity of the relationship between the two variables, and the pearson correlation coefficient measures the linearity of the relationship. Both vary between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic/linear relationship.

Dataset	TriQA	HS	RACE	WG	NQA	ClozeT	C3	CW	MMLU	CE	GSM	GSMC
Spearman	-0.996	-0.996	-0.977	-0.978	-0.984	-0.986	-0.988	-0.947	-0.804	-0.831	-0.975	-0.948
Pearson	-0.994	-0.994	-0.963	-0.988	-0.982	-0.985	-0.993	-0.972	-0.903	-0.884	-0.874	-0.829

loss (x-axis) in Figure 1, and provide the statistical measures of the two variables in Table 2. From the curves and statistics, we can see that the training loss is a good predictor of the performance on 12 downstream tasks.

- Generally, the task performance improves as the training loss goes down, regardless of the model sizes. On MMLU, C-Eval, GSM8K, and GSM8K-Chinese, all models of three sizes perform at the random level until the pre-training loss decreases to about 2.2, after which the performance gradually climbs as the loss decreases. Detailed analysis on this is shown in Section 3.
- Importantly, the performance-vs-loss data points of different model sizes fall on the same trending curve. That is, by ignoring the color differences (model sizes), the data points of different models are indistinguishable. For example, when the training loss falls around 2.00, the green and orange points on TriviaQA are indistinguishable. This indicates that the model performance on downstream tasks largely correlates with the pre-training loss, *regardless of the model size.*
- Both spearman and pearson correlation coefficients show that performance is strongly related to pre-training loss on TriviaQA, HellaSwag, RACE, WinoGrande, etc. The pearson correlation coefficients on these tasks specifically show that points from different models lie on the same trending curve. The relationship is weaker on MMLU, CEval, GSM8K, and GSM8K-Chinese, verifying the emergence of performance which we discuss in Section 3.

- Interestingly, we find that the overall training loss is a good predictor of performance on both English and Chinese tasks, although it is computed on a mixture of English and Chinese tokens. This implies that the learning dynamics of English and Chinese tokens are likely very similar during multilingual pre-training.

2.4 Training Token Count vs. Performance

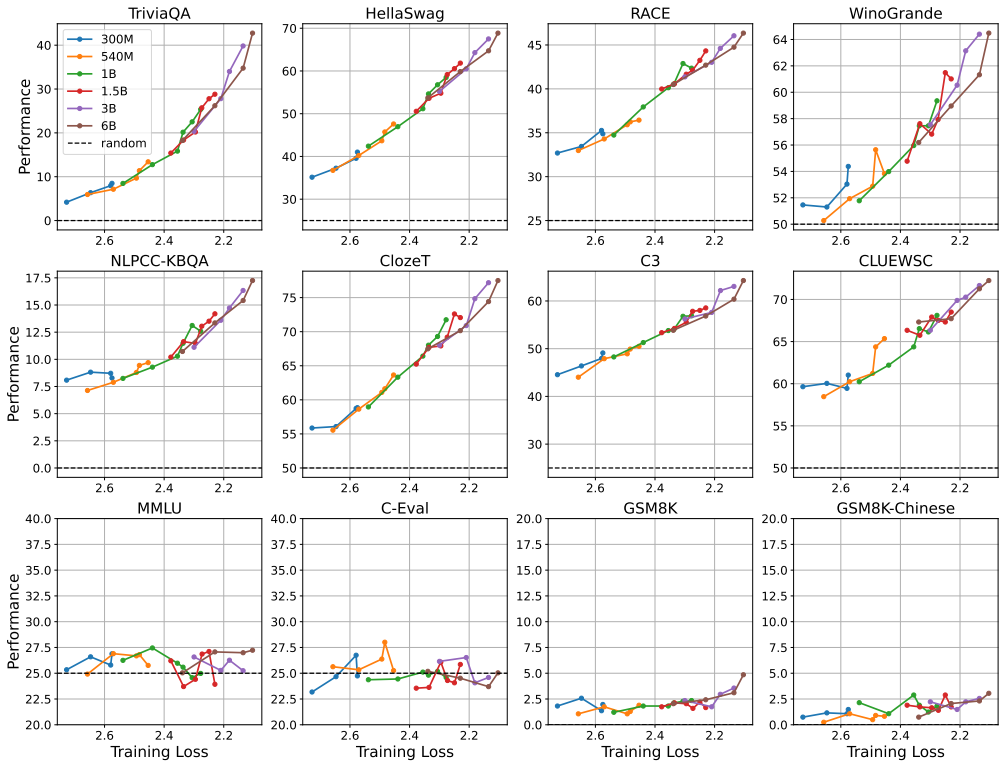

Figure 2: **The performance-vs-loss curves of smaller models pre-trained with different numbers of training tokens**. Each data point is the loss (x-axis) and performance (y-axis) of the final checkpoint of one model, i.e., each point corresponds to one model trained from scratch. We mark the results of random guess in black dashed lines.

Following the empirical experiments in scaling laws [22, 28, 23], we further pre-train 28 relatively smaller models with different numbers of training tokens. The model sizes range from 300M, to 540M, 1B, 1.5B, 3B, and to 6B, while the numbers of pre-training tokens range from 33B to 500B. Varying the number of pre-training tokens is necessary since to achieve optimal performance we need to set the cosine learning rate schedule to reach the minimum at the corresponding token count [28, 23]. The number of tokens used and hyperparameters for all models are shown in Table 5 (Appendix).

On each line, each data point represents the performance and pre-training loss of the corresponding model pre-trained completely from scratch with the certain token count (and learning rate schedule). We can see that similar to the observations from Figure 1, the data points of different models sizes and training tokens largely fall on the same trending curves. In other words, *the LMs with the same pre-training loss regardless of token count and model size exhibit the same performance on the 12 downstream tasks.*

Another similar observation is that the performance curves on MMLU, C-Eval, GSM8K, and GSM8K-Chinese do not yield an uptrend, meaning that the performance of these models on these four tasks are close to random (with fewer than 500B tokens). For simplicity, we only plot the performance of the latest checkpoint in each training in Figure 2. The complete performance curves with intermediate checkpoints of each model, in which we can observe the same trend but larger variance, are shown in Figure 5 (Appendix).

https://doi.org/10.52202/079017-1683

2.5 LLaMA's Loss vs. Performance

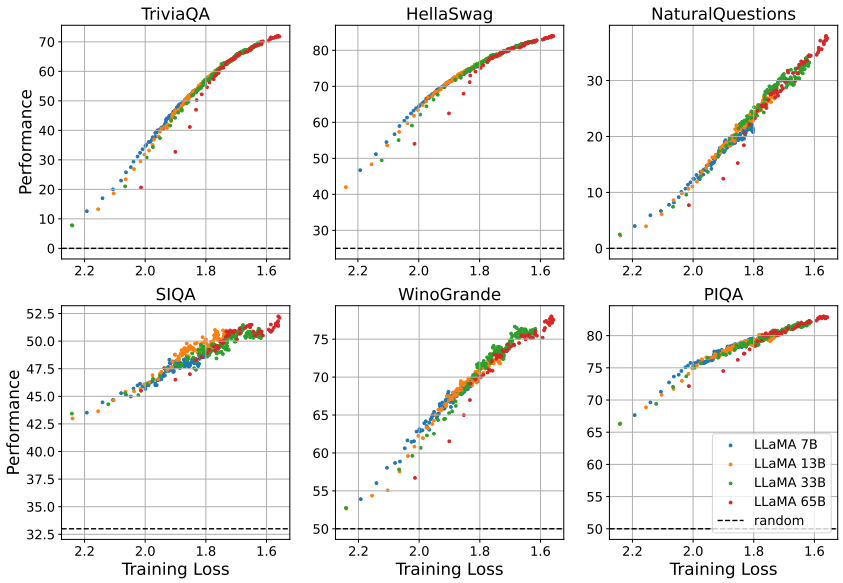

Figure 3: **The performance-vs-loss curves of LLaMA.** The values of performance and training loss are extracted from the figures in the original LLaMA paper [55]. Note that the LLaMA2 paper [56] does not cover such figures with related information.

To validate the generality of our observations, we analyze two different model series with required information made publicly available, i.e., LLaMA [55] and Pythia [3]. Compared to our models, LLaMA uses a pre-training corpus that excludes Chinese documents, leverages a different pre-training framework [37], and adopts a slightly different model architecture. Since the intermediate checkpoints of LLaMA are not available, we extract the pre-training loss and corresponding performance on six question answering and commonsense reasoning tasks from the figures in its original paper, and plot the points in Figure 3.

Excitingly, most data points from the LLaMA models with different sizes (7B, 13B, 33B, 65B) fall on the same upwards trend. This observation further confirm our conclusion that the model's pre-training loss can predict its performance on downstream tasks, regardless of model size and token count. Note that there is only one exception at the early stage of LLaMA-65B. We can see that when the training loss is higher than 1.8, LLaMA-65B performs worse than smaller models with the same training loss. Without access to its intermediate checkpoints, we unfortunately cannot further analyze the result. One possible explanation is that they use exponential smoothing on either the loss or downstream performance plots. Exponential smoothing would perturb the earlier points more than other points, potentially leading to this effect. Note that the outliers only constitute the initial 10% training tokens. The results for Pythia are shown in Appendix F, which also support our conclusion.

Observed from previous experiments and analysis, we can conclude that the pre-training loss is a good indicator of LMs' performance on downstream tasks, independent of model sizes, training tokens, languages, and pre-training frameworks.

3 Analysis of Different Tasks and Metrics

3.1 Performance Trends of Different Tasks

In Figures 1 and 2, we can separate the datasets into two groups: First, on TriviaQA, HellaSwag, RACE, WinoGrande, NLPCC-KBQA, ClozeT, CLUEWSC, and C3, the performance improves smoothly with decreased pre-training loss from the very beginning. Second, on MMLU, C-Eval, GSM8K, and GSM8K-Chinese, the performance remains flat when the loss is higher than a certain threshold. Once the pre-training loss is lower than this threshold, the performance starts to improve.

https://doi.org/10.52202/079017-1683

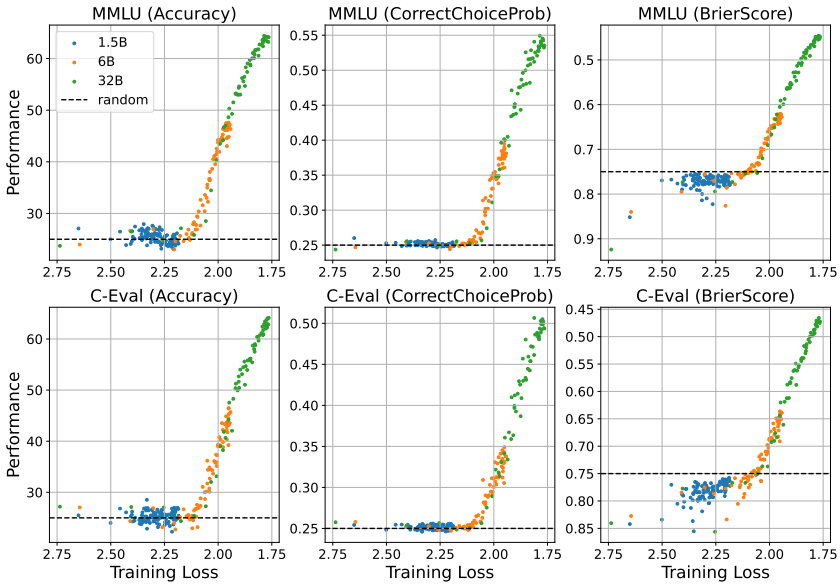

Figure 4: **The performance-vs-loss curves of different metrics on MMLU and C-Eval.** Accuracy: discontinuous; CorrectChoiceProb and BrierScore: continuous. We mark the result of random guess in black dashed lines.

The correlation coefficients in Table 2 also reveal the difference: coefficients in the first group are close to -1, indicating strong correlations, while correlations in the second group are weaker.

Take MMLU as an example of the second group, when the pre-training loss is higher than 2.2, the accuracy remains around 25%. Since each question in MMLU has four options, this means the model prediction is no better than random guessing. However, when the pre-training loss drops below 2.2, the accuracy increases as the loss decreases, similar to the trend observed in the first group of tasks. The performance trends of C-Eval, GSM8K, and GSM8K-Chinese follow a similar pattern. Despite differences in languages, tasks, prompting types, and answer forms among the four datasets are different, the thresholds for performance improvement are surprisingly all around 2.2.

RACE in the first group has a prompting format similar to MMLU: both consist of multi-choice examination questions with in-context demonstrations, but their performance curves are quite different. We hypothesis that it is the task difficulty that makes the difference. Tasks of the first group of datasets are easier than those of the second group. For example, RACE requires the model to select correct answers for questions about a given article, and HellaSwag lets the model to select the possible followup of a situation based on commonsense. In contrast, MMLU and C-Eval consist of questions designed for high school, college, or professional examinations, requiring a broader range of knowledge. GSM8K and GSM8K-Chinese are math word problems that were previously considered as impossible to be solved by pre-trained language models before Chain-of-Thought prompting.

The phenomenon can be related to grokking, which describes the improvement of performance from the random chance level to perfect generalization [40]. Power et al. [40] find that this improvement can occur well past the point of overfitting. In pre-training, the models are usually underfitting instead of overfitting overall. Since the pre-training corpus is a mixture of different documents, it is possible that the model already fits some patterns—such as numerical addition—in the data, while still underfitting the overall corpus.

Certainly, the observations on the second groups of datasets can also be related to emergent abilities [58], that is, abilities that only present in large models. According to the scaling law [28], with the number of training tokens fixed, the pre-training loss follows a power law with respect to model sizes. In other words, there is a monotonic relationship between model size and pre-training loss. For the second group of tasks, there is a threshold of model sizes that corresponds to the tipping point in the pre-training loss. When the model size exceeds this threshold, the model can exhibit performance above the random chance level.

https://doi.org/10.52202/079017-1683

3.2 Influence of Different Metrics

Schaeffer et al. [46] propose an alternative explanation of emergent abilities of LMs, that is, emergent abilities appear due to the researchers' choice of nonlinear or discontinuous metrics. The accuracy on multi-choice questions (e.g., MMLU) is discontinuous, since the score on a question is either 1 or 0. To validate this claim, we examine the intermediate checkpoints on MMLU and C-Eval with continuous metrics rather than discontinuous accuracy used in the original benchmarks. The first metric is the predicted probability of the correct answer, denoted as CorrectChoiceProb. The second one is the Brier Score [5] used in Schaeffer et al. [46]:

$$\text{BrierScore} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij} - \hat{y}_{ij})^2 \tag{2}$$

where $\hat{y}_{ij}$ is the predicted probability of sample i for class j and y_{ij} is the ground truth probability. The metric measures the prediction error and a lower value indicates better performance.

We plot the results measured by different metrics on MMLU and C-Eval in Figure 4. All three metrics—accuracy, correct choice probability, and Brier Score—show emergent performance improvements (value increase for the first two and decrease for the third) when the pre-training loss drops below a certain threshold. The Brier Score also decreases when the pre-training loss is above the threshold. However, the decrease of Brier Score does not always represent improvements on the task, since the Brier Score is related to not only the predicted probability of the correct answer but also the predicted probabilities of the incorrect answers. We find that the distribution of the correct answers is uniform in the four options in MMLU and C-Eval. The best Brier Score for a context-free predictor is achieved by always giving uniform probability to all the options. In this case, the Brier Score is equal to 0.75. Therefore, the performance in terms of Brier Score is no better than random guess before the loss reaches the threshold. This observation further confirms our previous conclusion. We discuss the contrary observations of Schaeffer et al. [46] and Xia et al. [61] in Appendix C.

We conclude that emergent abilities of language models occur when the pre-training loss reaches a certain tipping point, and continuous metrics cannot eliminate the observed tipping point.

4 Defining Emergent Abilities from the Loss Perspective

In previous sections, we show that 1) the pre-training loss is predictive of the performance of language modes on downstream tasks, and 2) some tasks exhibit emergent performance improvements from the random guess level when the pre-training loss drops below a certain threshold regardless of model size, token count, and continuity of metrics. Based on these observations, we give a new definition of emergent abilities from the pre-training loss perspective:

Definition. *An ability is emergent if it is not present in models with higher pre-training loss but is present in models with lower pre-training loss.*

The normalized performance on an emergent ability as a function of the pre-training loss L is:

$$\begin{cases} f(L) & \text{if } L < \eta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $f(L)$ is a monotonically decreasing function of L, η is the threshold, and the normalized performance of random guess is 0.

Next we will show how the new definition can be related to previously observed emergent abilities [58]. In Henighan et al. [22], they give the scaling relation for the loss with model size N when the number of training tokens D is fixed:

$$L(N) = L_\infty + \left(\frac{N_0}{N}\right)^{\alpha_N} \tag{4}$$

where L_∞ is the irreducible loss, and α_N is the coefficient. The equation shows that the loss of language models follows a power-law plus a constant. Combining Equation (3) and Equation (4), we can get the normalized performance as a function of the model size N

$$\begin{cases} f\left(L_\infty + \left(\frac{N_0}{N}\right)^{\alpha_N}\right) & \text{if } N \geq N_0 \cdot (\eta - L_\infty)^{-\frac{1}{\alpha_N}} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

From this equation, we can explain the emergent abilities observed in Wei et al. [58]: when model sizes are smaller than $N_0 \cdot (\eta - L_\infty)^{-1/\alpha_N}$, the normalized performance is zero. When model sizes exceed $N_0 \cdot (\eta - L_\infty)^{-1/\alpha_N}$, the increase in model size leads to a decrease of pre-training loss and an improvement in normalized performance.

5 Related Work

Relationship of Pre-training Loss and Task Performance. In the transfer learning setting, i.e. the language model is pre-trained on the general corpus and fine-tuned on supervised data of specific tasks, Tay et al. [54] find that models with the same pre-training loss can have different downstream performance after finetuning, due to inductive bias in model architectures such as Transformers and Switch Transformers. Tay et al. [53] further study the effect of model shapes on downstream fine-tuning. Liu et al. [33] also study the effect of inductive bias of model sizes and model algorithms on the relationship of pre-training loss and downstream performance after fine-tuning, but their theory only applies in the saturation regime, where the models are close to minimal possible pre-training loss. Instead, large language models today are generally under-trained [23, 55], far from the saturation regime. Overall, these studies focus on the pretrain-finetune paradigm, in which inducitive bias helps improve transferability, while we study prompted performance of large language models without finetuning [29, 6]. For the prompted performance of large language models, Xia et al. [61] claim that perplexity is a strong predictor of in-context learning performance, but the evidence is limited to the OPT model [66] and a subset of BIG-Bench [51]. Instead, Shin et al. [49] find that low perplexity does not always imply high in-context learning performance when the pre-training corpus changes. Gadre et al. [18] fits the relation of perplexity and the top-1 error averaged over many natural language tasks with a power law. Instead, we focus on the different relations of tasks and a small part of tasks that show emergency trends.

Emergent abilities. Wei et al. [58] propose the idea of emergent abilities, abilities that only present in large language models. This is similar to the claim of Ganguli et al. [19] that it is more difficult to predict the capacities of language models than to predict the pre-training loss. The existence of emergent abilities has been challenged. Hoffmann et al. [23] show that smaller language models trained with sufficient data can outperform undertrained larger language models, supported by follow-up models [55, 26, 56]. On the other hand, Schaeffer et al. [46] claim that emergent abilities are due to the discontinuous metrics used for evaluation, also found in Xia et al. [61]. Similarly, Hu et al. [24] propose to predict the performance of emergent abilities with the infinite resolution evaluation metric. In this paper we prove the existence of emergent abilities from the perspective of pre-training loss, even with continuous metrics.

6 Conclusion

Our paper proposes a new definition of emergent abilities of language models from the perspective of pre-training loss. Empirical results show that the pre-training loss is a better metric to represent the scaling effect of language models than model size or training compute. The performance of emergent abilities exhibits emergent increase when the pre-training loss falls below a certain threshold, even when evaluated with continuous metrics.

The new definition offers a precise characterization of the critical junctures within training trajectories where emergent abilities manifest. It encourages future studies to investigate the shifts in language models at these junctures, which facilitate the development of new capabilities.

7 Limitation

We study the relationship of pre-training loss and task performance across model sizes, training tokens, tasks, languages, prompting types, and answer forms. Factors we have not considered are model architectures and training algorithms. We analyze the performance-loss curves of LLaMA and Pythia with slightly different architectures, and find that the relationship holds for all the models. But there are fundamentally different model architectures, such as routed Transformers [16], and

 https://doi.org/10.52202/079017-1683

non-Transformer architectures [17, 39] beyond our consideration. Both our models and LLaMA use AdamW optimizer [35], while there are other optimizers for language model pre-training [48, 32].

The disadvantage of studying emergent abilities in the lens of pre-training loss is that the pre-training loss is affected by the tokenizer and the distribution of pre-training corpus. The values of pre-training loss of language models trained on different corpus are not directly comparable. One possible solution is to evaluate different language models on a public validation set with the normalized perplexity [43] to account for the different vocabulary sizes.

The paper should not be considered as a push to expand model sizes and data sizes of language models beyond current scales. It is not guaranteed that new tipping points emerge in larger scales. Also, instruction tuning [57, 45, 9, 34] can improve the zero-shot performance of language models on unseen tasks, including MMLU and GSM8K.

Acknowledgments and Disclosure of Funding

This work is supported by the Natural Science Foundation of China NSFC 62425601 and 62276148, a research fund from Zhipu, New Cornerstone Science Foundation through the XPLORER PRIZE and Tsinghua University (Department of Computer Science and Technology)-Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things (JCIIOT). Corresponding authors: Yuxiao Dong and Jie Tang.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4895–4901. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.298.

[2] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

[3] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 2023.

[4] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

[5] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

[6] 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. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[7] Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3349–3356. AAAI Press, 2020. doi: 10.1609/AAAI.V34I04.5736. URL https://doi.org/10.1609/aaai.v34i04.5736.

[8] 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. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

[10] Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake A. Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J. Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc'Aurelio Ranzato, Jack W. Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4057–4086. PMLR, 2022. URL https://proceedings.mlr.press/v162/clark22a.html.

[11] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

[13] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, April 2023. URL https://github.com/togethercomputer/RedPajama-Data.

[14] Yehuda Dar, Vidya Muthukumar, and Richard G. Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *CoRR*, abs/2109.02355, 2021. URL https://arxiv.org/abs/2109.02355.

[15] Nan Duan. Overview of the nlpcc-iccpol 2016 shared task: Open domain chinese question answering. In *Natural Language Understanding and Intelligent Applications*, pages 942–948. Springer International Publishing, 2016. ISBN 978-3-319-50496-4.

[16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

[17] Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=COZDy0WYGg.

[18] Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Alexandros G. Dimakis, Gabriel Ilharco, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. *CoRR*, abs/2403.08540, 2024.

[19] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom B. Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. Predictability and surprise in large generative models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1747–1764. ACM, 2022. doi: 10.1145/3531146.3533229. URL https://doi.org/10.1145/3531146.3533229.

[20] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[22] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701, 2020. URL https://arxiv.org/abs/2010.14701.

[23] 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. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

[24] Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, et al. Predicting emergent abilities with infinite resolution evaluation. *arXiv e-prints*, pages arXiv–2310, 2023.

[25] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322, 2023. doi: 10.48550/ARXIV.2305.08322. URL https://doi.org/10.48550/arXiv.2305.08322.

[26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

[27] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

[28] 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. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[29] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[30] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-2012. URL https://doi.org/10.18653/v1/d18-2012.

[31] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics, 2017. URL https://doi.org/10.18653/v1/d17-1082.

[32] Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *CoRR*, abs/2305.14342, 2023. doi: 10.48550/ARXIV.2305.14342. URL https://doi.org/10.48550/arXiv.2305.14342.

[33] Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23ao.html.

[34] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[36] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

[37] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-4009. URL https://doi.org/10.18653/v1/n19-4009.

[38] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[39] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional

language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28043–28078. PMLR, 2023. URL https://proceedings.mlr.press/v202/poli23a.html.

[40] Alethea Power, Yuri Burda, Harrison Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *CoRR*, abs/2201.02177, 2022. URL https://arxiv.org/abs/2201.02177.

[41] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

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

[43] Jihyeon Roh, Sang-Hoon Oh, and Soo-Young Lee. Unigram-normalized perplexity as a language model performance measure with different vocabulary sizes. *CoRR*, abs/2011.13220, 2020. URL https://arxiv.org/abs/2011.13220.

[44] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6399. URL https://doi.org/10.1609/aaai.v34i05.6399.

[45] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

[46] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *CoRR*, abs/2304.15004, 2023. doi: 10.48550/ARXIV.2304.15004. URL https://doi.org/10.48550/arXiv.2304.15004.

[47] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for*

Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1162. URL https://doi.org/10.18653/v1/p16-1162.

[48] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR, 2018. URL http://proceedings.mlr.press/v80/shazeer18a.html.

[49] Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woo-Myoung Park, Jung-Woo Ha, and Nako Sung. On the effect of pretraining corpora on in-context learning by a large-scale language model. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5168–5186. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.380. URL https://doi.org/10.18653/v1/2022.naacl-main.380.

[50] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556.

[51] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615, 2022. doi: 10.48550/ARXIV.2206.04615. URL https://doi.org/10.48550/arXiv.2206.04615.

[52] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:141–155, 2020. doi: 10.1162/TACL_A_00305. URL https://doi.org/10.1162/tacl_a_00305.

[53] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pretraining and finetuning transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[54] Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12342–12364. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.825.

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

[56] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony

Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

[57] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[58] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[60] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 94–106. Association for Computational Linguistics, 2017.

[61] Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. Training trajectories of language models across scales. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13711–13738. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.767. URL https://doi.org/10.18653/v1/2023.acl-long.767.

[62] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A chinese language understanding evaluation benchmark. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4762–4772. International Committee on Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.coling-main.419.

[63] Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, Zheni Zeng, Yuxian Gu, Kun Zhou, Xuancheng Huang, Wenhao Li, Shuhuai Ren, Jinliang Lu, Chengqiang Xu, Huadong Wang, Guoyang Zeng, Zile Zhou, Jiajun Zhang, Juanzi Li, Minlie Huang, Rui Yan, Xiaodong He, Xiaojun Wan, Xin Zhao, Xu Sun, Yang Liu, Zhiyuan Liu, Xianpei Han, Erhong Yang, Zhifang Sui, and Maosong Sun. CUGE: A chinese language understanding and generation evaluation benchmark. *CoRR*, abs/2112.13610, 2021. URL https://arxiv.org/abs/2112.13610.

 https://doi.org/10.52202/079017-1683

[64] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/p19-1472.

[65] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=-Aw0rrrPUF.

[66] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. doi: 10.48550/ARXIV.2205.01068. URL https://doi.org/10.48550/arXiv.2205.01068.

A Pre-training Settings

A.1 Pre-training Corpus

Source	Ratio
CommonCrawl	80.2%
Code	10.0%
Books	3.8%
Wikipedia	3.8%
Papers	1.6%
StackExchange	0.6%

Table 3: The ratio of different sources in the English corpus.

Our pre-training corpus is a mixture of English and Chinese documents. The ratio of English tokens to Chinese tokens in the pre-training corpus is 4:1. Both the English and Chinese corpora consist of webpages, wikipedia, books, and papers. The distribution of different sources in the English corpus is shown in Table 3. The distribution and processing pipeline are similar to Redpajama [13]. During pre-training, documents from Wikipedia and Books are trained for multiple epochs, but most documents (93.4% in the pre-training corpus) are never repeated.

We tokenize the data with the byte pair encoding (BPE) algorithm [47] in the SentencePiece package [30]. The vocabulary size is 65k.

A.2 Hyperparameters

The hyperparameters for training of 1.5B, 6B, and 32B models are shown in Table 4. The hyperparameters for training of smaller models are shown in Table 5. The sequence length is 2048 and the optimizer is AdamW [35] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$.

B Evaluation Settings

The evaluated splits and numbers of examples are summarized in Table 6. For English datasets, we follow Gopher [41] and Chinchilla [23]'s selection of evaluation splits. For Chinese datasets, we use the validation split when the ground labels are always available. For CLUEWSC, the size of the validation set is too small (100), so we combine the train and validation splits. GSM8K-Chinese is translated from GSM8K with machine translation and human proofreading.

C Are Emergent Abilities of Language Models a Mirage?

[46] claim that emergent abilities proposed in [58] are mainly a mirage caused by nonlinear and discontinuos metrics. [61] also support the idea.

[61] use the perplexity of correct options as the metric for BIG-Bench and find that the metric impproves smoothly on almost all the tasks of BIG-Bench. We argue that the perplexity of correct options is not the correct metric to evaluate the performance of multi-choice questions. The correct metric of multi-choice questions should reflect the ability of distinguishing correct options from incorrect options. The perplexity of correct options and incorrect options may decrease simultaneously. In fact, [61] already observe perplexity of incorrect options decreasing during pre-training and only at

Parameters	Tokens	d_model	d_hidden	n_heads	n_layers	Batch Size	Max LR
1.5B	3T	2048	6912	16	24	1344	5e-4
6B	3T	4096	13696	32	28	4224	4e-4
32B	2.5T	6656	22272	52	58	8832	3e-4

Table 4: Hyperparameters of pre-training of 1.5B, 6B, and 32B models.

https://doi.org/10.52202/079017-1683

Parameters	Tokens	d_model	d_hidden	n_heads	n_layers	Batch Size	Max LR
300M	67B	1152	3840	9	12	1152	2.8e-3
300M	125B	1152	3840	9	12	1152	2.8e-3
300M	250B	1152	3840	9	12	1152	2.8e-3
300M	500B	1152	3840	9	12	1152	2.8e-3
540M	33B	1536	5120	12	12	1152	2e-3
540M	66B	1536	5120	12	12	1152	2e-3
540M	125B	1536	5120	12	12	1152	2e-3
540M	250B	1536	5120	12	12	1152	2e-3
540M	500B	1536	5120	12	12	1152	2e-3
1B	33B	2048	6912	16	16	1152	1.5e-3
1B	67B	2048	6912	16	16	1152	1.5e-3
1B	125B	2048	6912	16	16	1152	1.5e-3
1B	250B	2048	6912	16	16	1152	1.5e-3
1B	500B	2048	6912	16	16	1152	1.5e-3
1.5B	67B	2048	6912	16	24	1152	1e-3
1.5B	100B	2048	6912	16	24	1152	1e-3
1.5B	125B	2048	6912	16	24	1152	1e-3
1.5B	250B	2048	6912	16	24	1152	1e-3
1.5B	375B	2048	6912	16	24	1152	1e-3
1.5B	500B	2048	6912	16	24	1152	1e-3
3B	67B	3072	10240	24	24	1152	7e-4
3B	125B	3072	10240	24	24	1152	7e-4
3B	250B	3072	10240	24	24	1152	7e-4
3B	500B	3072	10240	24	24	1152	7e-4
6B	33B	4096	13696	32	28	1152	4e-4
6B	67B	4096	13696	32	28	1152	4e-4
6B	125B	4096	13696	32	28	1152	4e-4
6B	250B	4096	13696	32	28	1152	4e-4

Table 5: Hyperparameters of pre-training of smaller models. Each line represents one model pre-trained completely from scratch with the certain number of tokens and its corresponding learning rate schedule.

Dataset	Evaluated Split	Num. Examples
TriviaQA	validation	11,313
HellaSwag	validation	10,042
RACE	test	4,934
WinoGrande	validation	1,267
MMLU	test	14,042
GSM8K	test	1,319
NLPCC-KBQA	validation	10,613
ClozeT	validation	938
CLUEWSC	train & validation	508
C3	validation	3,816
C-Eval	validation	1,346
GSM8K-Chinese	test	1,212

Table 6: Statistics of evaluation datasets.

the end of training that the perplexity of correct and incorrect options starts to diverge. This supports the existence of emergent abilities.

[46] use Brier Score [5] as the metric for BIG-Bench. We argue that increase in Brier Score does not always represent improvement of performance on the multi-choice task, since Brier Score is also related to the allocation of probabilities for incorrect options. For example, questions in the MMLU dataset have four options (A, B, C, and D) and the frequency of the four options as correct is equal.

Consider two models that give the same probability independent of questions. One model predicts $(1, 0, 0, 0)$ for the four options and the other model predicts $(0.25, 0.25, 0.25, 0.25)$. The Brier Score for the former is 1.5 while the Brier Score for the latter is 0.75. However, both models do not learn the relationship between questions and correct options at all. One can argue that the latter model better fits the distribution of correct options in the dataset, but the improvement is not as large as the different of 1.5 and 0.75. We should consider the Brier Score of 0.75 as the performance of the random guess baseline, and any decrease in Brier Score above 0.75 should not be considered as the real improvement on the task.

In Figure 6 of [46], they evaluate 4 tasks in BIG-Bench with the Brier Score metric and find that the emergent abilities disappear. We hypothesis that they normalize the Brier Score with the number of options in each question, otherwise the Brier Score of 0.25 on the swahili_english_proverbs task is too low for the smallest model. Four tasks have 2, 2, 4, 5 options in each question. The values of Brier Score for random guess baselines on the four tasks are 0.25, 0.25, 0.1875, and 0.16. Only the largest model surpasses the random guess baseline. This also supports the existence of emergent abilities.

D Complete Performance-vs-Loss Curves of Smaller Models

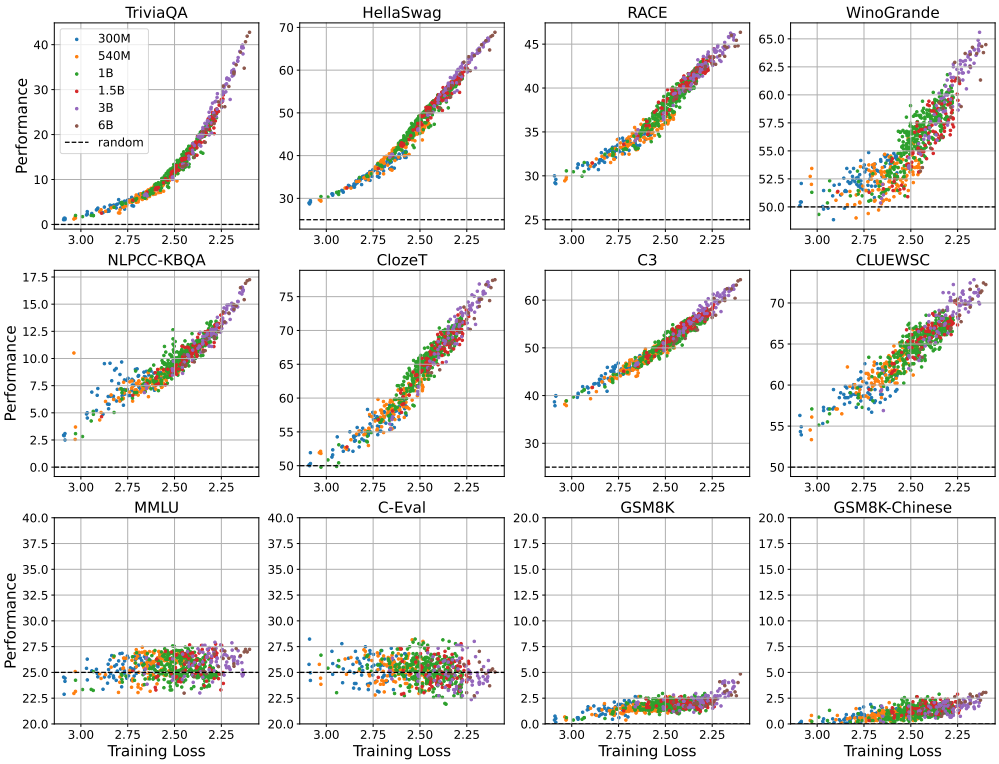

Figure 5: The complete performance-vs-loss curves of smaller models.

The performance-vs-loss curves for all the intermediate checkpoints are shown in Figure 5. The trend is the same as Figure 2, but with larger variance.

E Loss vs Compute as an Indicator of Performance

We show the performance-compute curves in Figure 6. Compared with Figure 1, we observe that points from different models do not fall on the same curves on most tasks. This proves that pre-training loss is a better indicator of task performance than compute.

https://doi.org/10.52202/079017-1683

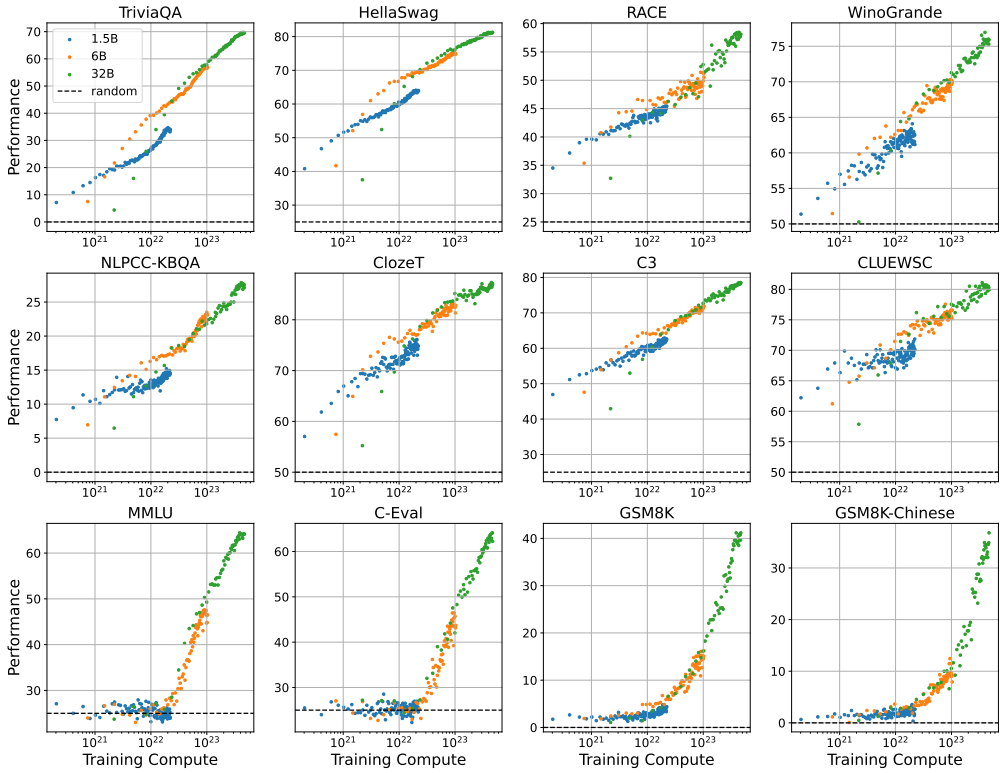

Figure 6: The performance-vs-compute curves of 1.5B, 6B, and 32B models.

F Pythia's Loss vs. Performance

To further support our conclusion, we plot the performance-loss curves of Pythia [3] in Figure 7. Pythia is a suite of open language models with intermediate checkpoints released. For downstream tasks, we select SciQ [60], LAMBADA [38], WinoGrande [44], ARC-Easy [11], ARC-Challenge [11], and PIQA [4] with reported performance in the official repository. We compute the cross-entropy loss of intermediate checkpoints on the corpus Pile [20]. From the plot, we can observe that the points from different models fall on the same curve on all the tasks. This supports our conclusion that pre-training loss is predictive of task performance.

However, neither Pythia nor LLaMA can be used to analyze the emergent abilities. The largest Pythia model fails to achieve performance above random chance on MMLU and GSM8K [2]. Instead, LLaMA has no intermediate checkpoints released and performance curves on MMLU and GSM8K are not available.

G Loss vs. Performance on BIG-bench

BIG-bench [51] is a series of diverse tasks designed to evaluate the capacities and limitations of pre-trained language models. Wei et al. [58] find that large language models exhibit emergent abilities on four tasks from BIG-Bench. Among the four tasks, the test set size of the figure-of-speech detection task is too small and the variance is too high. We evaluate the other three tasks in the same setting as Section 2.3 and the results are shown in Figure 8. With pretraining loss decreases along the x-axis, we can clearly observe the tipping point in the performance curves.

https://doi.org/10.52202/079017-1683

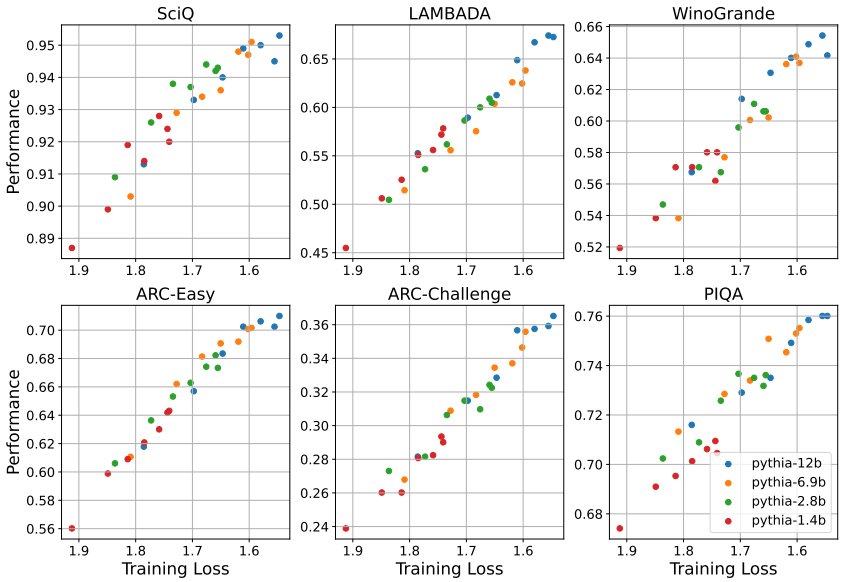

Figure 7: **The performance-vs-loss curves of Pythia.** The performance of Pythia is from the official repository and the loss is evaluated with the released checkpoints.

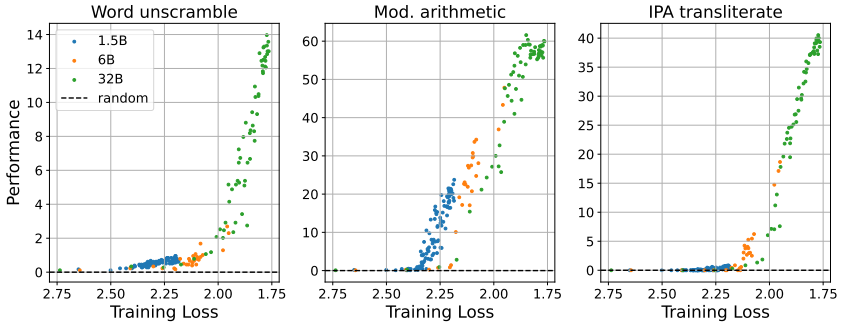

Figure 8: **The performance-vs-loss curves of 1.5B, 6B, and 32B models on 3 tasks in BIG-bench.** Each data point is the loss (x-axis) and performance (y-axis) of the intermediate checkpoint of one of the three models. We mark the results of random guess in black dashed lines.

H Compute Resources

All the models are trained on DGX-A100 GPU (8x80G) servers. The 1.5B, 6B, and 32B models in Section 2.3 take 8 days on 256 A100 GPUs, 8 days on 1024 A100 GPUs, and 20 days on 2048 A100 GPUs respectively. The small models in Section 2.4 take about 20 days on 256 A100 GPUs.

I Broader Impact

This paper finds that pre-training loss is predictive of downstream task performance and on some tasks the performance only begins to improve when the pre-training loss falls below a certain threshold. Combined with previous works on scaling laws [28, 22, 23], we can predict the amount of compute required to achieve a certain performance. This can be used to estimate the cost of training a large model.

The paper might encourage companies to expand model sizes and data sizes of language models beyond current scales to pursue new emergent abilities, leading to a waste of compute resources. We want to emphasize that the analysis of previous performance trends do not necessarily apply to the larger models.

https://doi.org/10.52202/079017-1683

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction include the main claims made in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The model are pre-trained with the open-source framework, Megatron-LM [1], and we provide the training hyperparameters. The datasets can be replaced with open-source counterparts, such as RedPajama.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

[1] https://github.com/NVIDIA/Megatron-LM

https://doi.org/10.52202/079017-1683

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The training code is Megatron-LM [2]. The pre-training dataset cannot be released since it contains proprietary content.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of pre-training are described in Appendix A and those of evaluation are described in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: It is too computationally expensive to train multiple pre-trained language models from scratch.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

[2] `https://github.com/NVIDIA/Megatron-LM`

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper does not include human subjects or participants.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Appendix I.

    Guidelines:

https://doi.org/10.52202/079017-1683

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks since it does not include release of data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the creators of code, datasets, and models used in the paper are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-1683

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.