Understanding Emergent Abilities of Language Models from the Loss Perspective

Anonymous ACL submission

Abstract

Recent studies have put into question the belief that emergent abilities (Wei et al., 2022b) 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 models with the same pre-training loss, but different model and data sizes, generate the same performance on various downstream tasks. 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 (Raffel et al., 2020; Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023; Zeng et al., 2023; Touvron et al., 2023a; OpenAI, 2023), leading to the widespread adoption of LM applications, e.g., ChatGPT. The success of such scaling is guided by scaling laws (Henighan et al., 2020; Kaplan et al., 2020; Clark et al., 2022; Hoffmann et al., 2022), 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 (Wei et al., 2022b) 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 number of tokens can outperform large models on tasks with claimed emergent abilities (Touvron et al., 2023a,b; Jiang et al., 2023). For example, LLaMA-13B with less compute (Touvron et al., 2023a) can outperform GPT-3 (175B) on MMLU (Hendrycks et al., 2021). Second, Schaeffer et al. (2023) 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.

Hoffmann et al. (2022) 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 (Liu et al., 2023b; Tay et al., 2023) or constrained its study to single models, tasks, or prompting methods (Shin et al., 2022; Xia et al., 2023).

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 fur-

ther verified by extracting and observing the performance and loss relationship of the open LLaMA models (Touvron et al., 2023a).

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 (Schaeffer et al., 2023; Xia et al., 2023), 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 (Henighan et al., 2020; Kaplan et al., 2020), 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 (Wei et al., 2022b), 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 The Pre-training Loss Predicts Task Performance?

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 (Touvron et al., 2023a) 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. For simplicity, we consider the Exact Match (EM) metric with single-token target. The score $\mathrm{EM}(\hat{y}, y)$ for the prediction $\hat{y}$ of the prompt x given the ground truth y is 1 if $\hat{y} = y$ and 0 otherwise. The expectation of $\mathrm{EM}(\hat{y}, y)$ is

$$\mathbb{E}[\mathrm{EM}(\hat{y}, y)] = P_{\mathrm{LM}}(y|x) = \exp(-\ell(y|x)) \quad (1)$$

where $\ell(y|x)$ is the cross entropy loss of the LM given the context x and the target y.

Note that 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 (Dar et al., 2021; Cao and Gu, 2020).

2.1 Pre-training Setting

All the models are pre-trained on a mixture of English and Chinese corpus. Both the English and Chinese corpora consist of webpages, wikipedia, books, and papers. The ratio of English to Chinese is 4:1 in the pre-training corpus. We tokenize the data with the byte pair encoding (BPE) algorith (Sennrich et al., 2016) with the SentencePiece package (Kudo and Richardson, 2018).

The model architecture is similar to LLaMA (Touvron et al., 2023a) with two differences: we use grouped-query attention (Ainslie et al., 2023) to replace the multi-query attention and we apply rotary position embedding on half the dimensions of the query and key vectors.

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 (Lai et al., 2017) for English. For Chinese, we build a closed-book QA

Dataset	Task	Prompting Type	Answer Form	Metric
		English datasets		
TriviaQA (Joshi et al., 2017)	Closed-book QA	Few-shot	Open-formed	EM
HellaSwag (Zellers et al., 2019)	Commonsense NLI	Zero-shot	Mulit-choice	Accuracy
RACE (Lai et al., 2017)	Reading Comprehension	Few-shot	Multi-choice	Accuracy
WinoGrande (Sakaguchi et al., 2020)	Coreference Resolution	Zero-shot	Multi-choice	Accuracy
MMLU (Hendrycks et al., 2021)	Examination	Few-shot	Multi-choice	Accuracy
GSM8K (Cobbe et al., 2021)	Math Word Problem	Few-shot CoT	Open-formed	EM
		Chinese datasets		
NLPCC-KBQA(Duan, 2016)	Closed-book QA	Few-shot	Open-formed	EM
ClozeT (Yao et al., 2021)	Commonsense NLI	Zero-shot	Multi-choice	Accuracy
CLUEWSC (Xu et al., 2020)	Coreference Resolution	Zero-shot	Multi-choice	Accuracy
C3 (Sun et al., 2020)	Reading Comprehension	Few-shot	Multi-choice	Accuracy
C-Eval (Huang et al., 2023)	Examination	Few-shot	Multi-choice	Accuracy
GSM8K-Chinese	Math Word Problem	Few-shot CoT	Open-formed	EM

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 (Wei et al., 2022c) as few-shot CoT and the original in-context learning prompting (Brown et al., 2020) as few-shot.

dataset based on NLPCC-KBQA (Duan, 2016) dataset following the TriviaQA format.

Commonsense Natural Language Inference (NLI): Selecting the most likely followup given an event description. We use the HellaSwag dataset (Zellers et al., 2019) for English and the ClozeT dataset in Yao et al. (2021) for Chinese.

Reading comprehension: Reading a given article or paragraph and answering questions about it. We use RACE (Lai et al., 2017) for English and C3 (Sun et al., 2020) 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 (Sakaguchi et al., 2020) for English and the CLUEWSC dataset (Xu et al., 2020) for Chinese.

Examination: Multiple-choice questions in examinations. For English, we use MMLU (Hendrycks et al., 2021), which includes mathematics, US history, computer science, law, and more. For Chinese, we use C-Eval (Huang et al., 2023) which comprises multiple-choice ranging 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 (Cobbe et al., 2021) dataset. For Chinese, we translate the questions and answers in GSM8K to Chinese, namely GSM8K-Chinese.

The prompting types cover few-shot (Brown et al., 2020), zero-shot, and few-shot chain-of-thought (CoT) (Wei et al., 2022c). 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 3 (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 loss (x-axis) in Figure 1. From the curves, 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 increases. 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

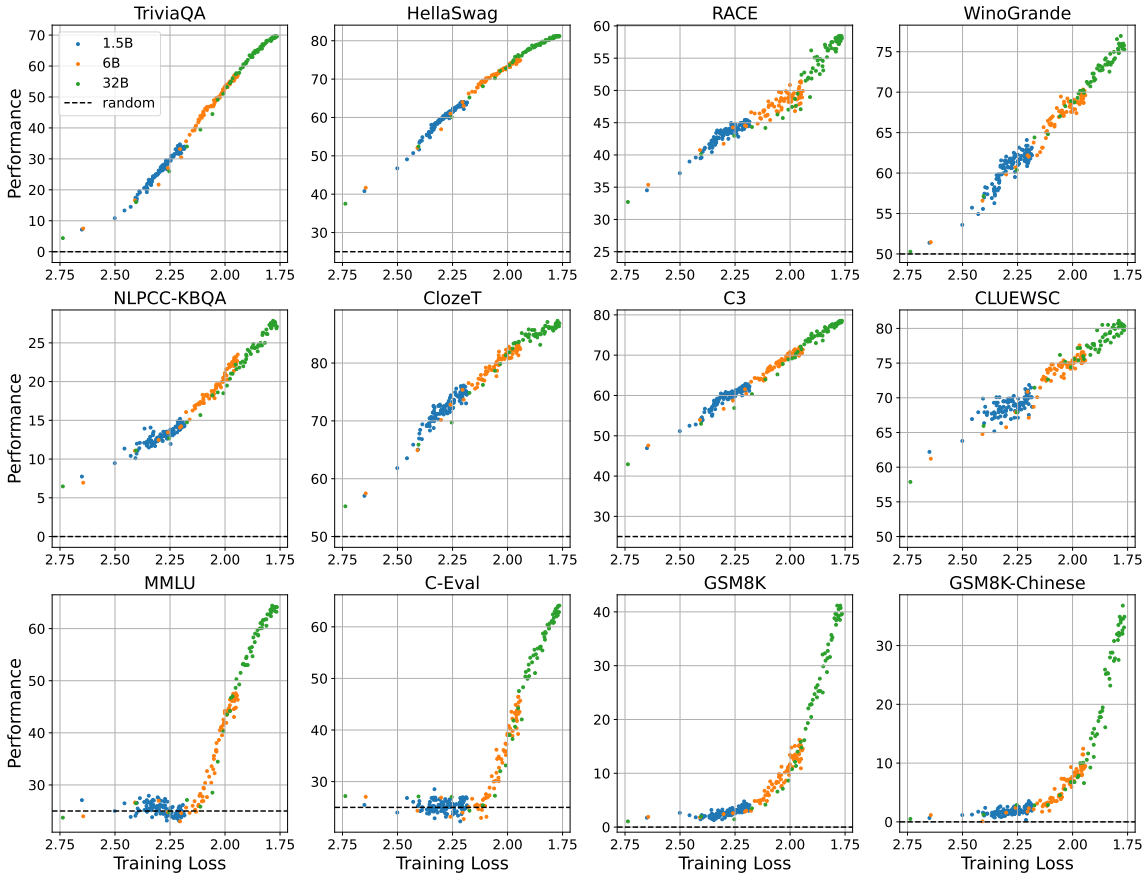

Figure 1: **The performance-vs-loss curves of 1.5B, 6B, and 32B models.** Each data point is the loss (x-axis) and performance (y-axis) of the intermediate checkpoint of one of the three models. We mark the results of random guess in black dashed lines.

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.*

- 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

Following the empirical experiments in scaling laws (Henighan et al., 2020; Kaplan et al., 2020; Hoffmann et al., 2022), 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. The learning rate schedule is set to reach the minimum at the corresponding token count, which is critical to the optimal performance (Kaplan et al., 2020; Hoffmann et al., 2022). The number of tokens used and hyperparameters for all models are shown in Table 4 (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 perfor-

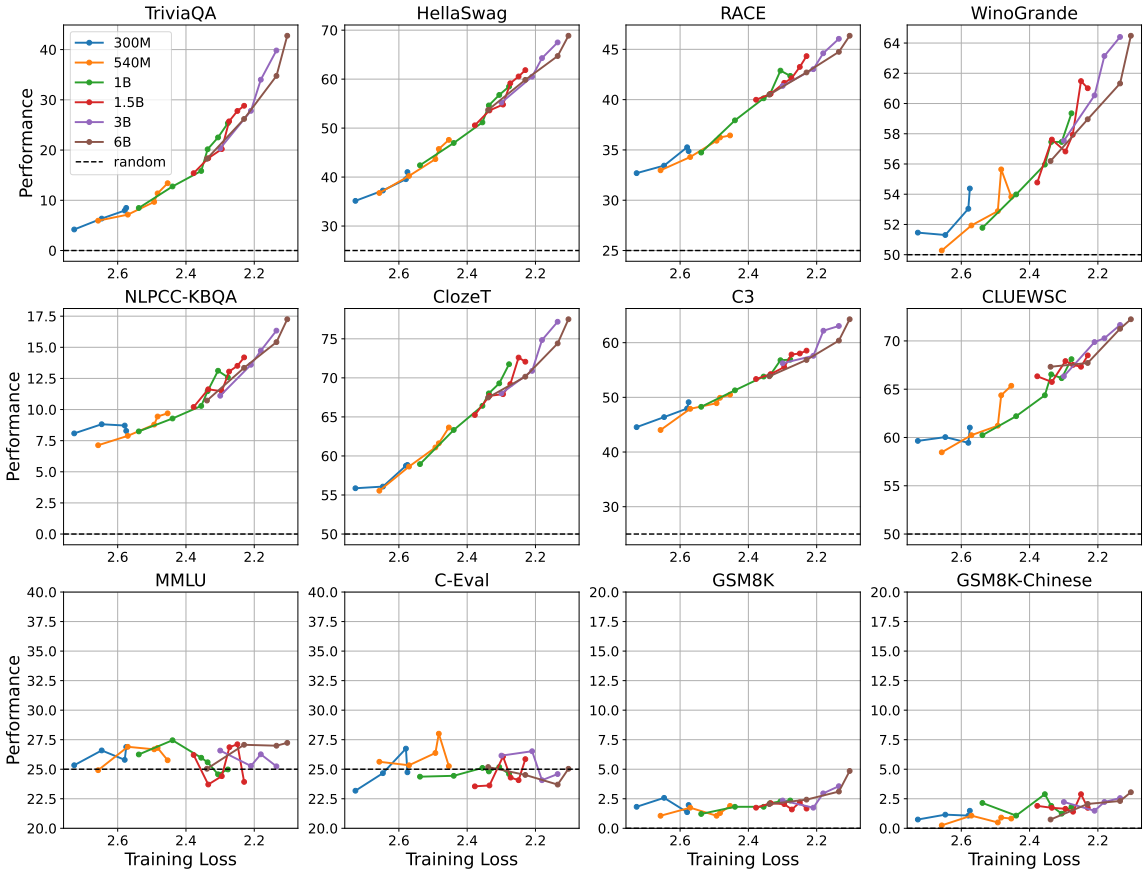

Figure 2: **The performance-vs-loss curves of smaller models pre-trained with different numbers of training tokens**. Each data point is the loss (x-axis) and performance (y-axis) of the final checkpoint of one model, i.e., each point corresponds to one model trained from scratch. We mark the results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

mance 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).

2.5 LLaMA's Loss vs. Performance

To validate the generality of our observations, we analyze a different model series with required information made publicly available, i.e., LLaMA (Touvron et al., 2023a). Compared to our models, LLaMA uses a pre-training corpus that excludes Chinese documents, leverages a different pre-training framework (Ott et al., 2019), 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 one only 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. Note that the outliers only constitute the initial 10% training tokens.

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,

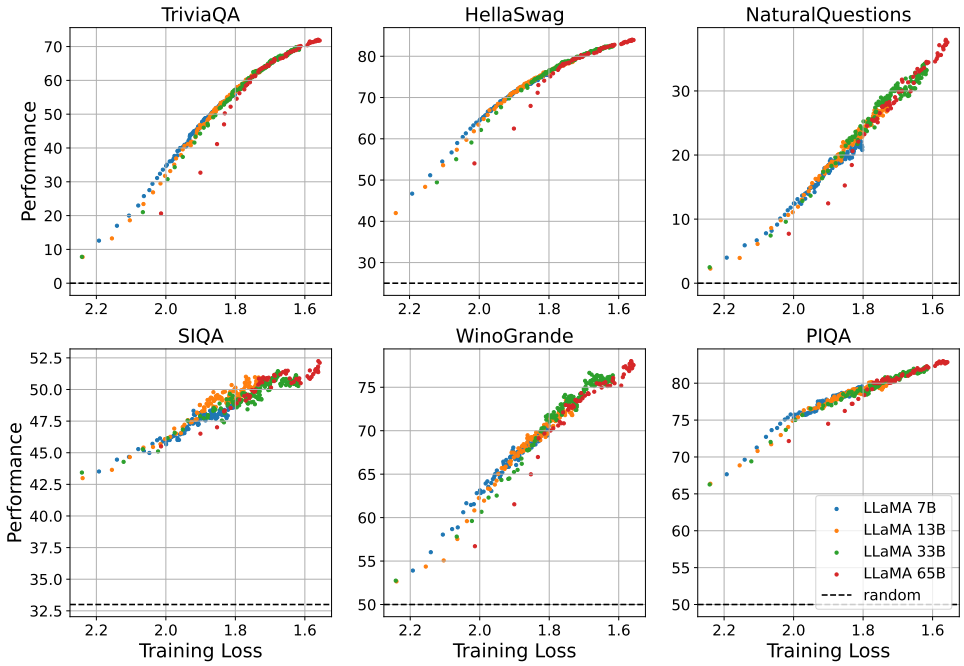

Figure 3: **The performance-vs-loss curves of LLaMA.** The values of performance and training loss are extracted from the figures in the original LLaMA paper (Touvron et al., 2023a). Note that the LLaMA2 paper (Touvron et al., 2023b) does not cover such figures with related information.

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.

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 are used to be considered as impossible to solve by pre-trained language models without 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 (Power et al., 2022). Power et al. (2022) 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 (Wei et al., 2022b), that is, abilities that only present in large models. According to the scaling law, with the number of training tokens fixed, the pre-training loss follows a power law with re-

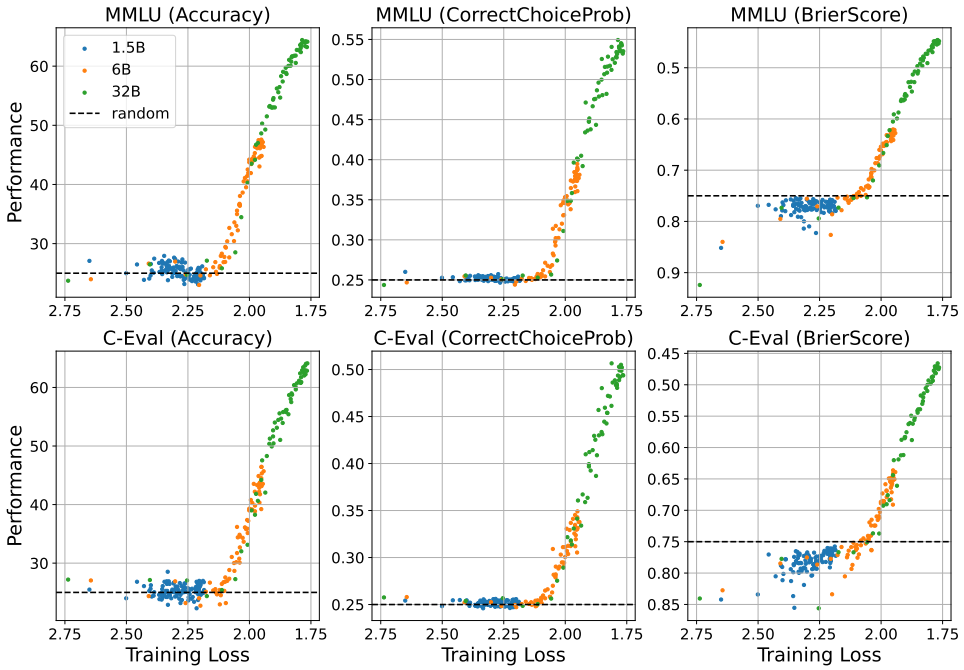

Figure 4: **The performance-vs-loss curves of different metrics on MMLU and C-Eval.** Accuracy: discontinuous; CorrectChoiceProb and BrierScore: continuous. We mark the result of random guess in black dashed lines.

spect 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.

3.2 Influence of Different Metrics

Schaeffer et al. (2023) 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 accuracy (discontinuous) used in the original benchmarks. The first metric is the predicted probability of the correct answer (CorrectChoiceProb). The second one is the Brier Score (Brier, 1950) used in Schaeffer et al. (2023):

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

where $\hat{y}_{ij}$ is the predicted probability of sample i for class j and y_{ij} is the ground probability.

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. (2023) and Xia et al. (2023) 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.

In Henighan et al. (2020), 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. (2022b): 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, Liu et al. (2023b); Tay et al. (2023) find that models with the same pre-training loss can have different downstream performance after finetuning, due to inductive bias in model sizes, model architectures, and training algorithms. For the prompted performance of large language models, Xia et al. (2023) claim that perplexity is a strong predictor of in-context learning performance, but the evidence is limited to the OPT model (Zhang et al., 2022) and a subset of BIG-Bench (Srivastava et al., 2022). Instead, Shin et al. (2022) find that low perplexity does not always imply high in-context learning performance when the pre-training corpus changes.

Emergent abilities. Wei et al. (2022b) propose the idea of emergent abilities, abilities that only present in large language models. This is similar to the claim of Ganguli et al. (2022) 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. (2022) show that smaller language models trained with sufficient data can outperform undertrained larger language models, supported by follow-up models (Touvron et al., 2023a; Jiang et al., 2023; Touvron et al., 2023b). On the other hand, Schaeffer et al. (2023) claim that emergent abilities are due to the discontinuos metrics used for evaluation, also found in Xia et al. (2023). Similarly, Hu et al. (2023) 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 perspecitve 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 performance on downstream tasks of language models, 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, a language model with a slightly different architecture, and fine that the relationship holds for the model family. But there are fundamentally different model architectures, such as routed Transformers (Fedus et al., 2022), and non-Transformer architectures (Fu et al., 2023; Poli et al., 2023) beyond our consideration. Both our models and LLaMA use AdamW optimizer (Loshchilov and Hutter, 2019), while there are other optimizers for language model pre-training (Shazeer and Stern, 2018; Liu et al., 2023a).

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 (Roh et al., 2020) 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, pre-training is not the only way to improve the performance of emergent abilities. For example, instruction tuning (Wei et al., 2022a; Sanh et al., 2022; Chung et al., 2022; Longpre et al., 2023) can improve the zero-shot performance of language models on unseen tasks, including the MMLU dataset. Future studies can analyze the acquisition of emergent abilities and lower the scale requirements.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: training generalized multi-query transformer models from multi-head checkpoints. In *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.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yuan Cao and Quanquan Gu. 2020. 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.

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. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

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.

2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

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. 2022. Unified scaling laws for routed language models. In *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.

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. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

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

Nan Duan. 2016. 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.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. 2023. 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.

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. 2022. 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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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. 2020. Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701.

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. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

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

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. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

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. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *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.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *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.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. 2023a. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *CoRR*, abs/2305.14342.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023b. Same pre-training loss, better downstream: Implicit bias matters for language models. In *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.

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. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *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.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *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.

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

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. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Jihyeon Roh, Sang-Hoon Oh, and Soo-Young Lee. 2020. Unigram-normalized perplexity as a language model performance measure with different vocabulary sizes. *CoRR*, abs/2011.13220.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. 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.

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, An-

drea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. 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.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *CoRR*, abs/2304.15004.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *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.

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. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *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.

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. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:141–155.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. 2023. Scaling laws vs model architectures: How does inductive bias influence scaling? In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12342–12364. Association for Computational Linguistics.

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. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. 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.

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. 2022b. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *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*.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *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.

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. 2020. CLUE: A chinese language understanding evaluation benchmark. In *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.

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. 2021. CUGE: A chinese language understanding and generation evaluation benchmark. *CoRR*, abs/2112.13610.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *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.

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. 2023. 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.

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. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

A Pre-training Hyperparameters

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

Table 2: The ratio of different sources in the English corpus.

The hyperparameters for training of 1.5B, 6B, and 32B models are shown in Table 3. The hyperparameters for training of smaller models are shown in Table 4. The sequence length is 2048 and the optimizer is AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$.

B Evaluation Dataset Statistics

The evaluated splits and numbers of examples are summarized in Table 5. For English datasets, we follow Gopher (Rae et al., 2021) and Chinchilla (Hoffmann et al., 2022)'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?

Schaeffer et al. (2023) claim that emergent abilities proposed in Wei et al. (2022b) are mainly a mirage caused by nonlinear and discontinuos metrics. Xia et al. (2023) also support the idea.

Xia et al. (2023) 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, Xia et al. (2023) already observe perplexity

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

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 4: 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 5: Statistics of evaluation datasets.

of incorrect options decreasing during pre-training and only at the end of training that the perplexity of correct and incorrect options starts to diverge. This supports the existence of emergent abilities.

Schaeffer et al. (2023) use Brier Score (Brier, 1950) 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 Schaeffer et al. (2023), they evaluate 4 tasks in BIG-Bench with the Brier Score metric and find that the emergent abilities disapper. 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 basenlines 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

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.

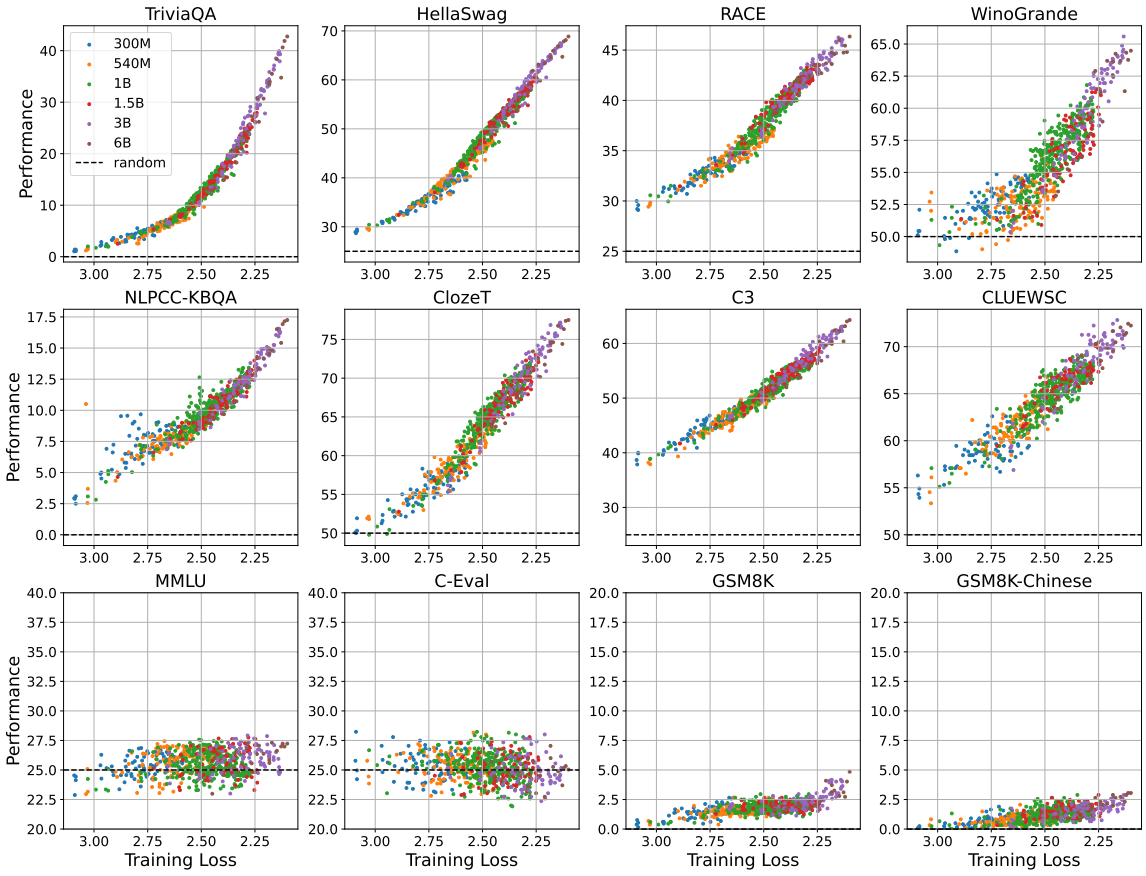

Figure 5: The complete performance-vs-loss curves of smaller models.

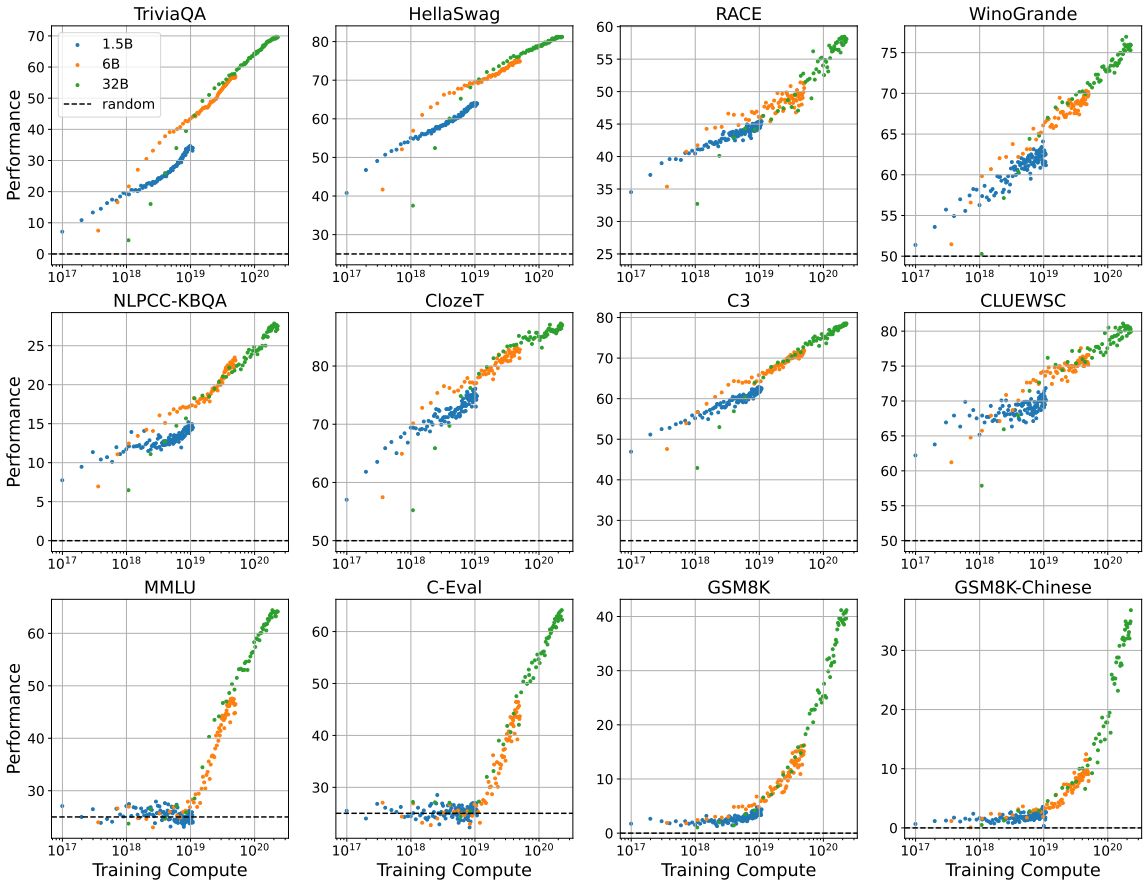

Figure 6: The performance-vs-compute curves of 1.5B, 6B, and 32B models.