U-SHAPED AND INVERTED-U SCALING BEHIND EMERGENT ABILITIES OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have been shown to exhibit *emergent abilities* in some downstream tasks, where performance seems to stagnate at first and then improve sharply and unpredictably with scale beyond a threshold. By dividing questions in the datasets according to difficulty level by average performance, we observe U-shaped scaling for hard questions, and inverted-U scaling followed by steady improvement for easy questions. Moreover, the emergence threshold roughly coincides with the point at which performance on easy questions reverts from inverse scaling to standard scaling. Capitalizing on the observable though opposing scaling trend on easy and hard questions, we propose a simple yet effective pipeline, called *Slice-and-Sandwich*, to predict both the emergence threshold and model performance beyond the threshold.

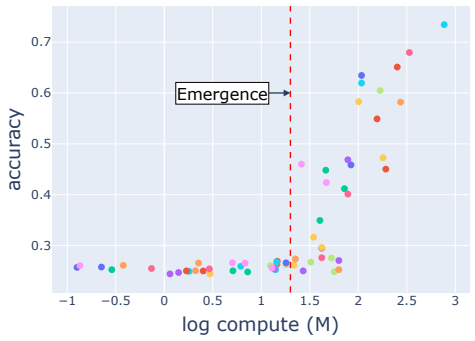
1 INTRODUCTION

Large language models (LLMs) (Team et al., 2023; Achiam et al., 2023; Brown, 2020; Touvron et al., 2023a;b; Workshop et al., 2022; Li et al., 2023; Jiang et al., 2024) have shown strong potential in various downstream applications (Jumper et al., 2021; Fawzi et al., 2022; Naveed et al., 2023; Kaddour et al., 2023). Though the training-loss scaling law has been well established (Kaplan et al., 2020; Hoffmann et al., 2022), the literature is inconclusive regarding how performance on downstream tasks scales. In particular, for certain downstream tasks (Srivastava et al., 2023; Lin et al., 2022a; Pilehvar & Camacho-Collados, 2019), LLMs seem to display *emergent abilities*: performance is stagnant even when model training compute scales up hundredfold, and then exhibits sharp improvement at a seemingly unpredictable critical threshold (Wei et al., 2022; Schaeffer et al., 2024a).

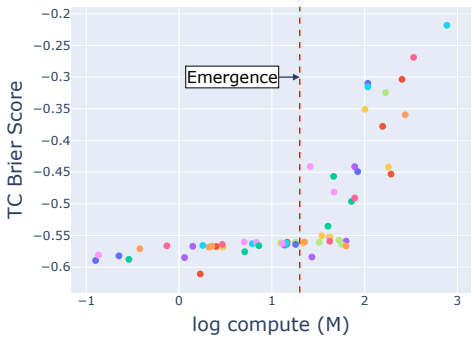
Some prior work (Schaeffer et al., 2024a;b; Lu et al., 2024) has identified crude performance metrics as a contributing factor to LLM's apparent emergent abilities because of its inability to capture improvements of smaller models. Hu et al. (2023) proposes the *PASSUNTIL* metric, according to which models slowly improve with scale instead of stagnating. Schaeffer et al. (2024a) finds that LLMs display emergent abilities mainly on string-match and multiple-choice tasks (Schaeffer et al., 2024a), for which the traditional performance measure of accuracy exhibits strong discontinuity. They propose using a continuous metric such as Brier Score (Brier, 1950) or linear metric such as token edit distance (TED) (Schaeffer et al., 2024a) to better predict LLM's scaling behavior for downstream tasks. Schaeffer et al. (2024b) further ranks several performance metrics in correlation with the model scale. On the other hand, Michaud et al. (2024) establish the quantization model of neural scaling to explain the emergent drop of cross-entropy loss from the aspect of next-token prediction.

Another focus of prior literature is the predictability of ability emergence on traditional metrics like accuracy, which is crucial for AI safety since a tool for task-wise emergence prediction can help us monitor and forecast LLMs' potential harmful capabilities, such as writing certain malicious code (Charan et al., 2023). Firstly, Wei et al. (2022) characterizes emergent abilities as *unpredictable* performance soar. Though some studies (Ruan et al., 2024; Gadre et al., 2024; Hu et al., 2023; Owen, 2024; Ye et al., 2023) have proposed pipelines to estimate task-specific scaling law, they usually incorporate models past the emergence threshold into the training set to fit a Sigmoid function and do not provide an explainable prediction of emergence abilities.

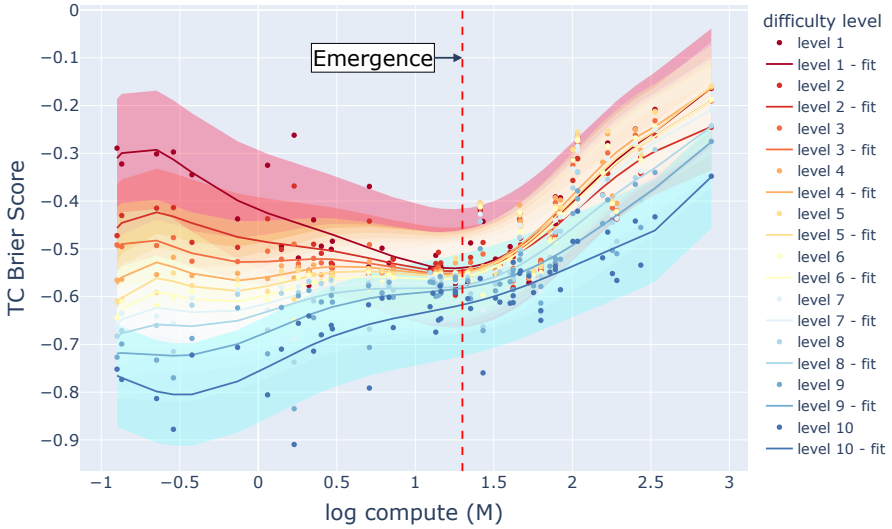
This paper contributes to both fronts of the literature's discussion on emergent abilities, especially for multiple-choice tasks. First, we propose a novel procedure to measure LLM's performance separately on groups of questions with different difficulty levels. Fig. 1 shows the evaluation result



(a) Accuracy vs. log compute (M).



(b) TC Brier Score vs. log compute (M).



(c) U-Shaped and inverted-U scaling with MMLU's questions clustered into 10 groups. Higher levels are harder questions.

Figure 1: The accuracy, Target-Conditioned (TC) Brier Score, U-shaped and inverted-U scaling on the MMLU benchmark (Hendrycks et al., 2021), with 56 LLMs used for evaluation and model details being in App. A. The TC Brier Score is our proposed performance measure to capture models' subtle performance change, which we detail in Sec. 2.1.2.

of 56 LLMs with diverse training compute on the MMLU benchmark, whose 14042 questions are clustered into 10 groups based on their difficulty levels, with higher levels denoting harder questions, measured by the *Target-Conditioned (TC) Brier Score*, our proposed continuous metric that has a high positive correlation with accuracy but can capture more nuanced model capability increase/decrease detailed in Sec. 2.1.2. With TC Brier Score, as shown in Fig. 1, we observe that model performance on hard questions exhibits U-shaped scaling (Wei et al., 2023; McKenzie et al., 2023), where it worsens with scale at first and then reverses to improve with scale. In contrast, performance on easy questions exhibits an inverted U-shape followed by steady improvement with scale, consistent with the previously reported deep double descent of testing loss (Nakkiran et al., 2021). Moreover, the point at which performance reverts from inverse to standard scaling roughly coincides with the emergence threshold beyond which model performance begins to soar. Our observation could explain why LLM's performance on some multiple-choice tasks stagnates for models below the emergence threshold: the scaling trend on easy questions offsets that on hard questions.

This observation of U-shaped and inverted-U scaling provides a basis to predict the forthcoming sharp increase in model performance, a defining feature of emergent abilities. We propose *Slice-and-Sandwich* pipeline, where we first group questions on a given downstream task by difficulty levels, use data before the emergence threshold to fit the performance on easy and hard questions separately,

then forecast performance on easy and hard questions separately beyond the emergence threshold. We show that *Slice-and-Sandwich* captures the performance soar well.

We summarize our contributions as follows:

- We demonstrate that, for some downstream tasks previously shown to display emergent abilities, under a proper continuous metric, LLM's performance exhibits opposing scaling trends: inverted-U vs. U-shape, on easy vs. hard questions below the emergence threshold, and steadily improves beyond the emergence threshold.
- Based on the observation of inverted-U vs. U-shape on easy vs. hard questions, we propose a simple yet effective pipeline, *Slice-and-Sandwich*, to forecast model performance past the emergence threshold. Experimental results on three iconic datasets show its effectiveness.

2 Scaling Trend by Difficulty Level: U-shape vs. Inverted-U

This section documents LLM's scaling trend by question difficulty level. Sec. 2.1 defines terminologies such as log compute, emergence threshold, and our performance metrics. Sec. 2.2 describes how we group questions by difficulty level. Sec. 2.3 presents and discusses the results of six iconic multiple-choice tasks with ability emergence.

2.1 Terminology

2.1.1 Log Compute and Emergence Threshold

For clearer visualization, in this paper, we refer to an LLM's log compute M as:

$$M = \log_{10}(\frac{C}{10^{21}}), \tag{1}$$

where $C \approx 6ND$ (Kaplan et al., 2020) is the total training compute (FLOPs) of an LLM, N is the number of model parameters, and D is the number of training tokens. We then recognize and mark the emergence threshold T manually as the log compute where the model accuracy exhibits a sharp improvement, as illustrated in Fig. 1a.

2.1.2 Continuous Performance Metrics

Prior work (Schaeffer et al., 2024a;b; Lu et al., 2024) has advocated for performance metrics that distinguish finer differences. One candidate metric is the Brier Score (Brier, 1950):

$$Brier = \frac{1}{K}\sum_{t=1}^{K}\sum_{i=1}^{C}\left(\hat{p}_{t,i} - p_{t,i}\right)^2, \tag{2}$$

where K is the number of samples and C is the number of classes. $p_{t,i}$ is 1 if the t-th sample belongs to class i, otherwise 0. $\hat{p}_{t,i}$ is the model's predicted probability of the t-th sample being class i.

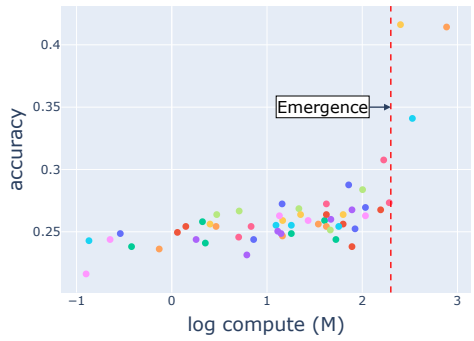
However, the Brier Score depends not only on the model's predicted probability of the target class (choice) but also on the predicted probability distribution of all classes. Since LLMs' confidence calibration is usually poor on multiple-choice questions (Li et al., 2024), we propose the *Target-Conditioned (TC) Brier Score* that depends only on the target label conditioned on the probability sum of all classes and has an opposite sign to Eq. 2 so that higher score means higher performance:

$$TC_Brier = -\frac{1}{K}\sum_{t=1}^{K}\left(\hat{p}^{con}_{t,c} - 1\right)^2, \tag{3}$$

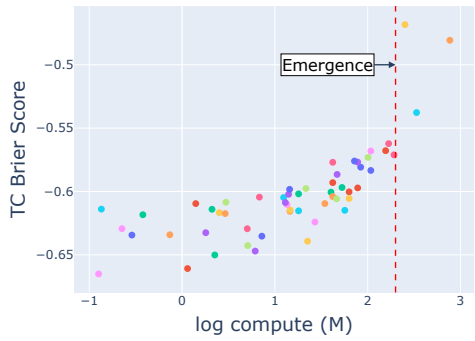
where $\hat{p}^{con}_{t,c}$ is the model's output probability on t-th sample's target class c conditional on available classes, i.e.,

$$\hat{p}^{con}_{t,c} = \frac{\hat{p}_{t,c}}{\sum\limits_{c' \in \text{ all classes}} \hat{p}_{t,c'}}, \tag{4}$$

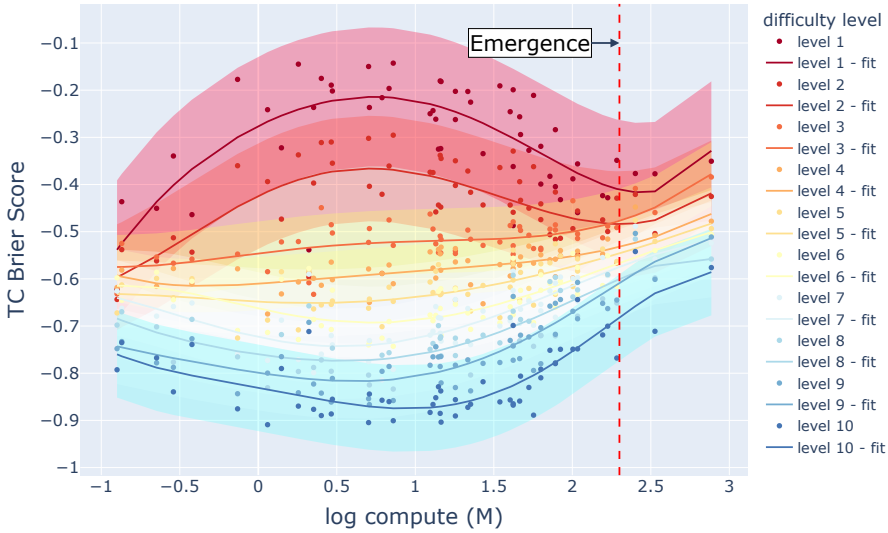
where $\hat{p}_{t,c}$ is the output probability on t-th sample's correct class c. The TC Brier Score is invariant under the output probability distribution of non-target classes. We discuss the effect of conditionality on the TC Brier Score in App. B.



(a) Accuracy vs. log compute (M).



(b) TC Brier Score vs. log compute (M).



(c) U-Shaped and inverted-U scaling.

Figure 2: The accuracy, TC Brier Score, U-Shaped and inverted-U scaling on the Persian-QA dataset in BIG-bench (Srivastava et al., 2023).

2.2 Grouping Questions by Difficulty Levels

2.2.1 Measuring Question Difficulty Level

For a sample question q of a downstream task, we define its difficulty level D_q to be the average performance on q based on the chosen performance metric (TC Brier Score in this paper) across all L LLM models smaller than the emergence threshold T for that downstream task:

$$D_q = \frac{1}{L}\sum_{i=1}^{L} TC_Brier_i^q, \tag{5}$$

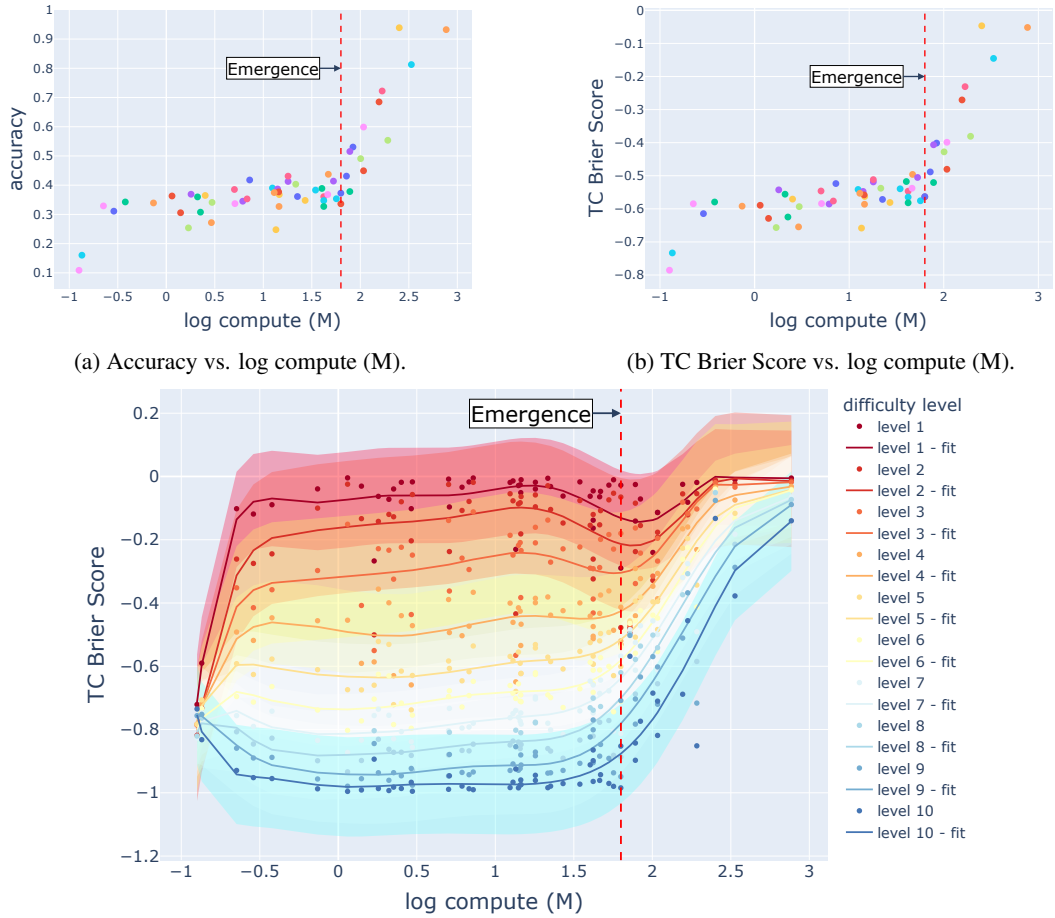
where $TC_Brier_i^q$ is the TC Brier Score of i-th LLM on sample question q, defined by Eq. 3.

2.2.2 Question Sorting and Grouping

Because model performance on individual questions is quite noisy, we group questions by difficulty levels. First, we sort questions by ascending difficulty level. Then, we evenly divide the sorted questions into G groups. Thus, each group has a different difficulty level.

2.3 U-Shaped and Inverted-U Scaling

Fig. 1a–3a show the scaling trend of accuracy on the MMLU, Persian-QA, and arithmetic datasets, with clear ability emergence demonstrated. In contrast, Fig. 1b–3b show the performance scaling



(a) Accuracy vs. log compute (M).

(b) TC Brier Score vs. log compute (M).

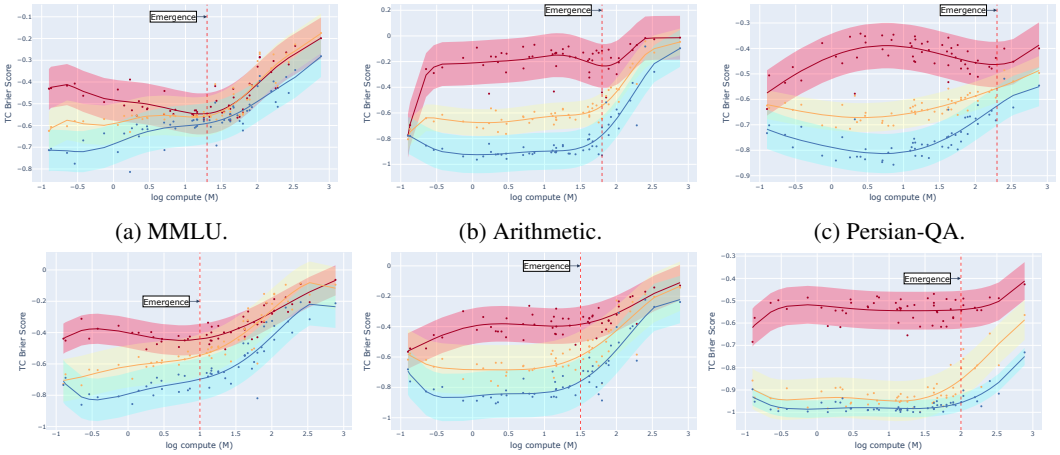
(c) U-shaped and inverted-U scaling.

Figure 3: The accuracy, TC Brier Score, U-shaped and inverted-U scaling on the arithmetic dataset in BIG-bench (Srivastava et al., 2023).

trend measured by the TC Brier Score. Though the scaling trend on the Persian-QA dataset is smoother, MMLU and arithmetic still exhibit a sharp increase of the TC Brier Score past the emergence threshold. Fig. 1c–3c show the TC Brier Score scaling trend with group number $G = 10$. Implementation details and model details are in App. A. Model performance on easier questions, such as difficulty level 1 in Fig. 1c and Fig. 2c, displays an inverted-U shape followed by steady improvement, i.e., performance first increases and then worsens with scale, followed by a second ascent, aligning with the previously reported deep double descent[1] on testing loss (Nakkiran et al., 2021). Moreover, the reversion from inverse scaling to standard scaling roughly coincides with the emergence threshold T. On the contrary, performance on hard questions, such as difficulty level 10 in Fig. 2c and Fig. 3c, displays a U-shaped scaling trend (Wei et al., 2023; McKenzie et al., 2023): model performance decreases with scale in early stage and increases with scale when M gets larger. Besides MMLU, arithmetic, and Persian-QA, Fig. 4 shows U-shaped vs. inverted-U scaling of Hindu knowledge, conceptual combinations, and analogical similarity datasets in Big-Bench (Srivastava et al., 2023), totaling six datasets, with $G = 3$. Detailed results of the three datasets are in App. C.

Overall, the scaling trend of a group transitions from that of the easiest group (inverted-U followed by steady ascent) to that of the hardest group (U-shape) as we move from the easiest group to the hardest group. Because the initial scaling trends of easy questions and hard questions roughly offset each other when aggregated across all difficulty levels, performance stagnates until the scaling trend

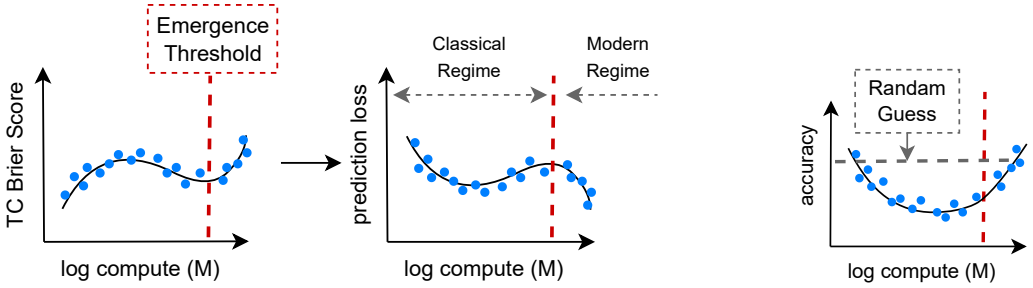
[1]The term "descent" in the original paper refers to the testing loss. Hence, for the sense of accuracy or Brier Score, it is an "ascent" of performance.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

(d) Hindu knowledge. (e) Conceptual combinations. (f) Analogical similarity.

Figure 4: U-shaped and inverted-U scaling on 6 datasets with emergent phenomenon, with group number $G = 3$. Except for MMLU, the other 5 tasks are from Big-Bench. Different levels of U-shaped and inverted-U scaling trends are demonstrated across the 6 tasks.



(a) Deep double descent. (b) U-shaped scaling.

Figure 5: Illustration of deep double descent (Nakkiran et al., 2021) on easy question groups and U-shaped scaling (Wei et al., 2023) on hard question groups under the TC Brier Score.

on easy questions reverts from inverse scaling to standard scaling, followed by a sharp improvement when performance on easy and hard questions both improve with scale. This could explain the emergent ability phenomenon reported in previous literature (Wei et al., 2022; Schaeffer et al., 2024a; Hu et al., 2023). More results of scaling trend on three non-emergent tasks in App. D, and U-shaped vs. inverted-U scaling measured by accuracy are in App. E.

3 Possible Explanation for U-shaped and Inverted-U Scaling

We provide a possible explanation for the initially opposing scaling trends (inverted-U vs. U-shaped) on easy vs. hard questions using the AI community's previous findings (Nakkiran et al., 2021; Wei et al., 2023; McKenzie et al., 2023) in deep neural networks (DNNs)' and specific LLMs' behaviors.

3.1 Scaling Trend of Easy Question Groups

As discussed in Sec. 2 and shown in Fig. 1–3, for a downstream task with emergent abilities, model performance on easy question groups first increases with scale, then decreases with scale, and finally reverts to increasing with scale. Fig. 5a illustrates the scaling trend if we flip the sign on the TC Brier Score so that a higher number means higher prediction loss. The pattern is then consistent with the deep double descent phenomenon identified in Nakkiran et al. (2021). In the context of testing error scaling law, Nakkiran et al. (2021) argues that initially, the bias-variance trade-off in the classical

Table 1: Examples of an easy and hard question in the MMLU benchmark. The Avg. Prob. is the average output probabilities before re-distribution over all models with log compute $M < 1.5$. Answer choices (classes) are underlined. In the hard question, small models overlook the negation "doesn't", giving choice C a high confidence, yet correct choice D a low confidence.

Question Description	Difficulty Level	Choices	Avg. Prob.
(conceptual physics, id = 44) The second law of thermodynamics tells us that heat doesn't flow from	level 10 (hardest group)	A. hot to cold ever B. cold to hot ever C. hot to cold without external energy D. cold to hot without external energy	A. 0.24 B. 0.29 C. 0.29 D. 0.18
(global facts, id = 66) In 1935 roughly how many Americans were in favor of Social Security act?	level 1 (easiest group)	A. 90% B. 70% C. 50% D. 30%	A. 0.44 B. 0.30 C. 0.17 D. 0.09

statistical learning theory (Hastie et al., 2009) applies, which forms the "classical regime": complex models suffer from "overfitting" and thus, once complexity exceeds a certain threshold, models become over-sensitive to sample noises, and the effect from such bigger variance dominates the effect of further reducing testing error. On the other hand, once the model is big enough (the "modern regime"), further increase in complexity allows the model to pick from more and more interpolating models that all fit the dataset, thereby improving performance and reducing testing error to near zero.

3.2 Scaling Trend on Hard Question Group

In contrast to the easy question groups, performance in hard question groups exhibits U-shaped scaling. McKenzie et al. (2023); Wei et al. (2023) have identified U-shaped scaling of LLM performance in some downstream tasks, as illustrated in Fig. 5b. Wei et al. (2023) provides a potential explanation for the initial inverse scaling: these tasks might contain a "distractor task" that attracts models to learn to solve at first, and thus larger models perform worse. One such example is the NeQA task (McKenzie et al., 2023) , which negates each multiple-choice question in the OpenBookQA dataset (Mihaylov et al., 2018) to examine whether models would be misled by the negation. It turns out that model performance would first decline from random guesses because of the attempt to answer the non-negation part of the question. Table. 1 shows such a question in our hard question group in the MMLU benchmark. For the question "The second law of thermodynamics tells us that heat doesn't flow from", small models (log compute $M < 1.5$) on average assign high confidence to choice C and lowest confidence to the correct choice D, and the former is the answer if removing negation "doesn't" from the original question.

4 Slice-and-Sandwich

4.1 Problem Formulation

We aim to predict the performance soar of traditional metrics before it happens. Specifically, we want to use only data before the emergent threshold to forecast the incidence of emergent abilities and the scaling trend past the emergent threshold. We compare the performance of our pipeline with the current iconic baseline of Sigmoid-based task-specific scaling law (Ye et al., 2023), which uses the Sigmoid function to regress accuracy on models' log compute M.

4.2 Pipeline Overview

Fig. 6 shows the overall pipeline of *Slice-and-Sandwich*. We use models smaller than the emergence threshold T as the training set. As performance no longer stagnates with scale once we group

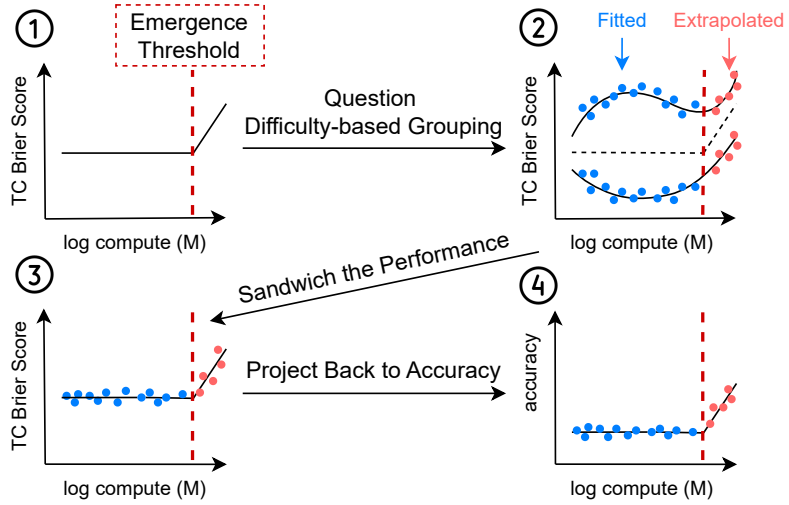


Figure 6: The overall pipeline of *Slice-and-Sandwich*. We group questions into different difficulty levels, fit each group's scaling trend, sandwich the overall performance to construct the scaling law on the linear metric, and finally project the scaling law back to the traditional metric.

questions by difficulty level, we fit the scaling trend of a continuous metric (TC Brier score in this paper) on the easiest question group and hardest question group separately and use the fitted scaling trend to forecast performance (measured in TC Brier Score) on easy and hard questions past T. We also use the training set to regress accuracy on the TC Brier Score and then use this estimated relation to convert the predicted TC Brier Score into predicted accuracy for models past the T.

4.3 Predicting Emergent Ability

4.3.1 Question Grouping

To reduce data noise, we group questions into $G = 3$ difficulty levels, as in Fig. 4, for *Slice-and-Sandwich* and denote the level 1, 2 and 3 questions as easy, medium, and hard question groups. The medium group's pattern is close to aggregating the scaling trend between easier and harder groups.

4.3.2 Fitting and Forecasting Scaling Trend of Easy vs. Hard Questions

We use simple polynomial regression to fit the scaling trend of the TC Brier Score of the easy and hard question groups separately, using models before the emergence threshold T. We denote by $F_e^c(x)$ and $F_h^c(x)$ the fitted scaling trend of the easy and hard question groups, respectively, where x is the log compute. We then use $F_e^c(x)$ and $F_h^c(x)$ to forecast performance (measured in TC Brier Score) on the easy and the hard question groups of models with log compute x above T.

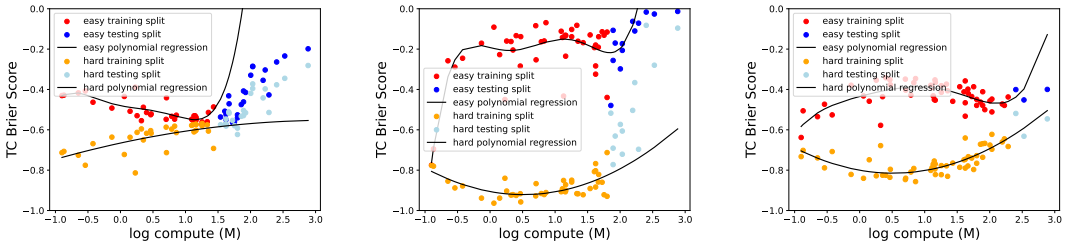
We use the average of performance on the easy group and the hard group to forecast aggregated performance measured in TC Brier Score:

$$F^c(x) = \frac{1}{2}(F_e^c(x) + F_h^c(x)), \tag{6}$$

as aggregate performance is sandwiched between performances in the easy and hard groups.

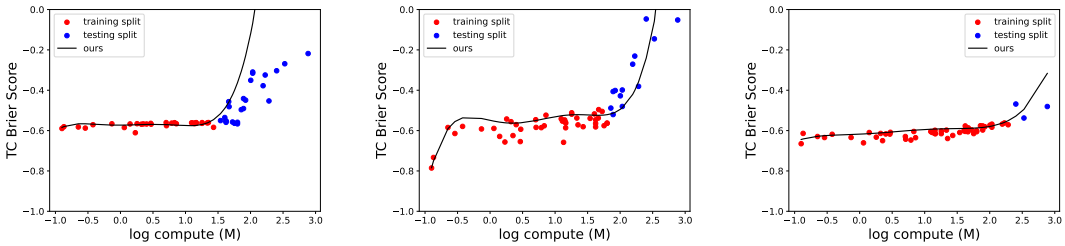
4.3.3 Obtaining Scaling Trend in Traditional Metric

Since what people usually care about ultimately are those traditional metrics such as accuracy (Hu et al., 2023), our last step is to project the forecast scaling trend in TC Brier Score, $F^c(x)$, back to scaling trend in accuracy, denoted by $F^t(x)$. One can replace TC Brier Score with other continuous metrics and accuracy with other traditional metrics. Specifically, we first estimate the relation between the continuous metric (TC Brier Score) and the traditional metric (accuracy) using models with log



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure 7: Data and polynomial fit for the easy and hard question groups on the MMLU, arithmetic, and Persian-QA datasets. The fitted trends encapsulate the actual trends.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure 8: The TC-Brier-Score-based scaling law on the MMLU, arithmetic, and Persian-QA datasets acquired by taking the average of fitted trends of easy and hard question groups in Fig. 7.

computes smaller than the emergent threshold T as the training set. We denote the estimated mapping from TC Brier Score to accuracy as $G(\cdot)$. Our forecast of scaling trend of accuracy is given by:

$$F^t(x) = G(F^c(x)) + C, \tag{7}$$

where C is a constant such that the average predicted accuracy of $F^t(x)$ on the training set is the same as the average true accuracy of all models in the training set.

5 Experiments

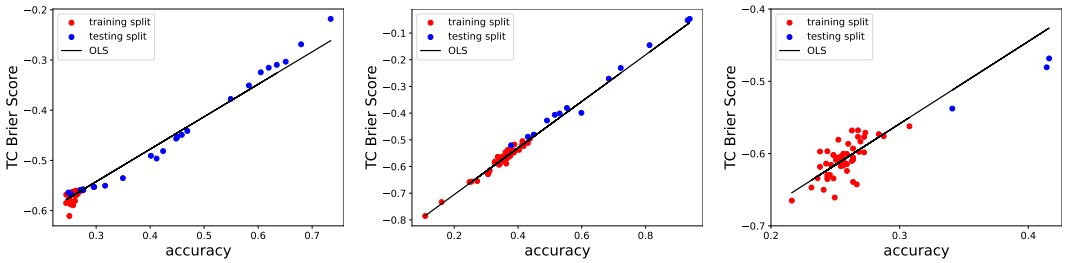
5.1 Fitting Scaling Trend of Easy Group and Hard Group

We adopt polynomial degree=2 and 5 for hard and easy questions, respectively, in response to our observation of U-shaped vs. inverted-U scaling. This parameter selection is based on our prior belief of polynomial regression's fitting powers to fit the deep double descent and U-shaped scaling. Experimental results on parameter robustness are in App. F.

Fig. 7 shows the fitted scaling trend of the easy and hard question groups on the MMLU, arithmetic, and Persian-QA datasets. Empirically, fitted trends on hard questions are either precise or underestimated, e.g., hard group of MMLU (Fig. 7a) and arithmetic (Fig. 7b), due to lower fitting power of degree 2; fitted trends on easy questions are precise or overestimated, e.g., easy group of MMLU (Fig. 7a), due to a more considerable fitting power of degree 5. However, they still encapsulate the overall trend. Therefore, as shown in Fig. 8, taking their average can decrease the deviation and still lead to a precise prediction of the actual scaling trend.

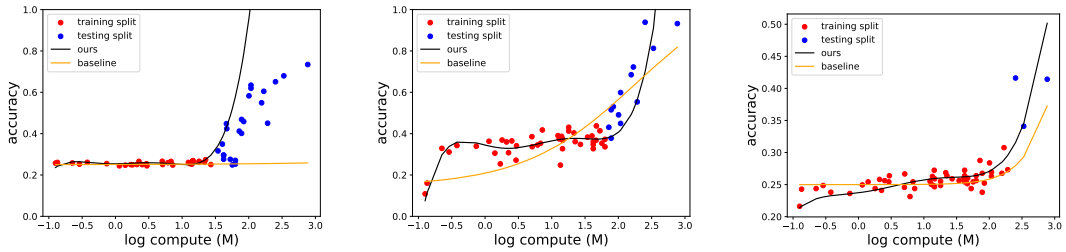
5.2 Relation Between Accuracy and Target-Conditioned (TC) Brier Score

Fig. 9 shows the close and almost linear relation between accuracy and TC Brier Score. As a result, simple ordinary least squares (OLS) regression using only models before the emergence threshold yields a precise mapping $G(\cdot)$ from TC Brier Score to accuracy.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure 9: The relation between accuracy and the TC Brier Score on the MMLU, arithmetic, and Persian-QA datasets. The mapping function $G(x)$ from the TC Brier Score to accuracy can be well-modeled using small models as the training set.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure 10: The accuracy-based scaling law on the MMLU, arithmetic, and Persian-QA datasets acquired by projecting the TC-Brier-Score-based scaling law back to accuracy-based scaling law by $G(x)$. Baseline is the Sigmoid-based regression (Owen, 2024).

5.3 Forecasting Scaling Trend in Accuracy

Finally, Fig. 10 shows the accuracy-based scaling trend, $F^t(x)$, obtained through Eq. 7 with $G(\cdot)$, together with the baseline of fitting performance measured in accuracy on the Sigmoid function (Owen, 2024; Ruan et al., 2024). Compared with the baseline that assumes the monotone Sigmoid scaling trend, our *Slice-and-Sandwich* better predicts and estimates the soaring performance by baking in more priors of the observed U-shaped and inverted-U scaling. For the MMLU benchmark, our approach captures the forthcoming soaring trend, whereas the baseline approach does not. For the arithmetic dataset, though the baseline provides a seemingly decent forecast, it does not capture the sharp increase in the improvement speed at all, whereas our approach does. In short, our *Slice-and-Sandwich* approach could be more explainable and capable of capturing the soaring trends of emergent abilities. A simple alternative method of *Slice-and-Sandwich* and its experimental results are in App. F.

6 Conclusions and Limitations

This work proposes to separately analyze LLM's task-specific scaling trends by question grouping based on difficulty level. For six multiple-choice tasks with emergent abilities, we demonstrate U-shaped scaling for hard questions and inverted-U scaling followed by steady improvement for easy questions. These findings provide insights into the potential causes of ability emergence. We then introduce the *Slice-and-Sandwich* pipeline to predict the ability emergence and forecast scaling trends thereafter. However, since emergent phenomena have been widely reported in many tasks across a bunch of LLM benchmarks, it might be hard to claim all of them must exhibit clear U-shaped vs. inverted-U scaling. Furthermore, this study primarily focuses on multiple-choice tasks. Applying our method to string-matching tasks requires identifying a continuous metric that differentiates easy questions from hard questions and is highly correlated with the traditional metric people are ultimately interested in. We demonstrate this point in our preliminary analysis for string-matching tasks in App. G, believing it a valuable avenue for future work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL `https://doi.org/10.5281/zenodo.5297715`.

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

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

PV Charan, Hrushikesh Chunduri, P Mohan Anand, and Sandeep K Shukla. From text to mitre techniques: Exploring the malicious use of large language models for generating cyber attack payloads. *arXiv preprint arXiv:2305.15336*, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Together Computer. Redpajama: an open dataset for training large language models, October 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

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

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL `https://zenodo.org/records/12608602`.

Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL `https://github.com/openlm-research/open_llama`.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Proceedings of Advances in neural information processing systems (NeurIPS)*, volume 35, 2022.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

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. *arXiv preprint arXiv:2001.08361*, 2020.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022a.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9019–9052, 2022b.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research (TMLR)*, 2023. ISSN 2835-8856.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Proceedings of Advances in neural information processing systems (NeurIPS)*, 36, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment (JSTAT)*, 2021(12):124003, 2021.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

David Owen. How predictable is language model benchmark performance? *arXiv preprint arXiv:2401.04757*, 2024.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Proceedings of Advances in neural information processing systems (NeurIPS)*, 36, 2024a.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier ai models with scale remained elusive? *arXiv preprint arXiv:2406.04391*, 2024b.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David

Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen,

Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023. ISSN 2835-8856.

Stability-AI. Stablelm: Stability ai language models, April 2023. URL `https://github.com/Stability-AI/StableLM`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, May 2023. URL `www.mosaicml.com/blog/mpt-7b`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

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. *Transactions on Machine Learning Research (TMLR)*, 2022. ISSN 2835-8856.

Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. Inverse scaling can become U-shaped. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Qinyuan Ye, Harvey Fu, Xiang Ren, and Robin Jia. How predictable are large language model capabilities? a case study on BIG-bench. In *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

SUPPLEMENTARY MATERIAL

A Implementation Details **16**
A.1 LLM Evaluation 16
A.2 Slice-and-Sandwich 16

B More Discussions on Brier Score **17**

C Scaling Trend by Question Difficulty Level for Other Emergent Tasks **19**

D Scaling Trend by Question Difficulty Level for Non-emergent Tasks **22**

E Scaling Trend by Question Difficulty Level on Accuracy **25**

F More Discussions on Slice-and-Sandwich **26**
F.1 Robustness Analysis 26
F.2 Hard Lift - A Simple Alternative Pipeline 27

G Preliminary Analysis for String-Match Tasks **29**

H Broader Impact **30**
H.1 Potential Positive Impacts 30
H.2 Potential Negative Impacts 30

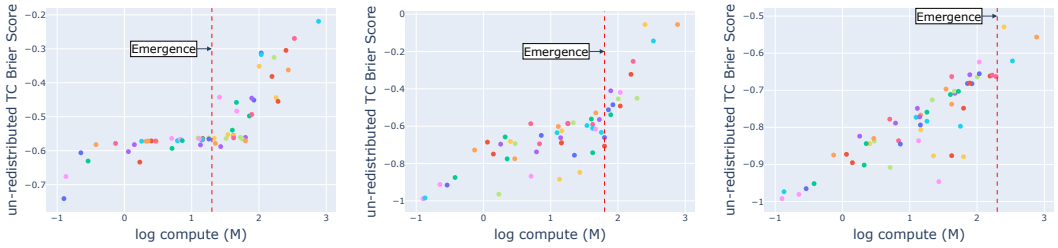
A IMPLEMENTATION DETAILS

A.1 LLM EVALUATION

We evaluate all datasets in this paper on the LM Evaluation Harness (Gao et al., 2024) platform. We adopt 56 models, including Gemma (Team et al., 2024), Llama (Touvron et al., 2023a), Llama-2 (Touvron et al., 2023b), RedPajama-INCITE (Computer, 2023), Yi (Young et al., 2024), StableLM (Stability-AI, 2023), MPT (Team, 2023), Falcon (Almazrouei et al., 2023), Pythia (Biderman et al., 2023), AMBER (Liu et al., 2023), Qwen (Bai et al., 2023), Qwen-1.5 (Bai et al., 2023), BLOOM (Workshop et al., 2022), DeepSeekMoE (Dai et al., 2024), OPT (Zhang et al., 2022), GPT-Neo (Black et al., 2021), Codegen (Nijkamp et al., 2023), XGLM (Lin et al., 2022b), and OpenLLaMA (Geng & Liu, 2023) families under FP16 precision. The evaluation time of each task varies from several hours to several days on 2 NVIDIA RTX A6000, depending on the question numbers and formats. We obtain each model's log compute through the released data by Ruan et al. (2024). We use $T = 1.5, 1.8$, and 2.3 as the emergence threshold for the MMLU, arithmetic, and Persian-QA dataset, respectively. We calculate question difficulty level q_d using models smaller than these thresholds. We adopt a 5-shot inference on the MMLU benchmark and a 2-shot inference on the arithmetic and Persian-QA datasets.

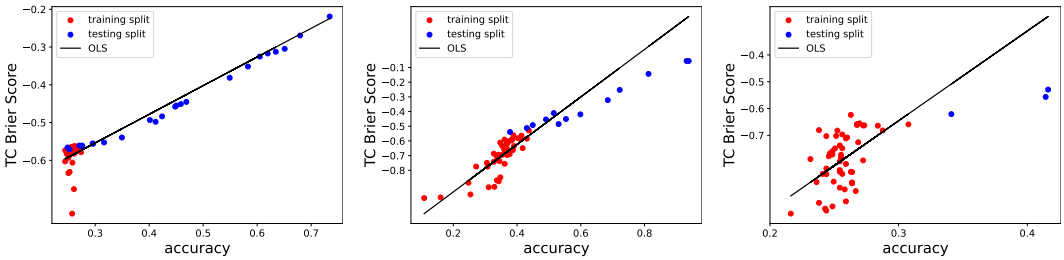
A.2 SLICE-AND-SANDWICH

We examine *Slice-and-Sandwich* on MMLU, arithmetic, and Persian-QA datasets with group number $G = 3$. Models smaller than $T = 1.5, 1.8$, and 2.3 in the MMLU, arithmetic, and Persian-QA datasets are the training set; other larger models are the testing set. We adopt polynomial regression to fit easy and hard question groups. Specifically, we adopt the polynomial order=5 and 2 for the easy and hard question groups, respectively.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure A11: The un-conditionalized TC Brier Score vs. log compute (M) on the MMLU, arithmetic, and Persian-QA datasets.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure A12: The relation between accuracy and un-conditionalized TC Brier Score on the MMLU, arithmetic, and Persian-QA datasets.

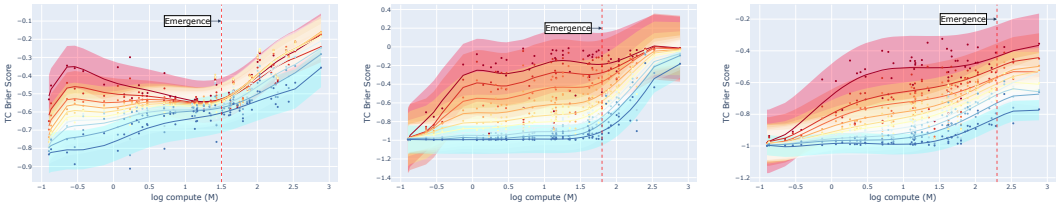
B More Discussions on Brier Score

In the main paper, we use the model's predicted probability of the correct class *conditional* on all classes to calculate the TC Brier Score (see Eq. 4). This section discusses the effect of such conditionalization. This section refers to the *un-conditionalized TC Brier Score* as the one without re-distributing output probabilities to all classes.

Fig. A11 shows the relationship between un-conditionalized TC Brier Score and log compute M on all three datasets. For the MMLU dataset, model performance still exhibits flat scaling before the emergence threshold and sharp improvement past the emergence threshold. For the arithmetic and Persian-QA datasets, the scaling trend does not show a sharp increase and is easier to forecast performance under the un-conditionalized TC Brier Score past the emergence threshold. This is consistent with the finding of (Schaeffer et al., 2024b) that the un-conditionalized measure is more correlated with the training compute than conditionalized ones. However, Fig. A12 shows that the un-conditionalized TC Brier Score is not as closely related to accuracy as the normal TC Brier Score for the arithmetic and especially the Persian-QA dataset. Table A2 corroborates this assertion by showing the correlation coefficient between accuracy and the normal/un-conditionalized TC Brier Score.

Table A2: Comparison of correlation coefficients between accuracy and TC Brier Score with and without conditionality on the MMLU, arithmetic, and Persian-QA datasets. "P.", "S.", and "K." stands for Pearson, Spearman, and Kendall, respectively. The TC Brier Score with conditionality, i.e., the one we adopt in the main paper, has a consistently stronger correlation with accuracy.

	MMLU			Arithmetic			Persian-QA		
Correlation coefficient	P.	S.	K.	P.	S.	K.	P.	S.	K.
un-conditionalized TC Brier Score	0.96	0.87	0.73	0.93	0.93	0.79	0.60	0.52	0.37
TC Brier Score	**0.99**	**0.91**	**0.79**	**1.00**	**0.97**	**0.88**	**0.88**	**0.68**	**0.52**



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

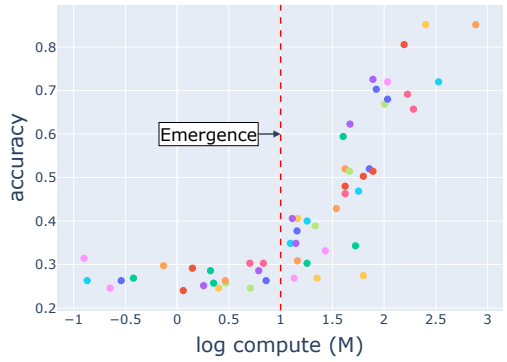
Figure A13: The U-shaped and inverted-U scaling with questions grouped and performances measured by the un-conditionalized TC Brier Score on the MMLU, arithmetic, and Persian-QA daatasets.

Fig. A13 shows the scaling trend by difficulty level for the un-conditionalized TC Brier Score. For the arithmetic and Persian-QA datasets, we no longer see inverse scaling on any intervals of log compute M. In fact, for both the arithmetic and the Persian-QA datasets, performance hovers around -1 on the hardest group, corresponding to the near-zero predicted probability of the correct class. For hard questions, the model's predicted probability on all classes is close to zero. Therefore, without conditionalizing on all classes, we cannot differentiate between an initial random guess and the distracted phase at larger model log computes where the model places a higher probability on an available incorrect class relative to the correct class, which yields the U-shaped scaling of normal TC Brier Score as discussed in the main paper.

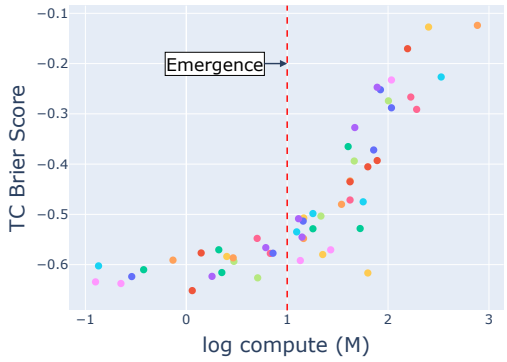
C Scaling Trend by Question Difficulty Level for Other Emergent Tasks

This section demonstrates U-shaped and inverted-U scaling on three more tasks with emergent abilities, besides the MMLU, arithmetic, and Persian-QA datasets in the main paper. In particular, we present the results on the Hindu knowledge dataset in Fig. A14, conceptual combinations dataset in Fig. A15, and analogical similarity dataset in Fig. A17. These datasets are all in BIG-bench (Srivastava et al., 2023).

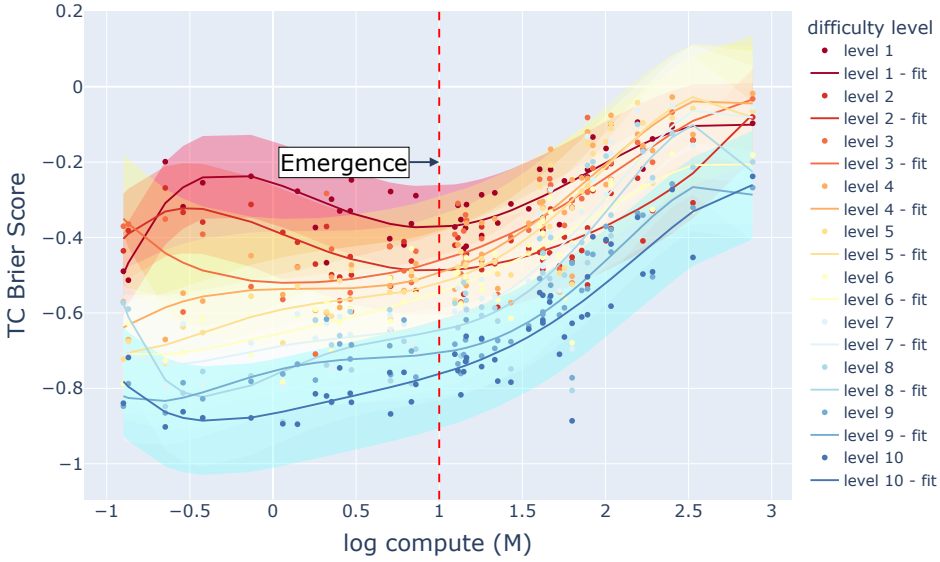
In particular, the Hindu knowledge and conceptual combinations datasets display the U-shaped scaling for easy question groups and inverted-U scaling hard question groups. The analogical similarity dataset also shows U-shaped scaling, albeit very mild, for hard question groups, and inverted-U scaling for easy question groups. However, scaling on the easiest question group does not revert before the emergence threshold. Scaling on the second easiest and the third easiest group reverts from inverse scaling to standard scaling way before the emergence threshold. The overall scaling trend for accuracy (see Fig. A17a) actually declines slightly with scale. However, Fig. A16 shows that, though a bit overestimated, we can still predict the forthcoming of emergent abilities using *Slice-and-Sandwich*, whereas the Sigmoid-based regression yields a flat line.



(a) Accuracy vs. log compute (M).

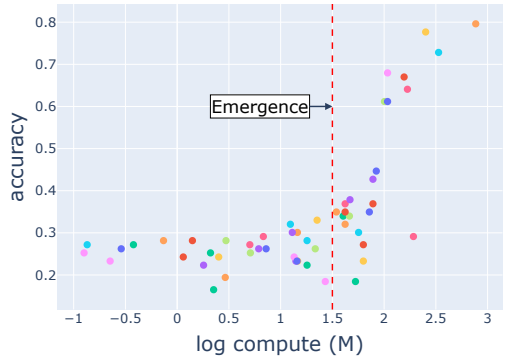


(b) TC Brier Score vs. log compute (M).

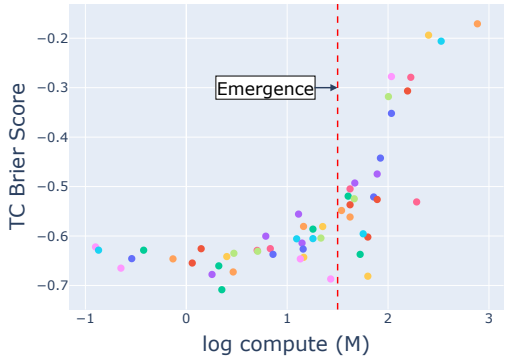


(c) U-Shaped and inverted-U scaling

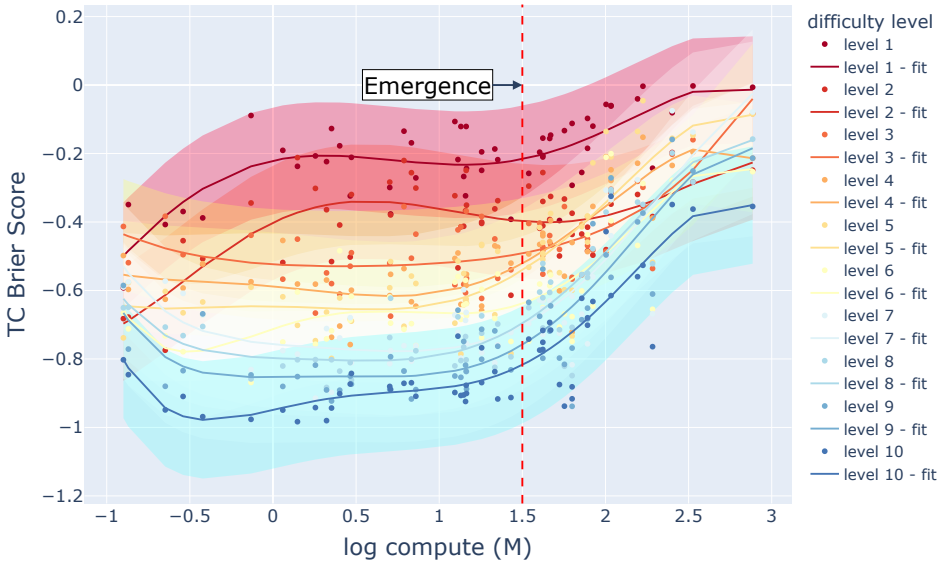
Figure A14: The accuracy, TC Brier Score, U-Shaped and inverted-U scaling on the Hindu knowledge dataset in BIG-bench (Srivastava et al., 2023).



(a) Accuracy vs. log compute (M).



(b) TC Brier Score vs. log compute (M).



(c) U-Shaped and inverted-U scaling

Figure A15: The accuracy, TC Brier Score, U-Shaped and inverted-U scaling on the conceptual combinations dataset in BIG-bench (Srivastava et al., 2023).

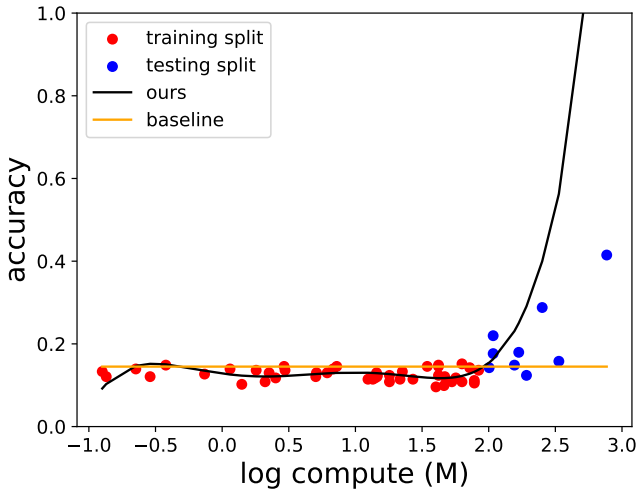
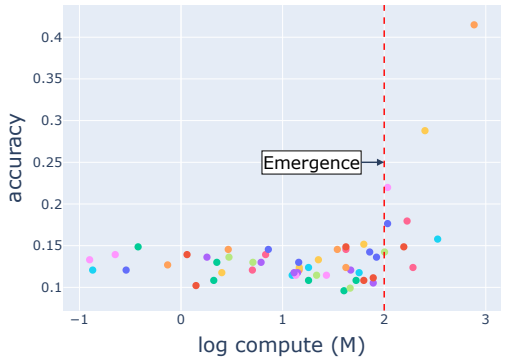
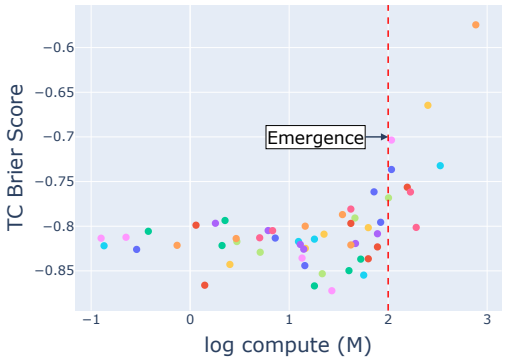


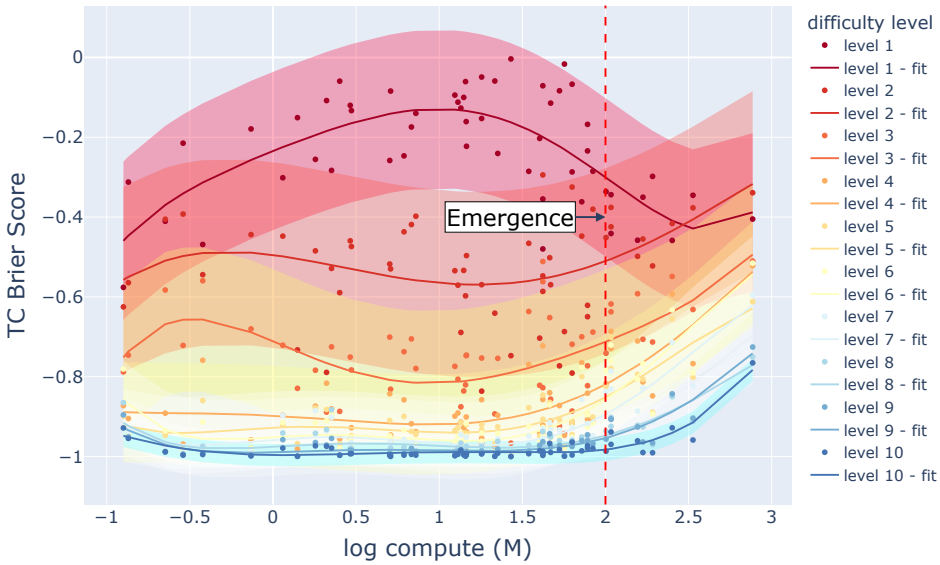
Figure A16: The accuracy-based scaling law on the analogical similarity dataset in BIG-bench (Srivastava et al., 2023).



(a) Accuracy vs. log compute (M).



(b) TC Brier Score vs. log compute (M).

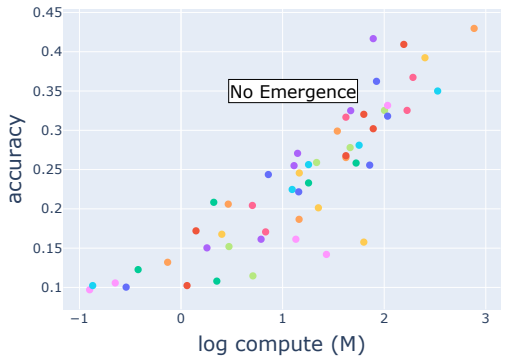


(c) U-Shaped and inverted-U scaling

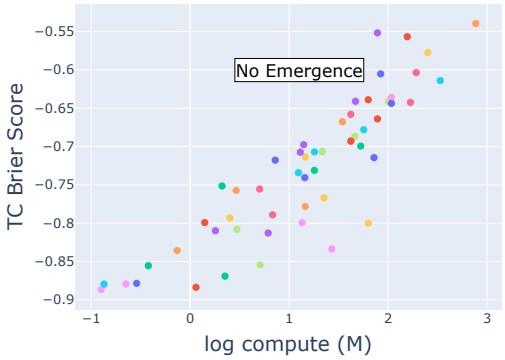
Figure A17: The accuracy, TC Brier Score, U-Shaped and inverted-U scaling on the analogical similarity dataset in BIG-bench (Srivastava et al., 2023).

D SCALING TREND BY QUESTION DIFFICULTY LEVEL FOR NON-EMERGENT TASKS

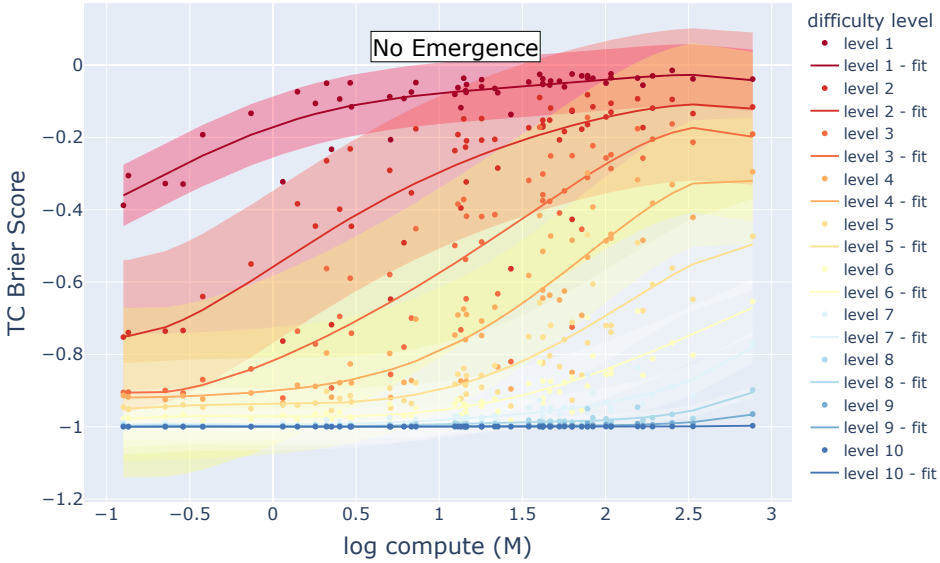
We apply the same procedure as in Sec. 2 to several multiple-choice tasks without emergent abilities, i.e., tasks for which performance improves consistently with scale. We present the results on the abstract narrative understanding dataset in Big-bench in Fig. A18, ARC dataset (Clark et al., 2018) in Fig. A19, and HellaSwag dataset (Zellers et al., 2019) in Fig. A20. Interestingly, we do not observe the U-shaped and inverted-U scaling as in the MMLU, arithmetic, and Persian-QA datasets. Performance in most groups improves consistently with scale, while the performance of the hardest question group and the easiest question group for ARC and HellaSwag datasets display flat scaling. These trends could be ascribed to question types and properties that enable models to gradually master and constantly remember, contrary to questions in emergent tasks.



(a) Accuracy vs. log compute (M).

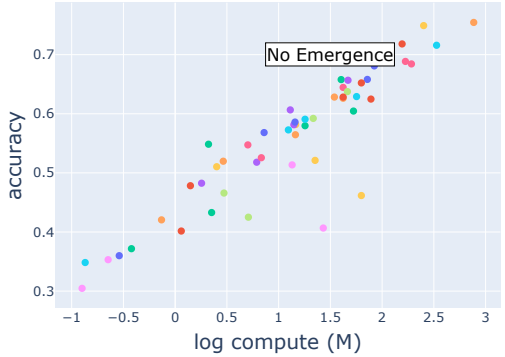


(b) TC Brier Score vs. log compute (M).

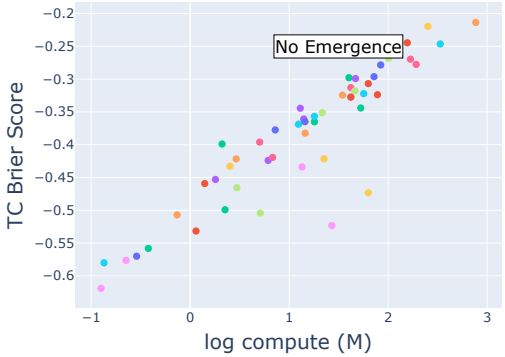


(c) Scaling trend.

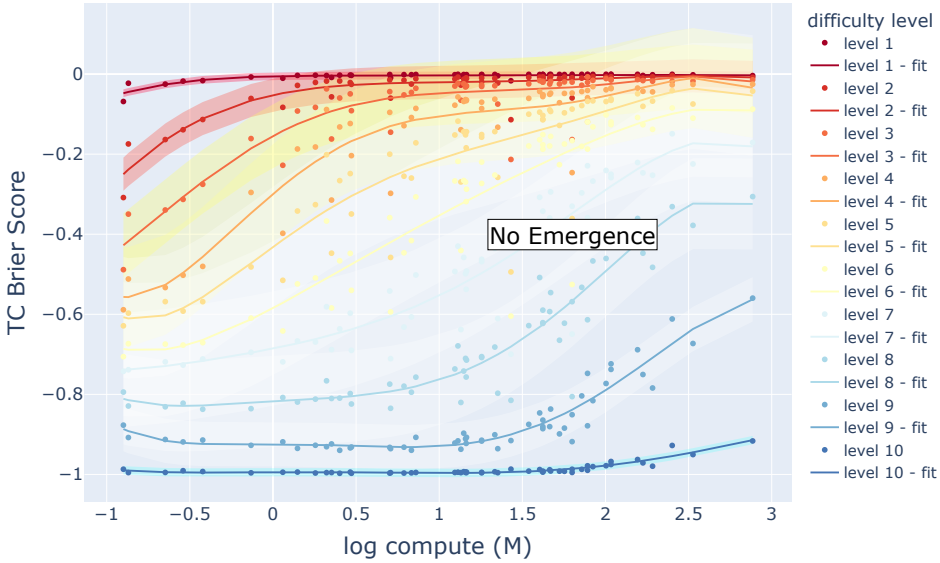
Figure A18: The accuracy, TC Brier Score, and scaling trend on the abstract narrative understanding dataset in BIG-bench (Srivastava et al., 2023).



(a) Accuracy vs. log compute (M).

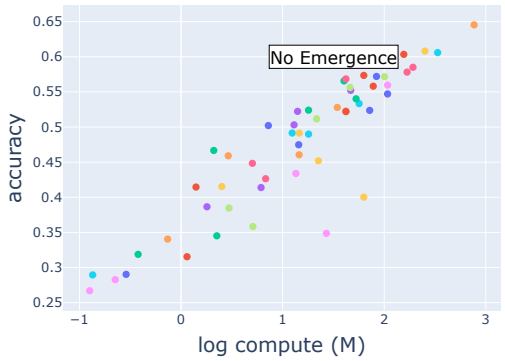


(b) TC Brier Score vs. log compute (M).

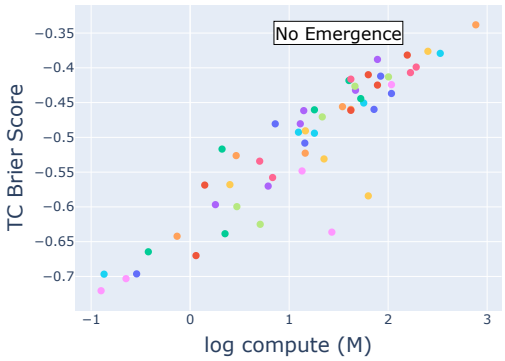


(c) Scaling trend.

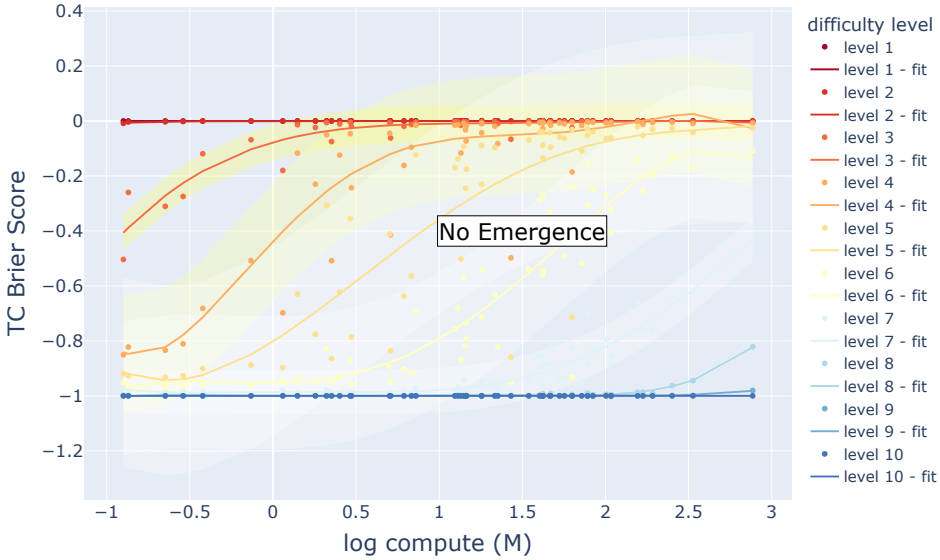
Figure A19: The accuracy, TC Brier Score, and scaling trend on the ARC dataset (Clark et al., 2018).



(a) Accuracy vs. log compute (M).

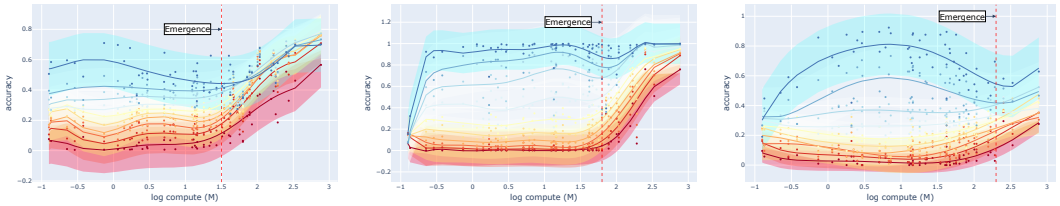


(b) TC Brier Score vs. log compute (M).



(c) Scaling trend.

Figure A20: The accuracy, TC Brier Score, and scaling trend on the HellaSwag dataset (Zellers et al., 2019).

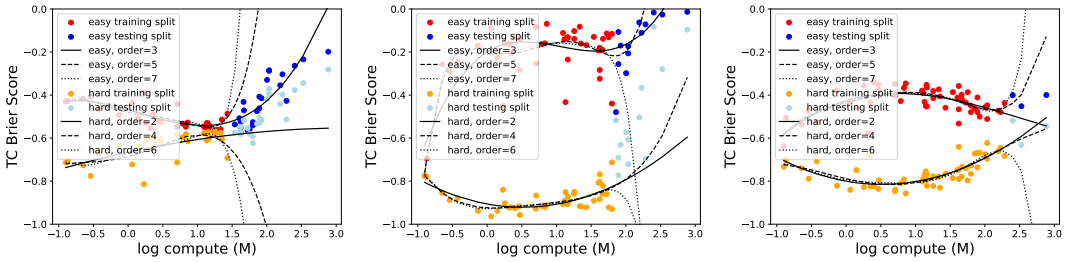


(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure A21: The U-shaped and inverted-U scaling of accuracy with group number $G = 10$.

E Scaling Trend by Question Difficulty Level on Accuracy

We apply the same procedure as in Sec. 2 with accuracy as the performance measure instead of the TC Brier Score. Specifically, we calculate question difficulty level using average accuracy over models before the emergence threshold and plot the scaling trend on accuracy for each difficulty group, as shown in Fig. A21. We still observe clear inverted-U scaling for the easiest question group followed by steady improvement after the emergence threshold. However, U-shaped scaling for the hard question groups becomes unclear. Specifically, the accuracy performance of hard question groups tends to stagnate after the initial performance drop. For instance, all three datasets' hardest hard question groups consistently stuck at near-zero accuracy, lower than the random guess. The worse-than-random performance can be explained by distracting questions, as discussed in Sec. 3.2. On the other hand, the mitigated U-shaped scaling might be due to the fact that accuracy does not directly capture the target class' confidence values. In other words, the accuracy-based procedure cannot demonstrate the models' learning process of first being distracted by questions and gradually overcoming the distraction, where the models' accuracies are all around zero.



(a) MMLU. (b) Arithmetic. (c) Persian-QA.

Figure A22: Data and polynomial fit of different degrees for easy and hard groups.

F More Discussions on Slice-and-Sandwich

F.1 Robustness Analysis

We present the robustness analysis of *Slice-and-Sandwich* regarding (1) the choice of order, (2) lower model log compute cutoff for the training set, and (3) group number G.

F.1.1 Effect of Polynomial Degree

Fig. A22 shows the polynomial fit of TC Brier Score for the easy group with degree=3, 5, and 7, and for the hard group with degree=2, 4, and 6. Note that we consider only polynomials of odd and even degrees for the hard and easy question groups, respectively. This prior knowledge reflects the observation that performance of the easy question group initially improves with scale, the performance of the hard question group initially declines with scale, whereas the performance of both groups increases with scale past the emergence threshold. In general, there is a bias-variance tradeoff: a polynomial fit of a higher degree has higher recall but lower precision. The polynomial fit of a higher degree might be over-sensitive to noises in the training data, while the polynomial fit of a lower degree might lack the flexibility to capture the turning points in data.

In Fig. A22, we find that the polynomial fit of degree 3 and 5 forecast the scaling trend of the easy question group well except polynomial fit of degree 3 for the Persian-QA dataset, while the polynomial fit of degree 2 forecasts the scaling trend of the hard question group well. On the other hand, polynomial fit of higher degrees, in particular, degree 7 for the easy question group and degrees 4 and 6 for the hard question group, do not forecast the scaling trend well. We leave it for future work to explore better functional forms to model U-shaped scaling for the hard group and inverted-U scaling with steady improvement (deep double descent) for the easy group.

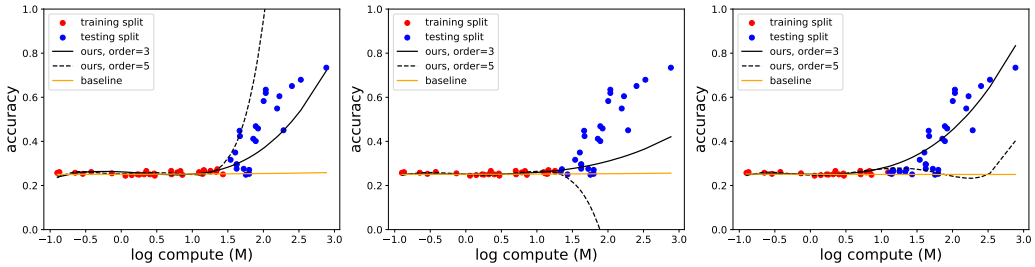
F.1.2 Effect of Log Compute Threshold for Train-Test Split

Fig. A23 shows the fitted scaling trend using different train-test splitting thresholds. For the hard question group, we use a polynomial fit of degree 2. For the easy question group, the polynomial fit of degree 3 is represented by a black solid line, and the polynomial fit of degree 5 is represented by a black dashed line.

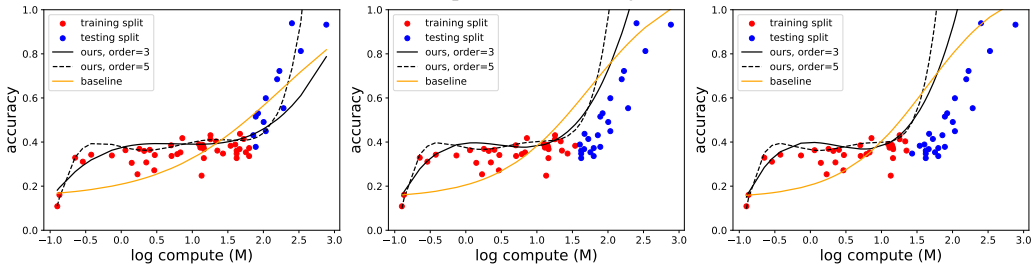
The forecast is reasonably robust to the train-test split. All capture the trend and display a similar shape to our original choice of train-test split threshold except for the case where threshold$= 1.3$ and degree$= 5$ for the MMLU dataset. The polynomial fit of degree 3 for Persian QA is too flat compared to data for all three thresholds and gets flatter as the threshold goes down. We leave it to future work to provide better guidelines as to the least upper bound of training data model log compute that still allows us to confidently predict the onset of emergent abilities.

F.1.3 Effect of Group Number

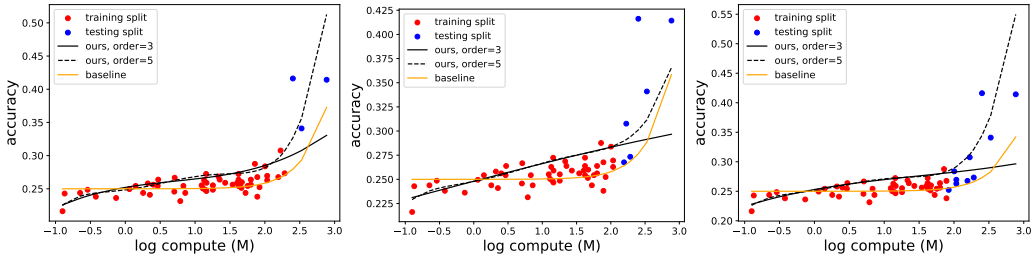
Fig. A24 shows the fitted scaling trends when splitting questions into different numbers of groups. We show group number $G = 3, 5$, and 7. Following the same procedure and degree parameter in the main paper, easier question groups, such as the groups of difficulty level 1 to 3 for $G = 7$, are fitted



(a) MMLU. Train-test split threshold being 1.5, 1.3, and 1.1.



(b) Arithmetic. Train-test split threshold being 1.8, 1.6, and 1.4.



(c) Persian-QA. Train-test split threshold being 2.3, 2.1, and 1.9.

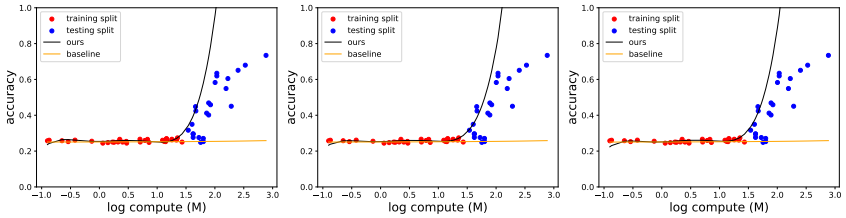
Figure A23: *Slice-and-Sandwich*'s results of accuracy-based scaling law under different train-test split thresholds. Solid lines are when order=3 is used for easy question fitting, and dashed lines are when order=5 is used.

by polynomial regression of degree=5 due to observed inverted-U scaling; harder question groups are fitted by degree=2 due to U-shaped scaling. Then Eq. 6 is modified to take the average of Brier-based fitting trends of all but the medium group and project the acquired Brier-based scaling law to the accuracy-based one. *Slice-and-Sandwich* shows its robustness under G. The robustness comes from similar fitting results among the same scaling types, such as inverted-U scaling, resulting in a similar final scaling law after taking their averages.

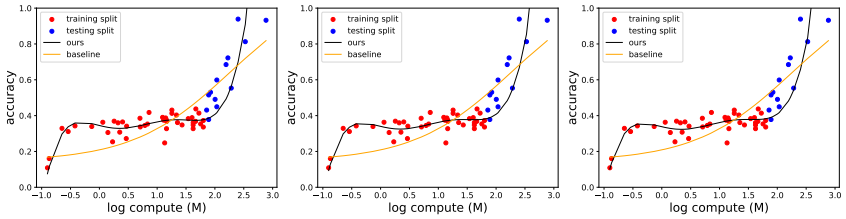
F.2 Hard Lift - A Simple Alternative Pipeline

As an alternative to *Slice-and-Sandwich*, we provide an even simpler pipeline called *Hard-Lift*. Specifically, we take the polynomial fit of degree 2 on TC Brier Score for the hard question group from *Slice-and-Sandwich* and lift it by a constant so the fitted TC Brier Score at the training set model log compute upper bound is equal to the true average. We use this to forecast the TC Brier Score of models past the emergence threshold. We then transform this predicted TC Brier Score back to predicted accuracy via the $G(\cdot)$ function as in *Slice-and-Sandwich*.

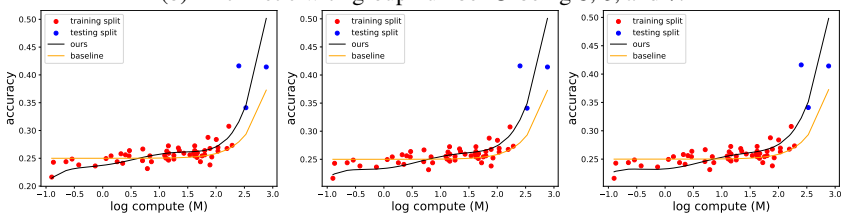
Fig. A25 shows the results of *Hard-Lift* under different log compute thresholds for train-test split as in Sec. F.1.2. *Hard-Lift* performs better than the baseline for MMLU (Fig. A25a) and Persian-QA (Fig. A25c) datasets, but worse than baseline for the arithmetic dataset. We believe this result reinforces our claim that analyzing difficulty-stratified scaling trends enables more explainable prediction of emergent abilities.



(a) MMLU with group number G being 3, 5, and 7.

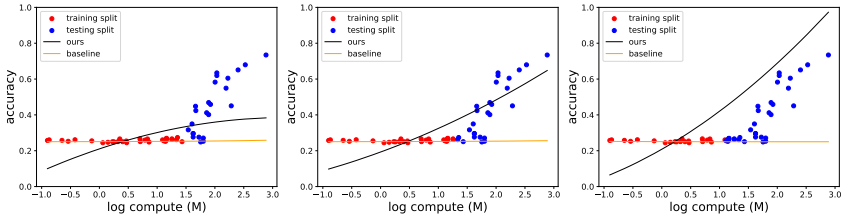


(b) Arithmetic with group number G being 3, 5, and 7.

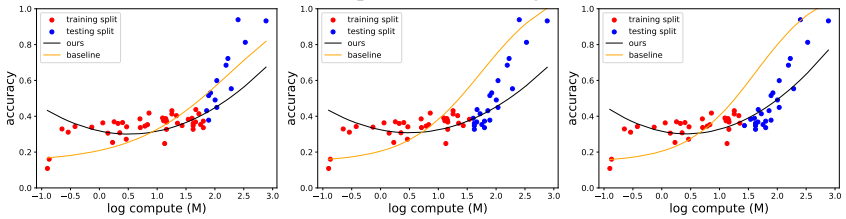


(c) Persian-QA with group number G being 3, 5, and 7.

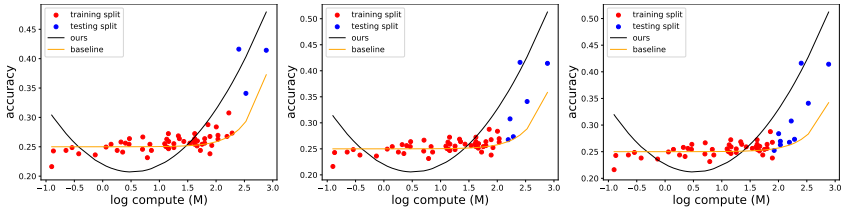
Figure A24: *Slice-and-Sandwich*'s results of accuracy-based scaling law under different group numbers G.



(a) MMLU. Train-test split threshold being 1.5, 1.3, and 1.1.

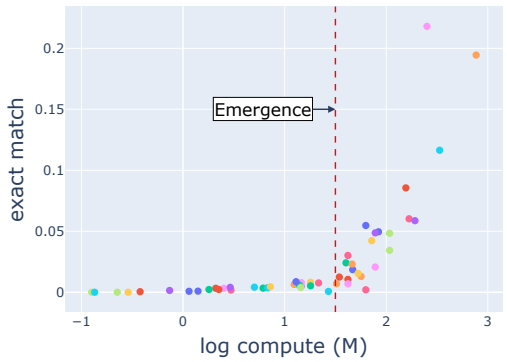


(b) Arithmetic. Train-test split threshold being 1.8, 1.6, and 1.4.

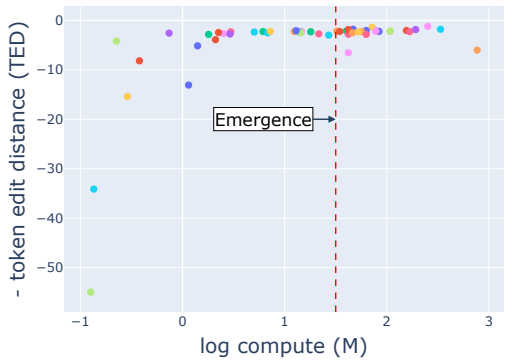


(c) Persian-QA. Train-test split threshold being 2.3, 2.1, and 1.9.

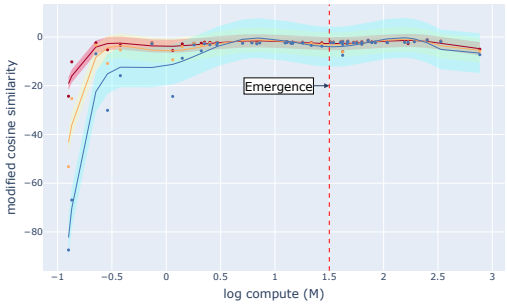
Figure A25: *Hard-Lift*'s results of accuracy-based scaling law under different train-test split thresholds.*Hard-Lift* uses order=2 for fitting hard question groups.



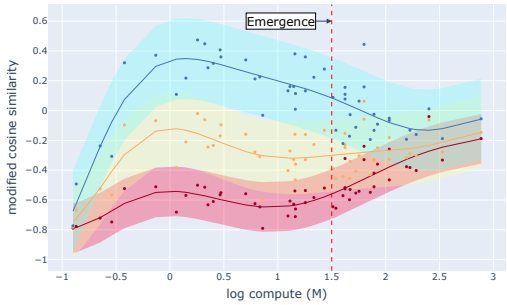
(a) Exact match accuracy vs. log compute (M).



(b) Token edit distance vs. log compute (M).

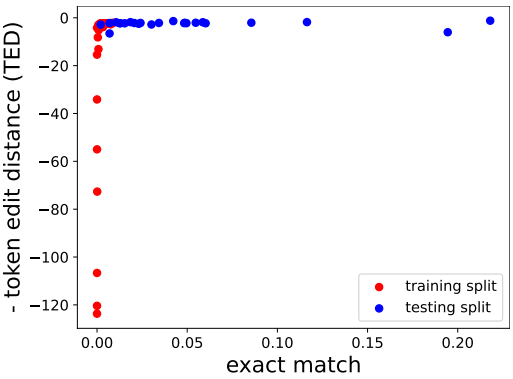


(c) token edit distance vs. log compute (M) with group number $G = 3$.

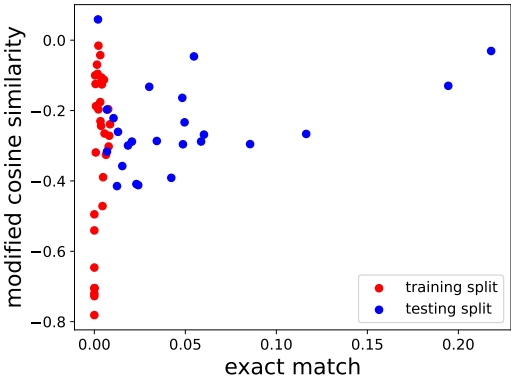


(d) Modified cosine similarity vs. log compute (M) with group number $G = 3$.

Figure A26: The exact match accuracy, token edit distance (TED), and cosine similarity score vs. log compute (M) on the word unscramble dataset in BIG-bench (Srivastava et al., 2023).



(a) Relation between token edit distance and exact match.



(b) Relation between modified cosine similarity and exact match.

Figure A27: Relations between token edit distance/modified cosine similarity and exact match on the word unscramble dataset in BIG-bench (Srivastava et al., 2023).

G PRELIMINARY ANALYSIS FOR STRING-MATCH TASKS

This section provides the preliminary analysis for the exact string match tasks.

Fig. A26a shows that the word unscramble dataset in BIG-bench (Srivastava et al., 2023) exhibits emergent abilities under the traditional metric: exact match accuracy. On the other hand, Fig. A26b shows that model performance measured by token edit distance (TED) as discussed in Schaeffer et al. (2024a) improves with scale steadily at first and then exhibit flat scaling.

We argue that TED is not a good measure of progress on a string match task. (1) It does not differentiate between easy and hard questions well. Fig. A26c shows that performance measured by TED on all three question groups is close for all model log computes above 0.5. A harder group's TED may be higher or lower than an easier group's. (2) It is not very correlated with exact match, the traditional metric that people are probably ultimately interested in (A27a).

One idea is to measure performance by modified cosine similarity (MCS):

$$MCS = \frac{F(s1) \cdot F(s2)}{\|F(s1)\|\|F(s2)\|} \cdot \mathbb{I}(s1 \subseteq s2), \tag{8}$$

where $s1$ is the model's output string, $s2$ is the answer string, $F(x)$ is CLIP (Radford et al., 2021)'s text encoder to project the string to the vector space, and $\mathbb{I}(x)$ is an indicator function having 1 if every single character of $s1$ is contained in $s2$, otherwise 0. MCS takes values in the interval $[-1, 1]$ and is good at differentiating questions by difficulty levels. Fig. A26d shows that MCS scaling curves of the easy, medium, and hard question groups are clearly ordered.

Interestingly, performance measured by MCS for all question groups exhibits inverted-U scaling followed by steady improvement. The only differences are the model log compute at which scaling reverts from inverse scaling to standard scaling and also how fast performance goes up/down. However, Fig. A27b shows that MCS is also poorly correlated with the exact match. Even if we can precisely predict the MCS of models above the emergence threshold, conversion back to exact match accuracy will be too noisy to be useful. We hope this section illustrates potential avenues for future work.

H Broader Impact

H.1 Potential Positive Impacts

This work identifies U-shaped and inverted-U Scaling of LLM performance once we group questions by difficulty level. We believe this observation can provide the AI community with a deeper understanding of emergent abilities. We also present a forecasting pipeline utilizing the above observation to detect the forthcoming performance soar, the ability of which we believe is crucial in preventing deployment in offensive applications.

H.2 Potential Negative Impacts

Given the limitations discussed in Sec. 6, we do not suggest predicting the forthcoming emergent abilities based on merely one of the methods we discuss. Multiple techniques should be used in parallel to prevent possible false positives or false negatives.