Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer [1] **Brando Miranda** [1] **Sanmi Koyejo** [1]

Abstract

Recent work claims that large language models display *emergent abilities*, abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. We present an alternative explanation for emergent abilities: that for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three ways: we (1) make, test and confirm predictions on the effect of metric choice using the InstructGPT/GPT-3 family; (2) make, test and confirm predictions about metric choices in a meta-analysis on BIG-Bench; and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities on vision tasks. These analyses provide evidence that alleged emergent abilities disappear with different metrics or better statistics. Our work challenging a popular conception speaks to challenges with accurately evaluating generative AI models.

1. Introduction

Emergent properties of complex systems have long been studied across disciplines, from physics to biology to mathematics. The idea of emergence was popularized by Nobel Prize-winning physicist P.W. Anderson's "More Is Different" (Anderson, 1972), which argues as the complexity of a system increases, new properties may materialize that cannot be predicted even from a precise quantitative understanding of the system's microscopic details. Recently, the idea of emergence gained significant attention in machine learning due to observations that large language models (LLMs) such as GPT (Brown et al., 2020), PaLM (Chowdhery et al., 2022) and LaMDA (Thoppilan et al., 2022) exhibit so-called "emergent abilities" (Wei et al., 2022; Ganguli et al., 2022; Srivastava et al., 2022; Brown et al., 2020).

The term "emergent abilities of LLMs" was recently defined as "abilities that are not present in smaller-scale models but are present in large-scale models; thus they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models" (Wei et al., 2022). We call into question the claim that LLMs possess emergent abilities, by which we specifically mean *sharp* and *unpredictable* changes in model outputs as a function of model scale on specific tasks. Our doubt stems from the observation that emergent abilities seem to appear only under metrics that nonlinearly or discontinuously scale any model's per-token error rate. As we later show, $> 92\%$ of emergent abilities on BIG-Bench tasks (Srivastava et al., 2022) hand-annotated by Wei (2022) appear under one of these two metrics:

$$\text{Multiple Choice Grade} \overset{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Exact String Match} \overset{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$$

This raises the possibility of an alternative explanation for the origin of LLMs' emergent abilities: emergent abilities are a mirage caused primarily by the researcher choosing a metric that nonlinearly or discontinuously deforms per-token error rates, and secondarily by possessing too few test data to accurately estimate the performance of smaller models, thereby causing smaller models to appear wholly unable to perform the task.

To communicate our alternative explanation, we present it as a simple mathematical model and demonstrate how it

[1]Department of Computer Science, Stanford University. Correspondence to: Rylan Schaeffer <rschaef@cs.stanford.edu>, Sanmi Koyejo <sanmi@cs.stanford.edu>.

ICML 2023 Workshop: Challenges of Deploying Generative AI.

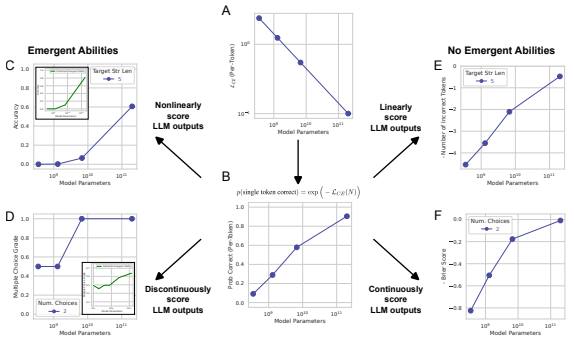

Figure 1. **Emergent abilities of large language models are created by the researcher's chosen metrics, not unpredictable changes in model behavior with scale.** (A) Suppose the per-token cross-entropy loss decreases monotonically with model scale, e.g., $\mathcal{L}_{CE}$ scales as a power law. (B) The per-token probability of selecting the correct token asymptotes towards 1. (C) If the researcher scores models' outputs using a nonlinear metric such as Accuracy (which requires a sequence of tokens to *all* be correct), the metric choice nonlinearly scales performance, causing performance to change sharply and unpredictably in a manner that qualitatively matches published emergent abilities (inset). (D) If the researcher instead scores models' outputs using a discontinuous metric such as Multiple Choice Grade (akin to a step function), the metric choice discontinuously scales performance, again causing performance to change sharply and unpredictably. (E) Changing from a nonlinear metric to a linear metric such as Token Edit Distance, scaling shows smooth, continuous and predictable improvements, ablating the emergent ability. (F) Changing from a discontinuous metric to a continuous metric such as Brier Score again reveals smooth, continuous and predictable improvements in task performance. Consequently, emergent abilities are created by the researcher's choice of metrics, not fundamental changes in model family behavior on specific tasks with scale. For a more complete exposition, see Schaeffer et al. (2023).

quantitatively reproduces the evidence offered in support of emergent abilities of LLMs. We then test our alternative explanation in three complementary ways: we (1) make, test and confirm three predictions based on our alternative hypotheses using the InstructGPT (Lowe & Leike, 2022) / GPT-3 (Brown et al., 2020) model family; (2) meta-analyze published benchmarks (Srivastava et al., 2022; Wei et al., 2022) to reveal that emergent abilities only appear for specific metrics, not for model families on particular tasks; (3) induce never-before-seen, seemingly emergent abilities in multiple architectures across various vision tasks by intentionally changing the metrics used for evaluation.

2. Alternative Explanation for Emergent Abilities

What might cause smooth, continuous, predictable changes in model family performance to appear sharp and unpredictable? The researcher's choice of a nonlinear or discontinuous metric can distort the model family's performance to appear sharp and unpredictable. To expound, suppose that within a model family, the test loss falls smoothly, continuously and predictably with the number of model parameters. One reason to believe this is the phenomenon known as neural scaling laws: empirical observations that networks exhibit power law scaling in the test loss as a function of training dataset size, number of parameters or compute (Hestness et al., 2017; Rosenfeld et al., 2019; Henighan et al., 2020; Kaplan et al., 2020; Gordon et al., 2021; Hernandez et al., 2021; Jones, 2021; Zhai et al., 2022; Hoffmann et al., 2022; Clark et al., 2022; Neumann & Gros, 2022). For concreteness, suppose we have a model family of different numbers of parameters $N > 0$ such that the per-token cross entropy falls as a power law with the number of parameters N for constants $c > 0, \alpha < 0$ (Fig. 1A):

$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha}$$

Note we do not require this particular functional form to hold; rather, we use it for illustrative purposes. Let V denote the set of possible tokens, $p \in \Delta^{|V|-1}$ denote the true but unknown probability distribution, and $\hat{p}_N \in \Delta^{|V|-1}$ denote the N-parameter model's predicted probability distribution. In practice, p is unknown, so we substitute a one-hot distribution of the observed token v^* to compute the per-token cross entropy as a function of number of parameters N:

$$\mathcal{L}_{CE}(N) = -\log \hat{p}_N(v^*)$$

A model with N parameters then has a per-token probability of selecting the correct token (Fig. 1B):

$$p(\text{single token correct}) = \exp\Big(-\mathcal{L}_{CE}(N)\Big)$$

Suppose the researcher then chooses a metric that requires selecting L tokens correctly. For example, our task might be L-digit integer addition, and a model's output is scored 1 if all L output digits exactly match all target digits with no additions, deletions or substitutions, 0 otherwise. If the probability each token is correct is independent[1], the probability of scoring 1 is:

$$\text{Accuracy}(N) \approx p_N(\text{single token correct})^{\text{num. of tokens}}$$

This choice of metric nonlinearly scales performance with increasing token sequence length. When plotting performance on a linear-log plot, one sees a sharp, unpredictable emergent ability on longer sequences (Fig. 1C) that closely matches claimed emergent abilities (inset). What happens if

[1]While the independence assumption is not true, the approximation yields results qualitatively matching the observed emergence claims.

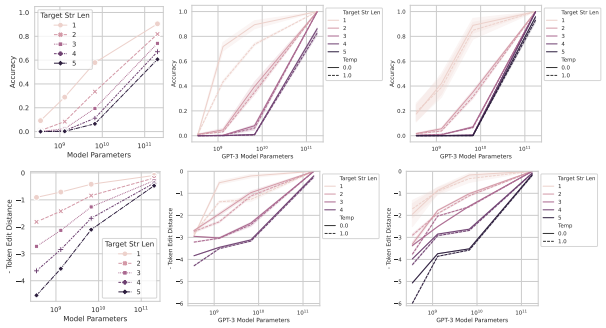

Figure 2. **Claimed emergent abilities evaporate upon changing the metric.** Left to Right: Mathematical Model, 2-Integer 2-Digit Multiplication Task, 2-Integer 4-Digit Addition Task. Top: When performance is measured by a nonlinear metric (e.g., Accuracy), the InstructGPT/GPT-3 family's performance appears sharp and unpredictable on longer target lengths. Bottom: When performance is instead measured by a linear metric (e.g., Token Edit Distance), the family exhibits smooth, predictable improvements.

the researcher switches from a nonlinear metric like Accuracy, under which the per-token error rate scales geometrically in target length (App. A.3), to an approximately linear metric like Token Edit Distance, under which the per-token error rate scales quasi-linearly in target length (App. A.2)?

$$\text{Token Edit Distance}(N) \approx L\left(1 - p_N(\text{single token correct})\right)$$

The linear metric reveals smooth, continuous, predictable changes in model performance (Fig. 1E). Similarly, if the researcher uses a discontinuous metric like Multiple Choice Grade, the researcher can find emergent abilities (Fig. 1D), but switching to a continuous metric like Brier Score removes the emergent ability (Fig. 1F). In summary, sharp and unpredictable changes with increasing scale can be fully explained by three interpretable factors: (1) the researcher choosing a metric that nonlinearly or discontinuously scales the per-token error rate, (2) having insufficient resolution to estimate model performance in the smaller parameter regime, with resolution set by $1/$test dataset size, and (3) insufficiently sampling the larger parameter regime.

3. Analyzing InstructGPT/GPT-3's Emergent Arithmetic Abilities

Previous papers prominently claimed the GPT (Brown et al., 2020; Lowe & Leike, 2022) family[2] displays emergent abilities at integer arithmetic tasks (Ganguli et al., 2022; Srivastava et al., 2022; Wei et al., 2022) (Fig. 1E). We chose these tasks as they were prominently presented, and we focused on the GPT family due to it being publicly queryable. To

[2]As of 2023-03-15, 4 models with 350M, 1.3B, 6.7B, 175B parameters are available via the OpenAI API.

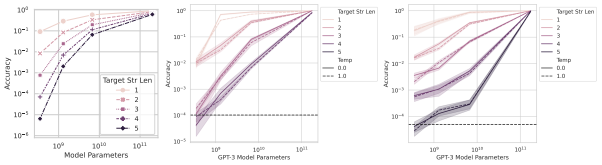

Figure 3. **Claimed emergent abilities evaporate upon using better statistics.** Left to Right: Mathematical Model, 2-Integer 2-Digit Multiplication Task, 2-Integer 4-Digit Addition Task. Generating additional test data to increase the resolution reveals that even on Accuracy, the InstructGPT/GPT-3 family's performance is above chance and improves in a smooth, continuous, predictable manner that qualitatively matches the mathematical model.

test predictions made by our mathematical model, we collected outputs from the InstructGPT/GPT-3 family on two tasks: 2-shot multiplication between two 2-digit integers and 2-shot addition between two 4-digit integers.

Prediction: Emergent Abilities Disappear With Different Metrics On both tasks, the GPT family displays emergent abilities if the target has 4 or 5 digits and if the metric is Accuracy (Fig. 2, top) (Brown et al., 2020; Ganguli et al., 2022; Wei et al., 2022). However, if one changes from nonlinear Accuracy to linear Token Edit Distance *while keeping the models' outputs fixed*, the family's performance smoothly, continuously and predictably improves with increasing scale (Fig. 2, bottom). This confirms our prediction and supports our alternative that the source of emergent abilities is the researcher's choice of metric, *not changes in the model family's outputs*. Increasing the length of the target string from 1 to 5 predictably decreases the family's performance in an approximately quasilinear manner.

Prediction: Emergent Abilities Disappear With Better Statistics Our second prediction is that even on nonlinear metrics such as Accuracy, smaller models do not have 0 accuracy, but rather have non-zero, above-chance accuracy *commensurate with choosing accuracy as the metric*. To properly measure models' accuracy, we increased the resolution by generating additional test data and found that on both arithmetic tasks, all models in the InstructGPT/GPT-3 family achieve above-chance accuracy (Fig. 3). This confirms our second prediction. Increasing the length of the target string from 1 to 5 predictably decreases the family's performance in an approximately geometric manner.

4. Meta-Analysis of Claimed Emergent Abilities

Analyzing the GPT family is possible because the models are publicly queryable. However, other model families claimed to exhibit emergent abilities are not publicly queryable, nor are their generated outputs publicly available,

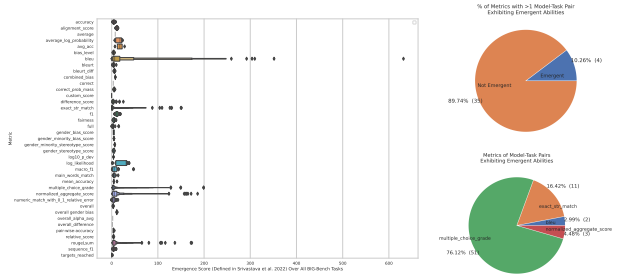

Figure 4. **Emergent abilities appear only for specific metrics, not task-model families.** (A) *Possible* emergent abilities appear with *at most* 5 out of 39 BIG-Bench metrics. (B) Hand-annotated data by Wei (2022) reveals emergent abilities appear only under 4 metrics. (C) $> 92\%$ of emergent abilities appear under one of two metrics: Multiple Choice Grade and Exact String Match.

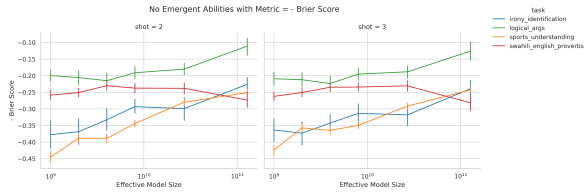

Figure 5. **Changing the metric when evaluating task-model family pairs causes emergent abilities to disappear.** LaMDA displays emergent abilities when measured under the discontinuous Multiple Choice Grade (not shown) that disappear when evaluated under a continuous BIG-Bench metric: Brier Score.

meaning we are limited to analyzing the published results themselves (Ganguli et al., 2022; Wei et al., 2022; Wei, 2022) contained in BIG-Bench (Srivastava et al., 2022).

Prediction: Emergent Abilities Should Appear with Metrics, not Task-Model Families If emergent abilities are real, one should expect task-model family pairs to show emergence for all reasonable metrics. However, if our alternative explanation is correct, we should expect emergent abilities to appear only under certain metrics. To test this, we analyzed on which metrics emergent abilities appear.

34 of 39 preferred metrics in BIG-Bench display no possible emergent abilities according to the emergence score introduced by Srivastava et al. (2022) (Fig. 4A). The remaining 5 are nonlinear or discontinuous, e.g., Exact String Match, Multiple Choice Grade, ROUGE-L-Sum (App. A.4). Because emergence score only *suggests* emergence, we also analyzed hand-annotated task-metric-model family triplets (Wei, 2022), and found emergent abilities appear with 4 metrics (Fig. 4B), with 2 metrics accounting for $> 92\%$ of claimed emergent abilities (Fig. 4C): Multiple Choice Grade (discontinuous) and Exact String Match (nonlinear).

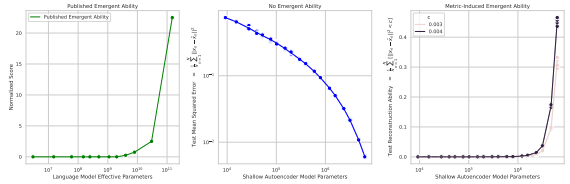

Figure 6. **Induced emergent reconstruction ability in shallow nonlinear autoencoders.** (A) A published emergent ability (Srivastava et al., 2022). (B) Shallow nonlinear autoencoders trained on CIFAR100 (Krizhevsky, 2009) display smoothly decreasing mean squared reconstruction error. (C) Using a newly defined Reconstruction$_c$ metric (Eqn. 1) induces an unpredictable change.

Prediction: Changing Metric Removes Emergent Abilities We focused on the LaMDA family (Thoppilan et al., 2022) because its outputs are available through BIG-Bench. For our analysis, we identified tasks on which LaMDA displays emergent abilities with Multiple Choice Grade, then asked whether LaMDA still displays emergent abilities on the same tasks with a different BIG-Bench metric: Brier Score (Brier et al., 1950). Brier Score is a strictly proper scoring rule for predictions of mutually exclusive outcomes. LaMDA's emergent abilities on the discontinuous Multiple Choice Grade disappeared when we changed the metric to the continuous Brier Score (Fig. 5).

5. Inducing Emergent Abilities in Networks on Vision Tasks

To demonstrate how emergent abilities can be induced by the researcher's choice of metric, we show how to produce emergent abilities in deep networks of various architectures: fully connected, convolutional, self-attentional. We focus on vision tasks because abrupt transitions in models' capabilities have not been observed to the best of our knowledge.

We first induce an emergent ability to reconstruct images in shallow nonlinear autoencoders trained on CIFAR100 (Krizhevsky, 2009). To emphasize that the sharpness of the metric is responsible for emergent abilities, we intentionally define a discontinuous metric that measures a network's ability to reconstruct a dataset as the average number of test data with squared reconstruction error below threshold c:

$$\text{Reconstruction}_c \overset{\text{def}}{=} \frac{1}{N} \sum_n \mathbb{I}\Big[||x_n - \hat{x}_n||^2 < c\Big] \qquad (1)$$

The autoencoder family displays smoothly decreasing squared reconstruction error as the number of units increases (Fig. 6B). Under our newly defined metric, the autoencoder family exhibits a sharp and seemingly unpredictable reconstruction ability (Fig. 6C) that qualitatively matches published emergent abilities (Fig. 6A). For the convolutional and transformer examples, see App. B.

References

Anderson, P. W. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A. K., Richemond, P. H., McClelland, J., and Hill, F. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, 2022.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Clark, A., De Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., et al. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pp. 4057–4086. PMLR, 2022.

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N., et al. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764, 2022.

Gordon, M. A., Duh, K., and Kaplan, J. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, 2021.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jones, A. L. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lowe, R. and Leike, J. Aligning language models to follow instructions. 2022. URL https://openai.com/research/instruction-following.

Neumann, O. and Gros, C. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage?, 2023.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wei, J. 137 emergent abilities of large language models. 2022. URL https://www.jasonwei.net/blog/emergence.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

A. Approximate Behavior of Metrics on Sequential Data

How do different metrics behave when used to measure autoregressive model outputs? Precisely answering this question is tricky and possibly analytically unsolvable, so we provide an approximate answer here.

Notationally, we consider N test data of length L (here, length is measured in tokens) with targets denoted $t_n \stackrel{\text{def}}{=} (t_{n1}, t_{n2}, ...t_{nL})$, the autoregressive model has a true-but-unknown per-token error probability of $\epsilon \in [0, 1]$ and the model outputs prediction $\hat{t}_n \stackrel{\text{def}}{=} (\hat{t}_{n1}, \hat{t}_{n2}, ...\hat{t}_{nL})$. This assumes that the model's per-token error probability is constant, which is empirically false, but modeling the complex dependencies of errors is beyond our scope.

A.1. Per-Token Error Probability is Resolution-Limited

Note that because we have N test data, each of length L, our resolution for viewing the per-token error probability ϵ is limited by $1/NL$. Here, resolution refers to "the smallest interval measurable by a scientific instrument; the resolving power." To explain what resolution means via an example, suppose one wants to measure a coin's probability of yielding heads. After a single coin flip, only two outcomes are possible (H, T), so the resolution-limited probability of heads is either 0 or 1. After two coin flips, four outcomes are possible (HH, HT, TH, TT), so the resolution-limited probability of heads is now one of $0, 0.5, 1$. After F coin flips, we can only resolve the coin's probability of yielding heads up to $1/F$. Consequently, we introduce a resolution-limited notation:

$$a_b \stackrel{\text{def}}{=} a \text{ rounded to the nearest integer multiple of } 1/b \tag{2}$$

A.2. Token Edit Distance

We first consider an adaptation of the Levenshtein (string edit) distance for models that function on tokens rather than characters, an adaptation we term the *token edit distance*. The token edit distance between two token sequences $t_n, \hat{t_n}$ is defined as the integer number of additions, deletions or substitutions necessary to transform t_n into $\hat{t}_n$ (or vice versa).

$$\text{Token Edit Distance}(t_n, \hat{t}_n) \stackrel{\text{def}}{=} \text{Num Substitutions} + \text{Num. Additions} + \text{Num. Deletions} \tag{3}$$

$$= \sum_{\ell=1}^{L} \mathbb{I}[t_{n\ell} \neq \hat{t}_{n\ell}] + \text{Num. Additions} + \text{Num. Deletions} \tag{4}$$

$$\geq \sum_{\ell=1}^{L} \mathbb{I}[t_{n\ell} \neq \hat{t}_{n\ell}] \tag{5}$$

The expected token edit distance is therefore:

$$\mathbb{E}[\text{Token Edit Distance}(t_n, \hat{t}_n)] \geq \mathbb{E}[\sum_{\ell=1}^{L} \mathbb{I}[t_{n\ell} \neq \hat{t}_{n\ell}]] \tag{6}$$

$$= \sum_{\ell=1}^{L} p(t_{n\ell} \neq \hat{t}_{n\ell}) \tag{7}$$

$$\approx L(1 - \epsilon) \tag{8}$$

The resolution-limited expected token edit distance is therefore:

$$\mathbb{E}[\text{Token Edit Distance}(t_n, \hat{t}_n)]_{NL} \geq L\Big(1 - \epsilon_{NL}\Big) \tag{9}$$

From this, we see that the expected token edit distance scales approximately linearly with the resolution-limited per-token probability. The real rate is slightly higher than linear because additions and deletions contribute an additional non-negative cost, but modeling this requires a model of how likely the model is to overproduce or underproduce tokens, which is something we do not currently possess.

A.3. Accuracy

$$\text{Accuracy}(t_n, \hat{t}_n) \overset{\text{def}}{=} \mathbb{I}[\text{No additions}]\,\mathbb{I}[\text{No deletions}] \prod_{l=1}^{L} \mathbb{I}[t_{nl} = \hat{t}_{nl}] \tag{10}$$

$$\approx \prod_{l=1}^{L} \mathbb{I}[t_{nl} = \hat{t}_{nl}] \tag{11}$$

As with the Token Edit Distance (App. A.3), we ignore how likely the language model is to overproduce or underproduce tokens because we do not have a good model of this process. Continuing along,

$$\mathbb{E}[\log \text{Accuracy}] = \sum_{l} \mathbb{E}[\log \mathbb{I}[t_{nl} = \hat{t}_{nl}]] \tag{12}$$

$$\leq \sum_{l} \log \mathbb{E}[\mathbb{I}[t_{nl} = \hat{t}_{nl}]] \tag{13}$$

$$\approx L \log(1 - \epsilon) \tag{14}$$

Taking an approximation that would make most mathematicians cry:

$$\mathbb{E}[\text{Accuracy}] \approx \exp(\mathbb{E}[\log \text{Accuracy}]) \tag{15}$$

$$= (1 - \epsilon)^L \tag{16}$$

$$\tag{17}$$

This reveals that accuracy **approximately** falls geometrically with target token length. The resolution-limited expected accuracy is therefore:

$$\mathbb{E}[\text{Accuracy}]_{NL} = (1 - \epsilon)^L{}_{NL} \tag{18}$$

From this we can see that choosing a nonlinear metric like Accuracy is affected significantly more by limited resolution because Accuracy forces one to distinguish quantities that decay rapidly.

A.4. ROUGE-L-Sum

Another BIG-Bench metric (Srivastava et al., 2022) is ROUGE-L-Sum (Lin, 2004), a metric based on the longest common subsequence (LCS) between two sequences. Section 3.2 of (Lin, 2004) gives the exact definition, but the key property is that ROUGE-L-Sum measures the "union" LCS, which means "stitching" together LCSs across the candidate and multiple references. As explained in the original paper: if the candidate sequence is $c = w_1 w_2 w_3 w_4 w_5$, and if there are two reference sequences $r_1 = w_1 w_2 w_6 w_7 w_8$ and $r_2 = w_1 w_3 w_8 w_9 w_5$, then $LCS(r_1, c) = w_1 w_2$ and $LCS(r_2, c) = w_1 w_3 w_5$, then the *union* -LCS of c, r_1, r_2 is $w_1 w_2 w_3 w_5$, with length 4. Intuitively, this disproportionately benefits models with smaller error rates because their mistakes can be "stitched" across multiple references; this is confirmed in simulation (Fig. 7).

B. Inducing Emergent Abilities in Networks on Vision Tasks

B.1. Emergent Classification of MNIST Handwritten Digits by Convolutional Networks

We begin by inducing an emergent classification ability in a LeNet convolutional neural network family (LeCun et al., 1998), trained on the MNIST handwritten digits dataset (LeCun, 1998). This family displays smoothly increasing test accuracy as the number of parameters increase (Fig. 8B). To emulate the accuracy metric used by emergence papers (Ganguli et al., 2022;

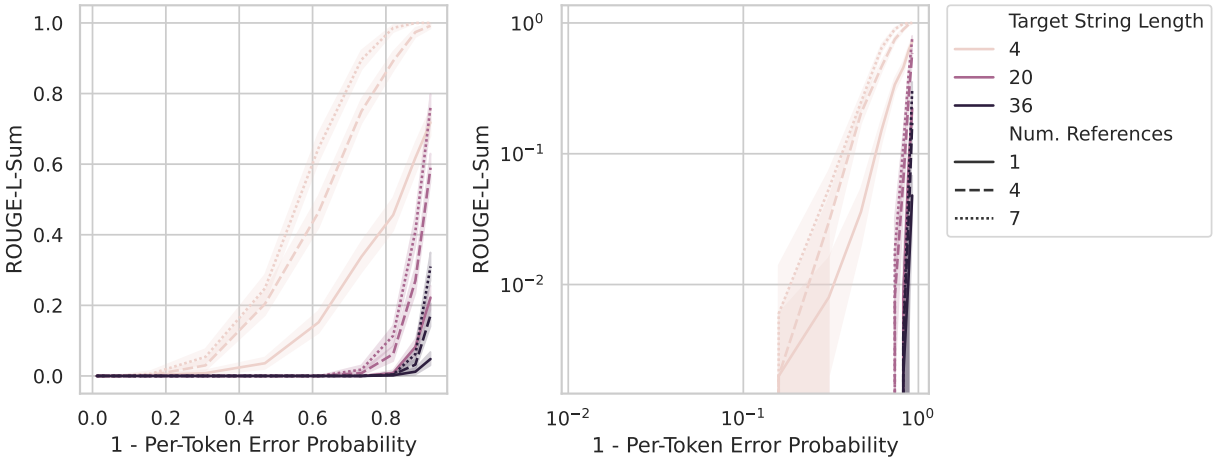

Figure 7. **ROUGE-L-Sum is a sharp metric.** Simulations show that as the per-token error probability slightly increase (e.g. from 0.05 to 0.1), the ROUGE-L-Sum metric sharply falls.

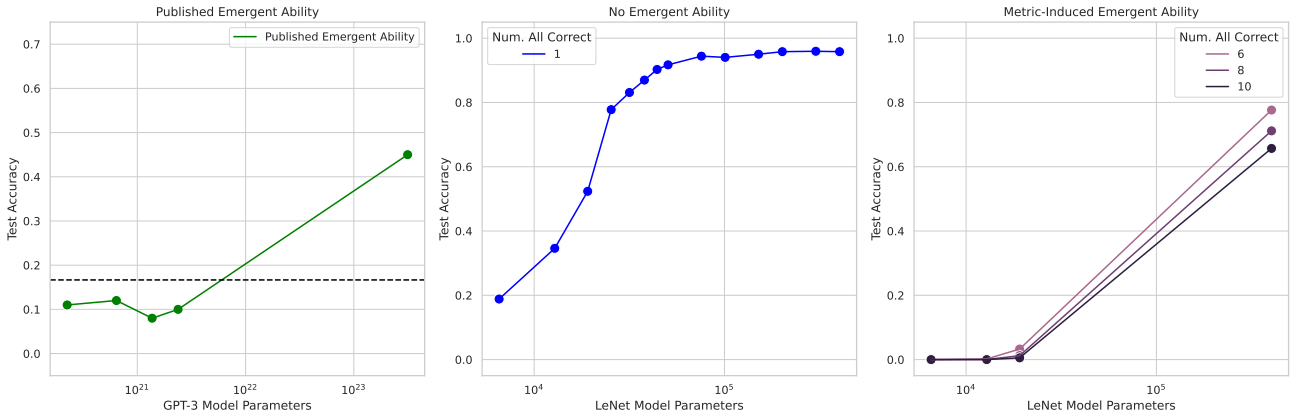

Figure 8. **Induced emergent MNIST classification ability in convolutional networks.** (A) A published emergent ability from the BIG-Bench Grounded Mappings task (Wei et al., 2022). (B) LeNet trained on MNIST (LeCun, 1998) displays a predictable, commonplace sigmoidal increase in test accuracy as model parameters increase. (C) When accuracy is redefined as correctly classifying K out of K independent test data, this newly defined metric induces a seemingly unpredictable change.

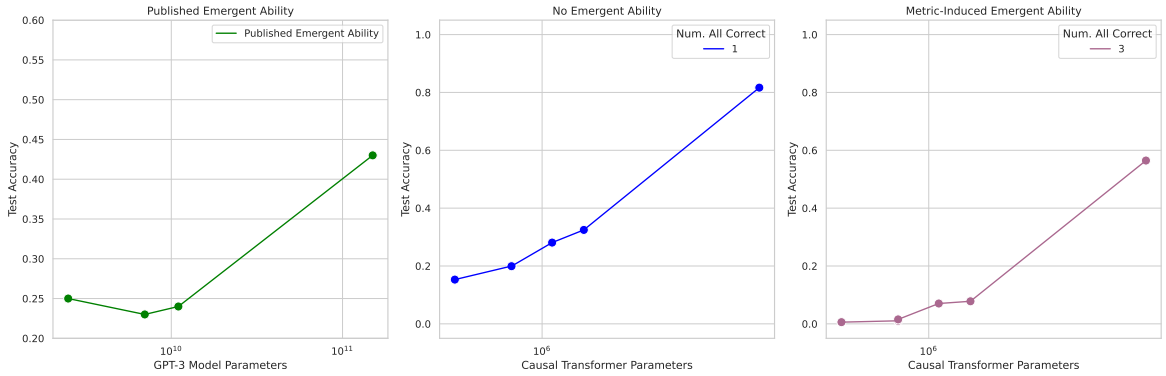

Figure 9. **Induced emergent classification ability in autoregressive Transformers.** (A) A published emergent ability on the MMLU benchmark (Ganguli et al., 2022). (B) Autoregressive transformers trained to classify Omniglot images display increasing accuracy with increasing scale. (C) When accuracy is redefined as classifying *all* images correctly, a seemingly emergent ability appears.

Wei et al., 2022; Srivastava et al., 2022), we use *subset accuracy*: 1 if the network classifies K out of K (independent) test data correctly, 0 otherwise. Under this definition of accuracy, the model family displays an "emergent" ability to correctly classify sets of MNIST digits as K increases from 1 to 5, especially when combined with sparse sampling of model sizes (Fig. 8C). This convolutional family's emergent classification ability qualitatively matches published emergent abilities, e.g., at the BIG-Bench Grounded Mappings task (Wei et al., 2022) (Fig. 8A).

B.2. Emergent Classification of Omniglot Characters by Autoregressive Transformers

We next induce emergent abilities in Transformers (Vaswani et al., 2017) trained to autoregressively classify Omniglot handwritten characters (Lake et al., 2015), in a setup inspired by recent work (Chan et al., 2022): Omniglot images are embedded by convolutional layers, then sequences of embedded image-image class label pairs are fed into decoder-only transformers. We measure image classification performance on sequences of length $L \in [1, 5]$, again via *subset accuracy*: 1 if all L images are classified correctly (Fig. 9B), 0 otherwise. Causal transformers display a seemingly emergent ability to correctly classify Omniglot handwritten characters (Fig. 9C) that qualitatively matches published emergent abilities (Fig. 9A).