Are Emergent Abilities of Large Language Models a Mirage?

NeurIPS 2023 Outstanding Main Track Paper

Schaeffer et. al.
Stanford University

Presenter: Hawon Jeong

Introduction

Emergent Abilities

- "Abilities that are not present in smaller-scale models but are present in large-scale models"
- Emergent abilities observed in LLMs such as GPT, PaLM, and LaMDA
- Two defining properties of emergent abilities
  - Sharpness
  - Unpredictability
- Research questions
  - What controls *which* abilities will emerge?
  - What controls *when* abilities will emerge?
  - How can we make desirable abilities emerge *faster*?

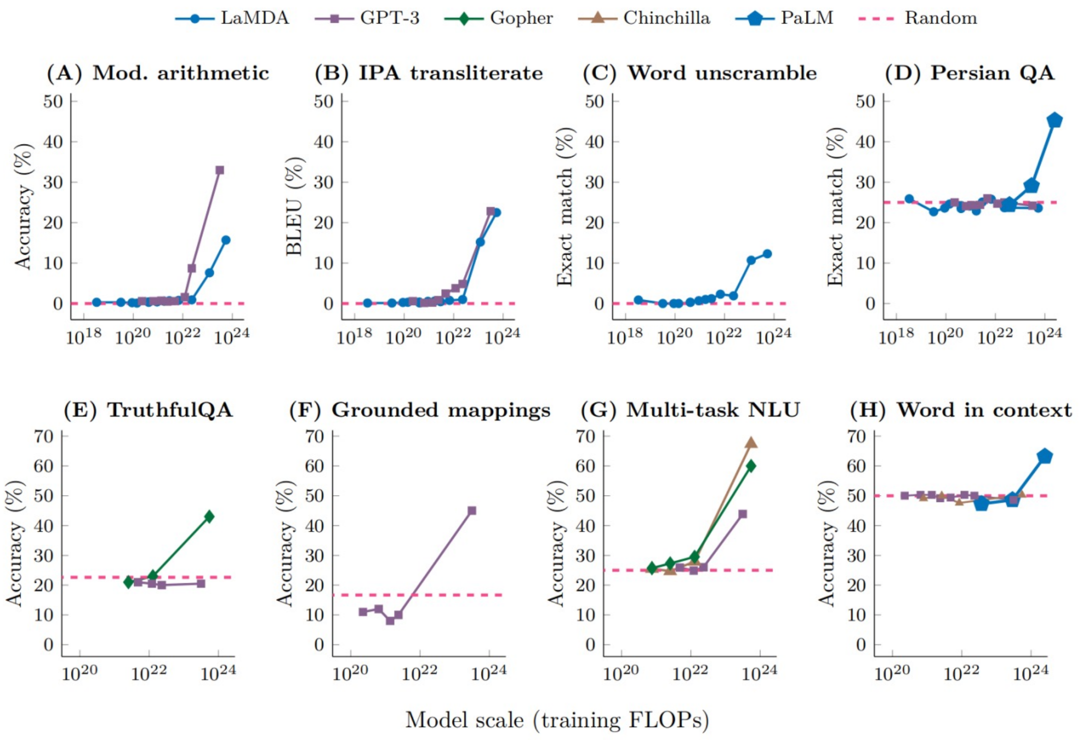


Figure 1: **Emergent abilities of large language models**. Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [33].

Introduction

Observation

- Many emergent abilities seem to appear only under metrics that nonlinearly or discontinuously scale the model's per-token error rate
- For instance, >92% of emergent abilities on BIG-Bench occur for 1 of the following metrics:

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

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

- "Emergent" abilities might not be due to fundamental changes in models with scale, but due to **researchers' choice of metrics**

Alternative Explanation for Emergent Abilities

—— Toy model

1. Suppose that test loss falls with model scale (e.g., params, data, compute)

2. Assume that each model's per-token cross entropy falls as a power law with the number of parameters N

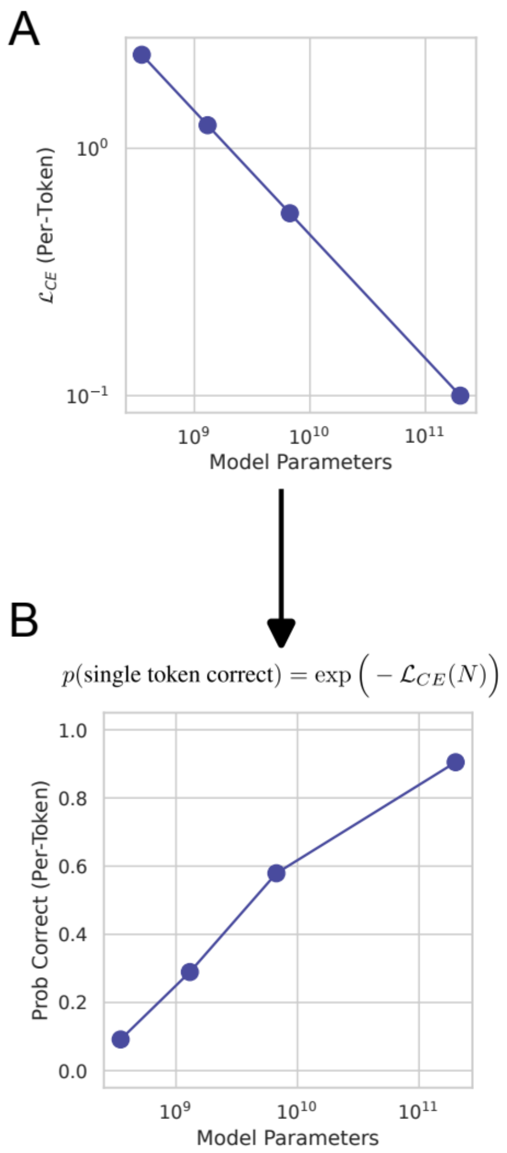
$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha} \overset{\text{def}}{=} -\sum_{v \in V} p(v) \log \hat{p}_N(v)$$

Substitute a one-hot distribution of observed token

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

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

$$\text{Accuracy}(N) \approx p_N(\text{single token correct})^{\text{num. of tokens}} = \exp\Big(-(N/c)^{\alpha}\Big)^{L}$$



Alternative Explanation for Emergent Abilities

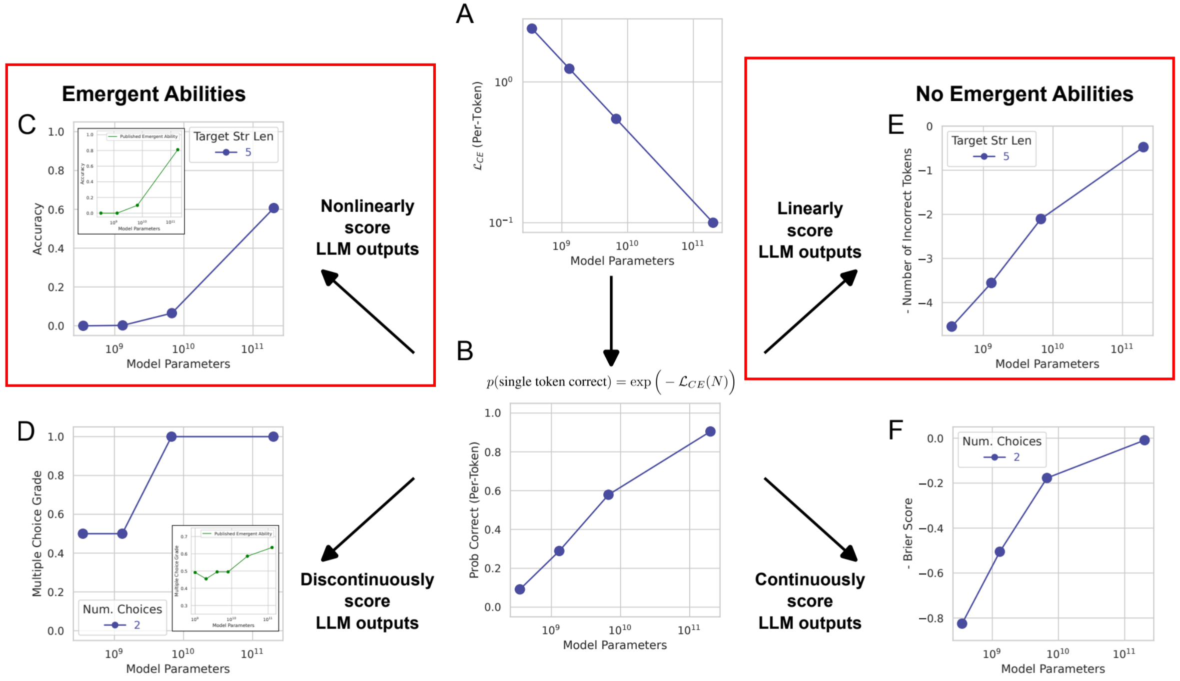
Toy model

3. Choose metrics: nonlinear v.s. linear

- Task: 2-integer k-digit addition

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

$$\begin{aligned} \text{Token Edit Distance}(N) & \\ &\approx L\Big(1 - p_N(\text{single token correct})\Big) \\ &= L\Big(1 - \exp\big(-(N/c)^{\alpha}\big)\Big) \end{aligned}$$



Alternative Explanation for Emergent Abilities

—— Toy model

3. Choose metrics: nonlinear v.s. linear

- **Token Edit Distance**

$$
\begin{aligned}
\text{Token Edit Distance}(t_n, \hat{t}_n) &\stackrel{\text{def}}{=} \text{Num Substitutions} + \text{Num. Additions} + \text{Num. Deletions} \\
&= \sum_{\ell=1}^{L} \mathbb{I}[t_{n\ell} \neq \hat{t}_{n\ell}] + \text{Num. Additions} + \text{Num. Deletions} \\
&\geq \sum_{\ell=1}^{L} \mathbb{I}[t_{n\ell} \neq \hat{t}_{n\ell}]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\text{Token Edit Distance}(t_n, \hat{t}_n)] &\geq \mathbb{E}[\sum_{\ell=1}^{L} \mathbb{I}[t_{n\ell} \neq \hat{t}_{n\ell}]] \\
&= \sum_{\ell=1}^{L} p(t_{n\ell} \neq \hat{t}_{n\ell}) \\
&\approx L(1-\epsilon)
\end{aligned}
$$

Alternative Explanation for Emergent Abilities

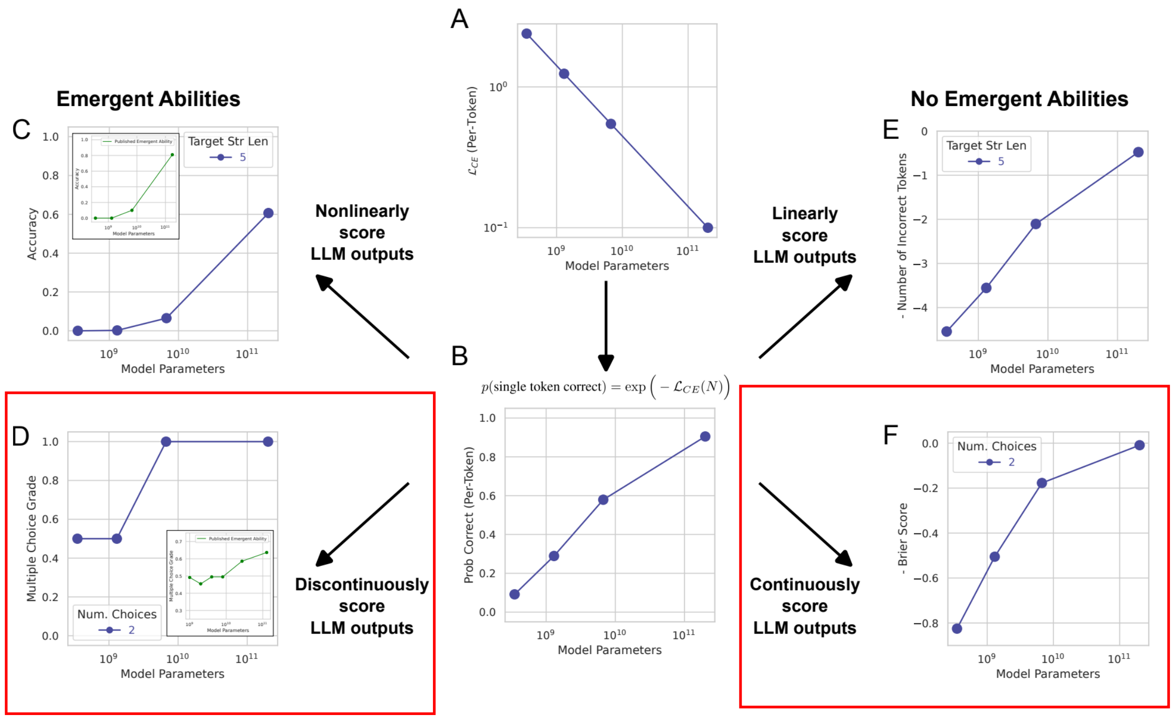
Toy model

3. Choose metrics: discontinuous v.s. continuous

- Task: Choosing 1 of 2 multiple choice options

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

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$



Alternative Explanation for Emergent Abilities

3 factors of emergent abilities

1) Chosen metric that nonlinearly or discontinuously scales the per-token error rate
2) Having insufficient test data to estimate the performance of smaller models
3) Insufficiently sampling the larger parameter regime

➡️

1. Test on InstructGPT / GPT-3 model family (350M, 1.6B, 6.7B, 175B)
2. Meta-analyzing published benchmarks (Google BIG-Bench)
3. Inducing seemingly emergent abilities in networks on vision tasks

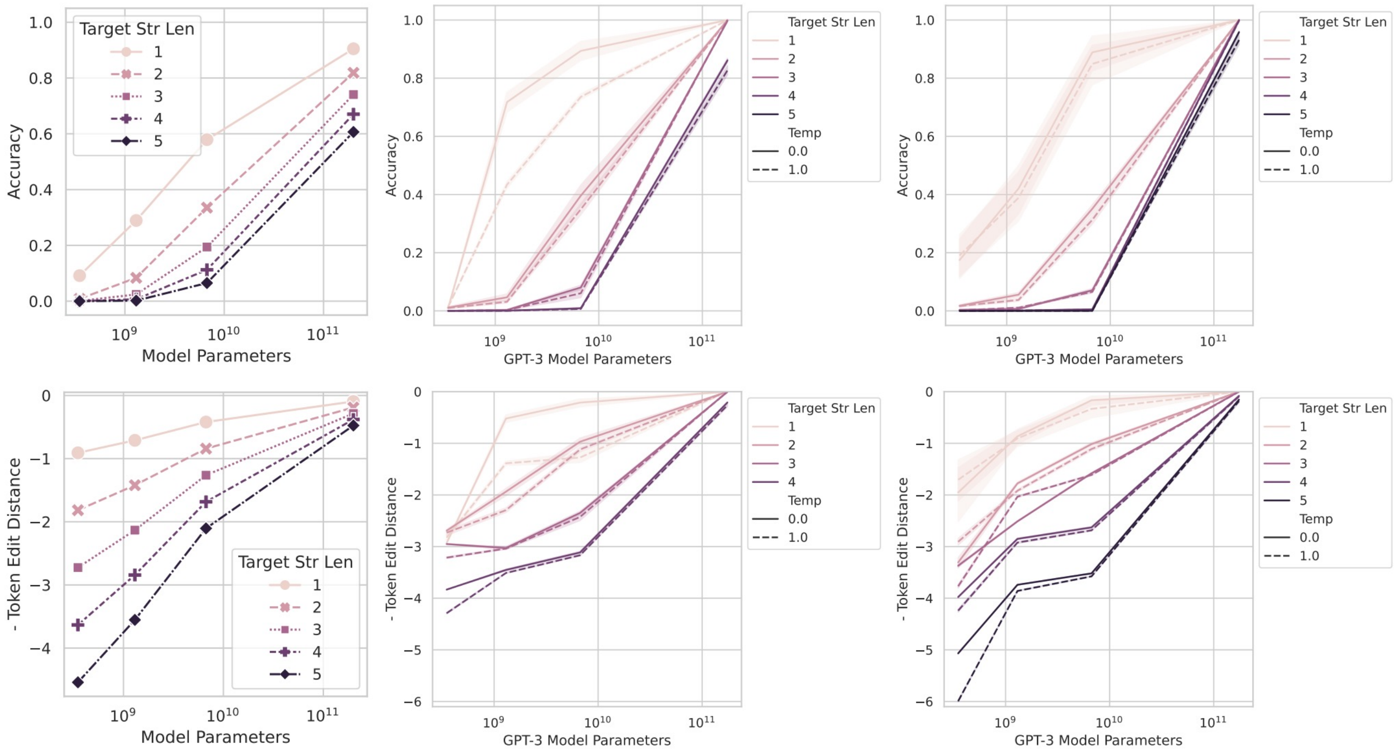
Experiment 1

Analyzing InstructGPT/GPT-3's Emergent Arithmetic Abilities

- **Prediction: Emergent abilities disappear with different metrics**

2-digit integer multiplication

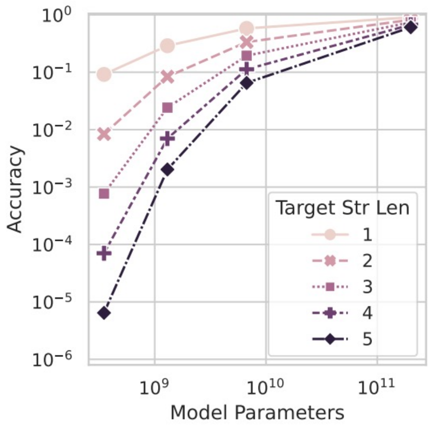
4-digit integer addition

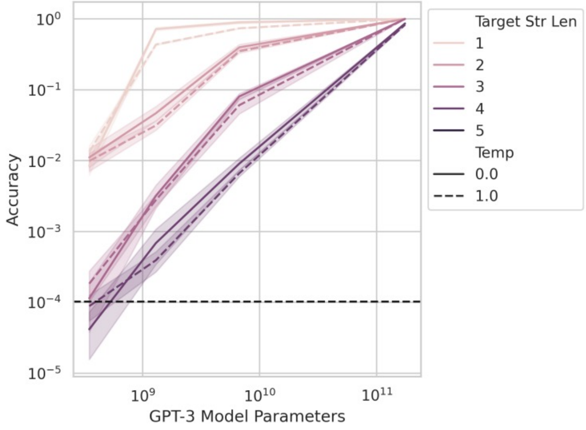


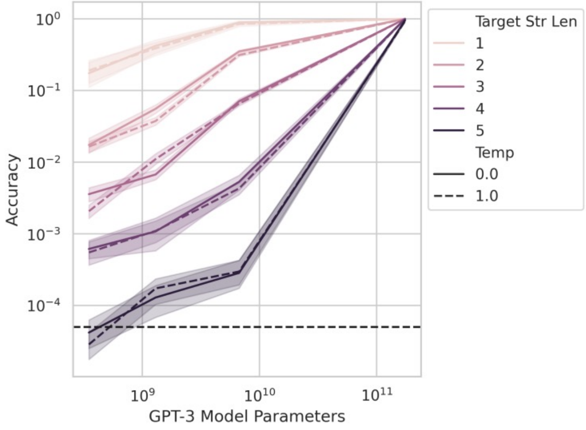
Experiment 1

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

- **Prediction: Emergent abilities disappear with better statistics**
  - Generating additional test data







Smaller models do not have zero accuracy

Experiment 2

Meta-Analysis of Claimed Emergent Abilities

- **Prediction: Emergent abilities should appear with specific *Metrics*, not *Task-Model Families***
  - In BIG-Bench paper, many LMs(GPT-3, Chinchilla, PaLM, LaMDA, ⋯) display emergent abilities.

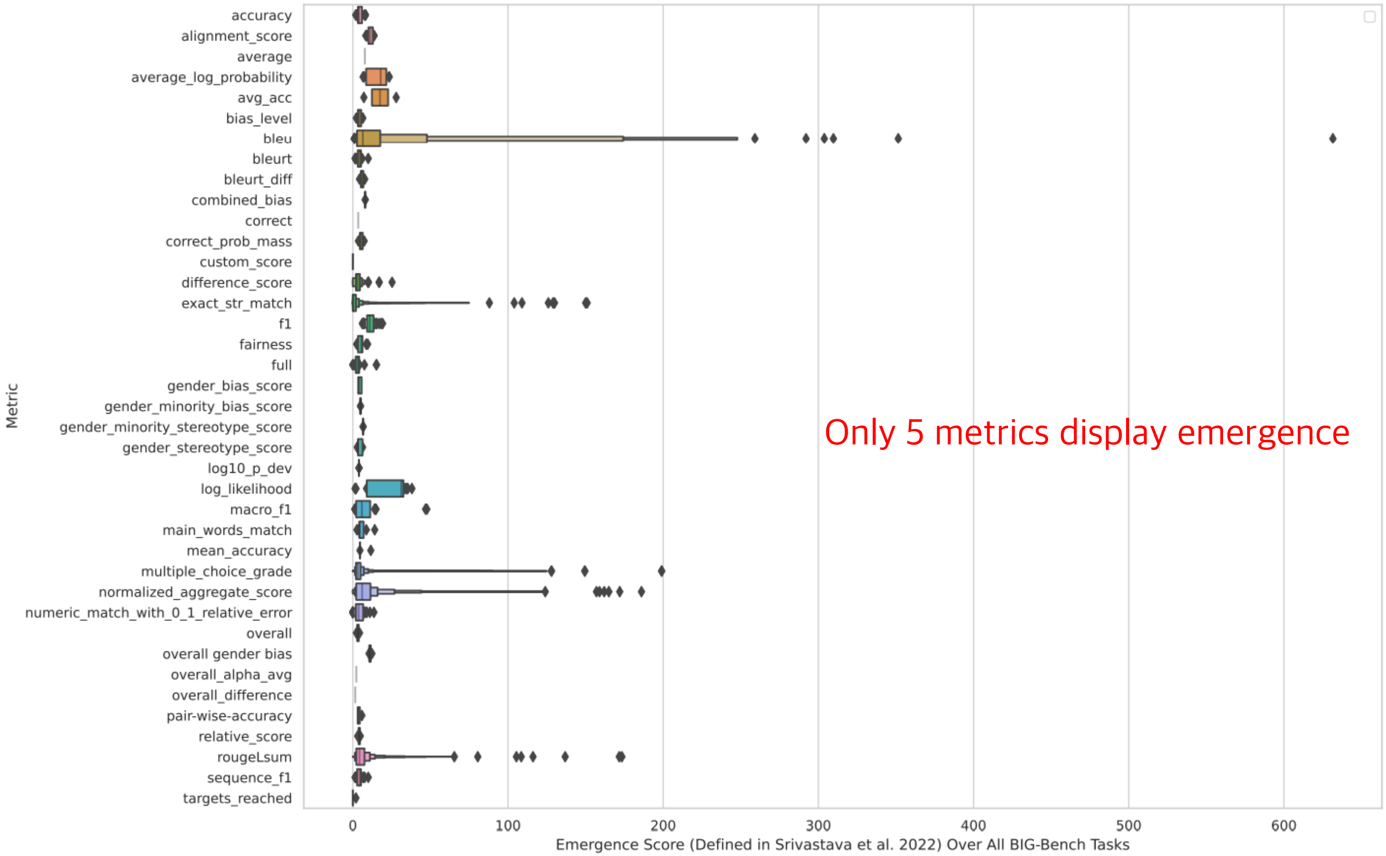
  - Consider 1 family
  - x_i: model scales, sorted s.t. $x_i < x_{i+1}$
  - y_i: model performance (on specific task-metric)

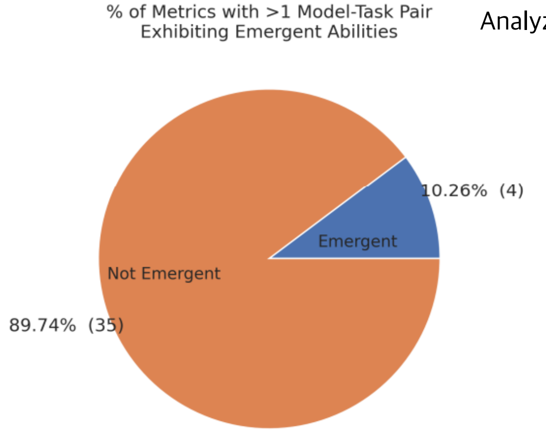
$$\text{Emergence Score}\left(\left\{(x_n, y_n)\right\}_{n=1}^{N}\right) \stackrel{\text{def}}{=} \frac{\operatorname{sign}(\arg\max_i y_i - \arg\min_i y_i)(\max_i y_i - \min_i y_i)}{\sqrt{\operatorname{Median}(\{(y_i - y_{i-1})^2\}_i)}}$$

Experiment 2

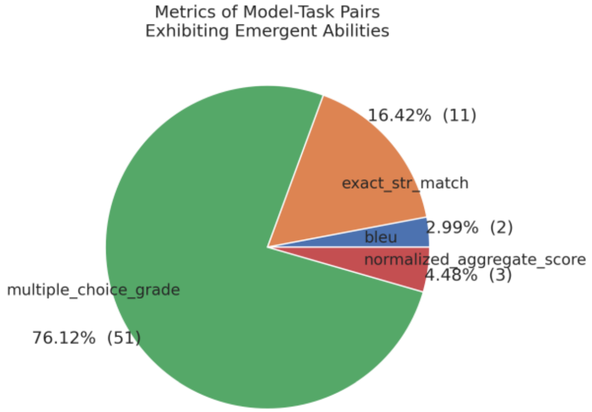
Meta-Analysis of Claimed Emergent Abilities

- **Prediction: Emergent abilities should appear with specific *Metrics*, not *Task-Model Families***





Analyze with [1]



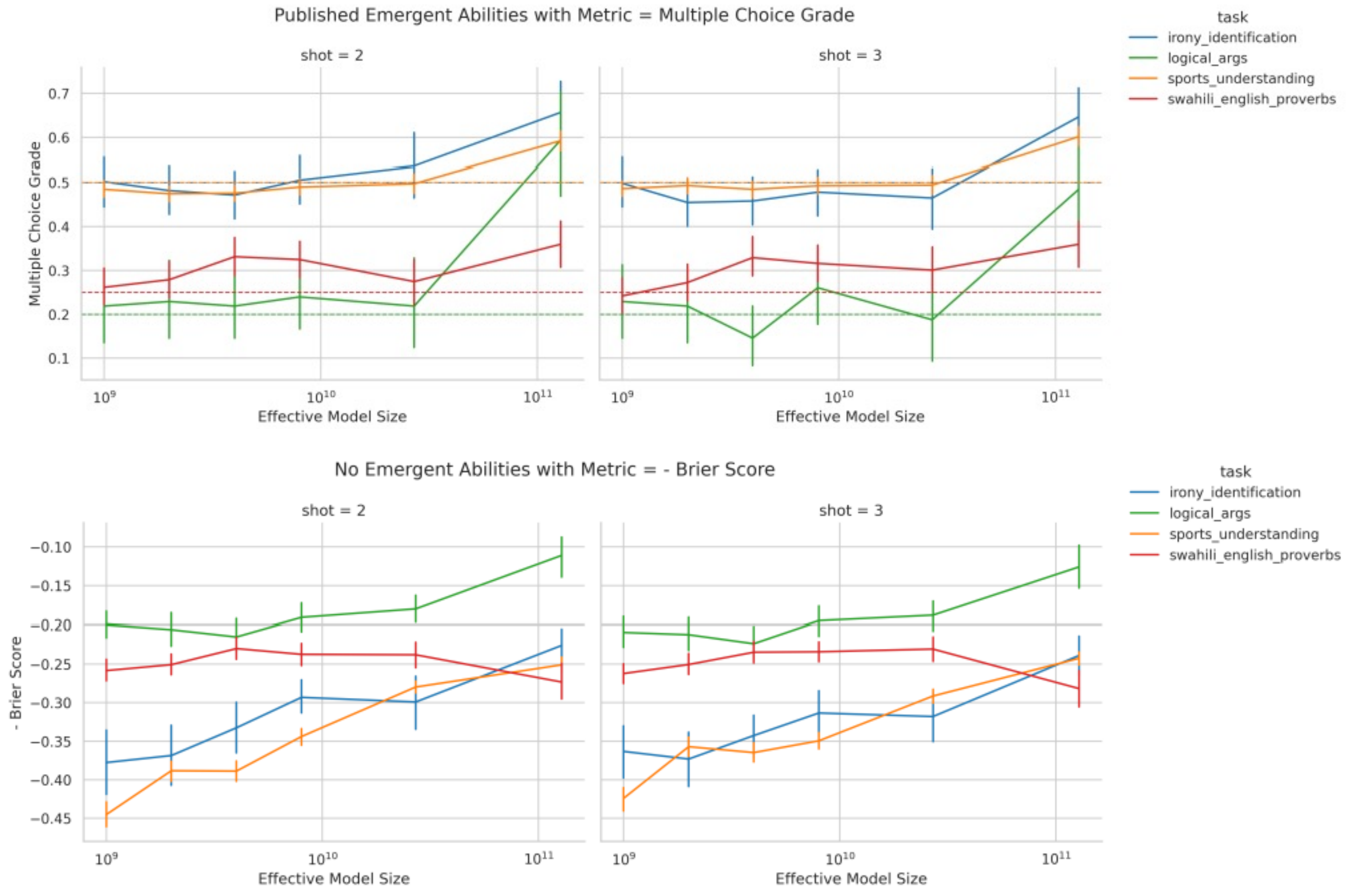
[1] Jason Wei. 137 emergent abilities of large language models. 2022.

Experiment 2

Meta-Analysis of Claimed Emergent Abilities

- **Prediction: Changing Metric Removes Emergent Abilities**
  - LaMDA family (available in BIG-Bench)
  - Continuous BIG-Bench metric: Brier Score

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$



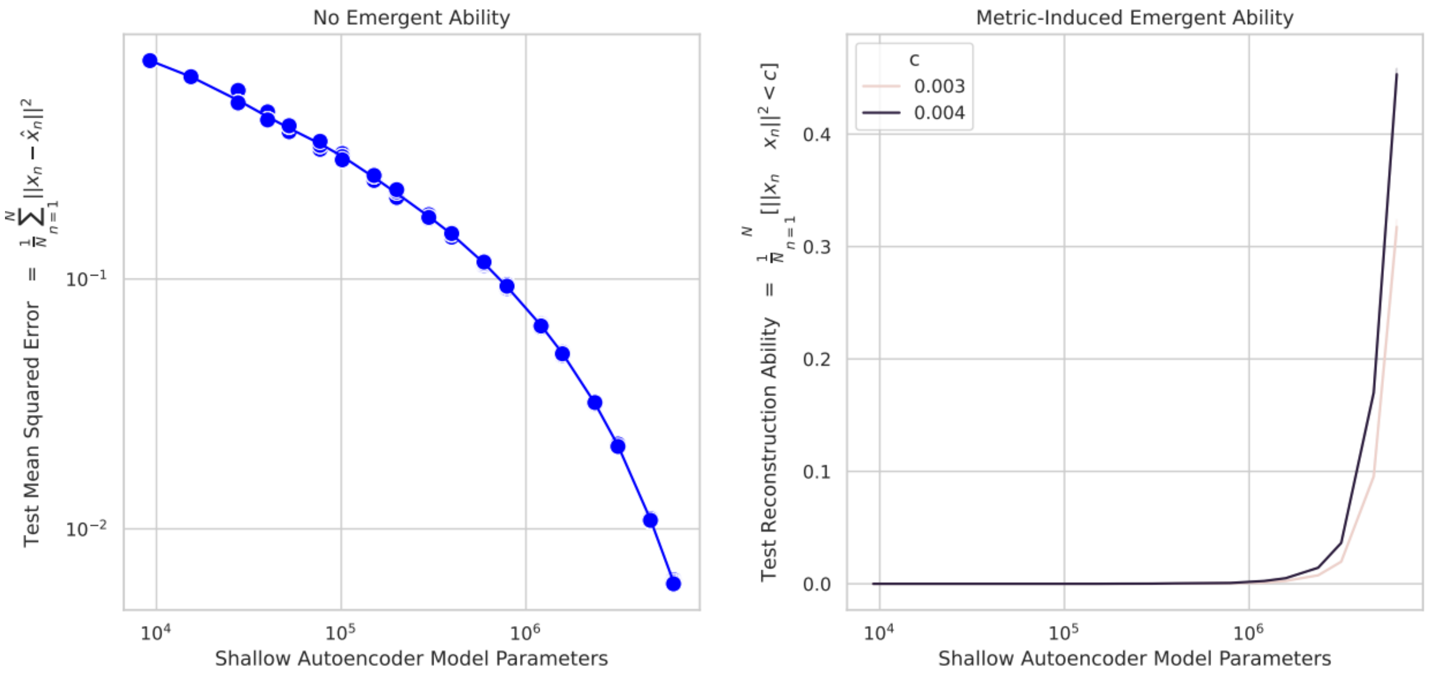
Experiment 3

Inducing Emergent Abilities in Networks on Vision Tasks

- Abrupt transitions in vision models' capabilities have not been observed.
- Producing emergent abilities in various architectures: fc, convolutional, self-attentional
- **Emergent reconstruction on CIFAR100**

Intentionally define a **discontinuous** metric:

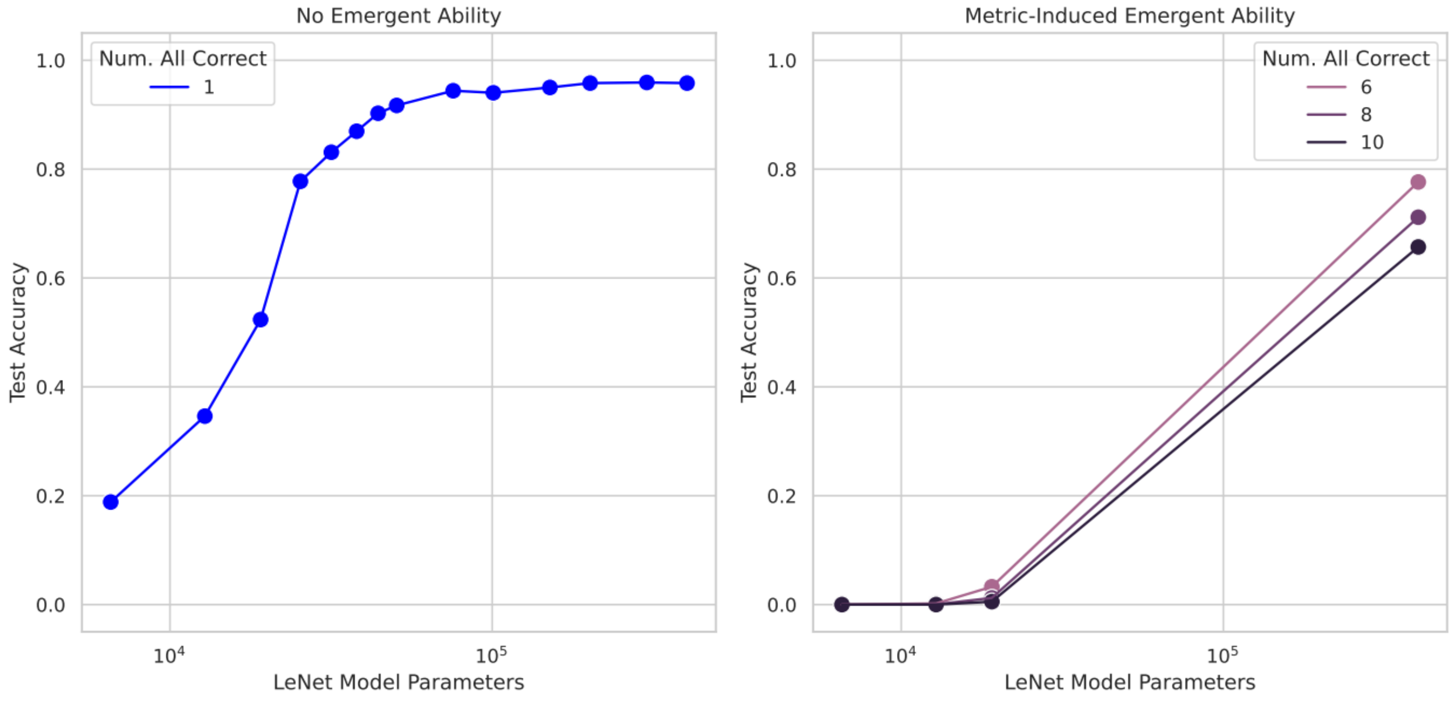
$$\text{Reconstruction}_c\left(\{x_n\}_{n=1}^N\right) \stackrel{\text{def}}{=} \frac{1}{N}\sum_n \mathbb{I}\left[||x_n - \hat{x}_n||^2 < c\right]$$



Experiment 3

Inducing Emergent Abilities in Networks on Vision Tasks

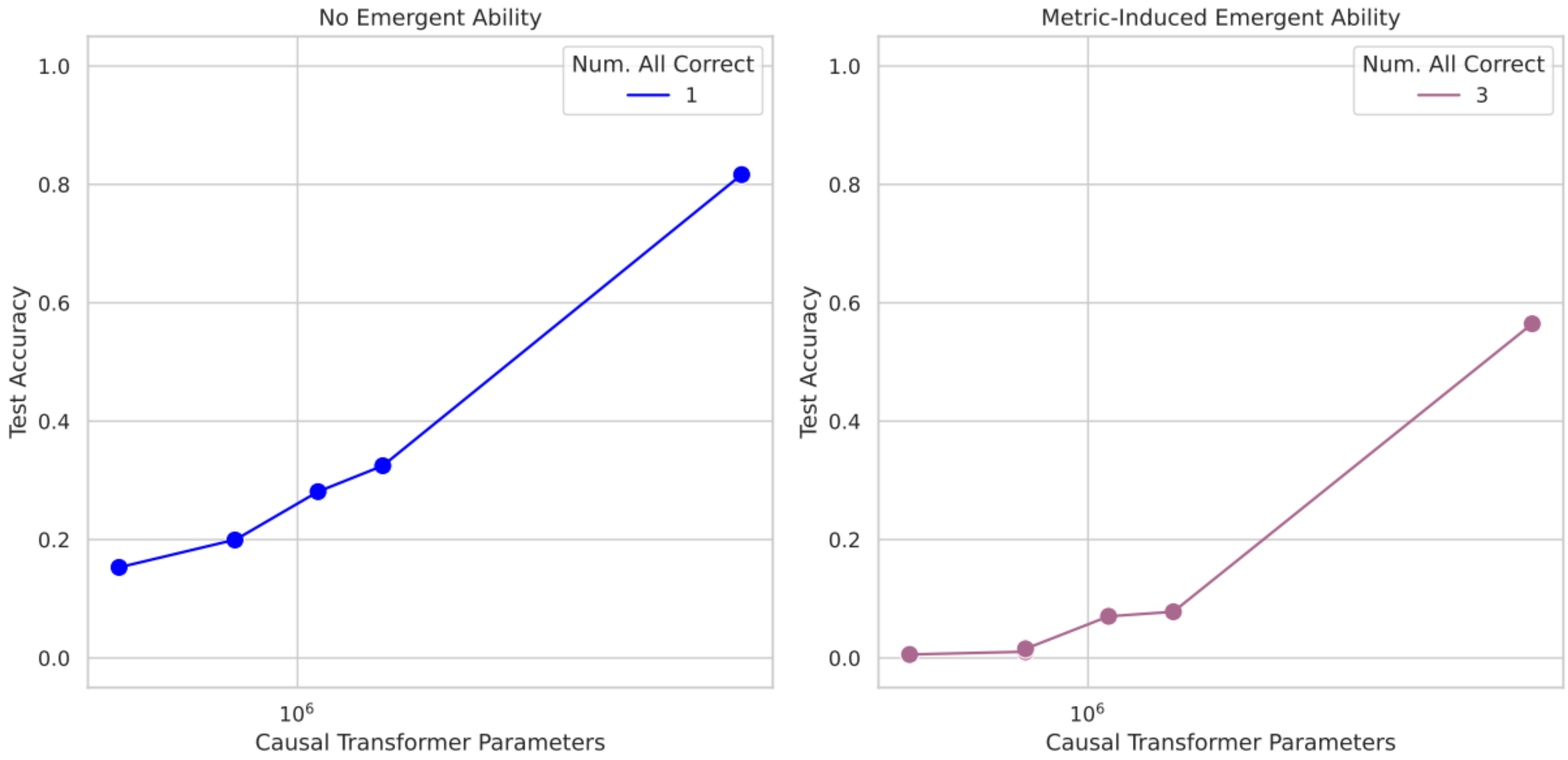
- **Emergent classification of MNIST**



Experiment 3

Inducing Emergent Abilities in Networks on Vision Tasks

- **Emergent classification of Omniglot charactors**



Discussion & Conclusion

- Emergent abilities may be creations of the researcher's choices, not a fundamental property of the model family on the specific task.
- A task and a metric are distinct and meaningful choices when constructing benchmarks.
- When choosing metrics, one should consider the metric's effect on the per-token error rate.
- When making claims about capabilities of large models, including proper control is critical.
- Scientific progress can be hampered when models and their outputs are not made public for independent scientific investigation.