Are Emergent Abilities of Large Language Models a Mirage?

Authors: Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Outstanding main paper at Neurips 2023

Presented for MLBBQ by Riyasat Ohib

This slide deck is based on the presentation by the original author of the paper Brando Miranda available at:
Are Emergent Abilities of Large Language Models a Mirage? - IEEE @ Stanford Unviersty (youtube.com)

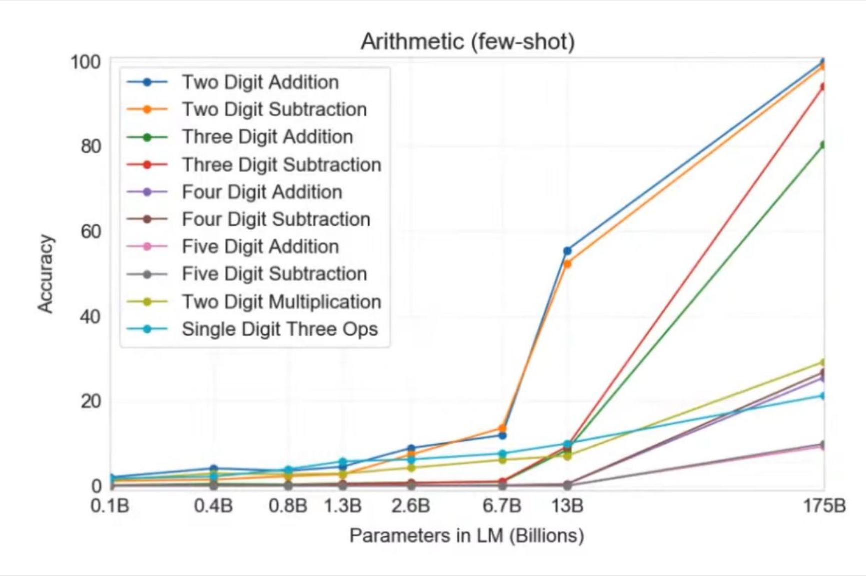
Main Takeaways

Some previous work argues that the improvements of LLMs capabilities are "emergent" at scale.

- The paper helps clarify & articulate clearly their definition of "emergent" as
  - Unpredictable
  - Sharp Jump
- They show **Emergent Abilities due to Scale** aren't what they seem
- They argue that the claimed "emergent capabilities" at scale are predictable and not always sharp
  - They show that the sharpness is a **fundamental property of their specific metric** (e.g. exact match) **not a fundamental property of the model.**
  - The sharpness or jump in performance is **predictable and expected**
  - In addition, they show that by using alternative "soft metrics" (e.g. edit distance) one can even remove sharpness.
- The paper is not making statements about "emergent" capabilities that the model was not explicitly trained for.
- The paper is not telling how to score the models, but enabling us to be aware of the trade-offs and helps us not to be surprised,

This slide deck is based on the presentation by the original author of the paper Brando Miranda available at:
Are Emergent Abilities of Large Language Models a Mirage? - IEEE @ Stanford Unviersty (youtube.com)

Background: What is Emergence?

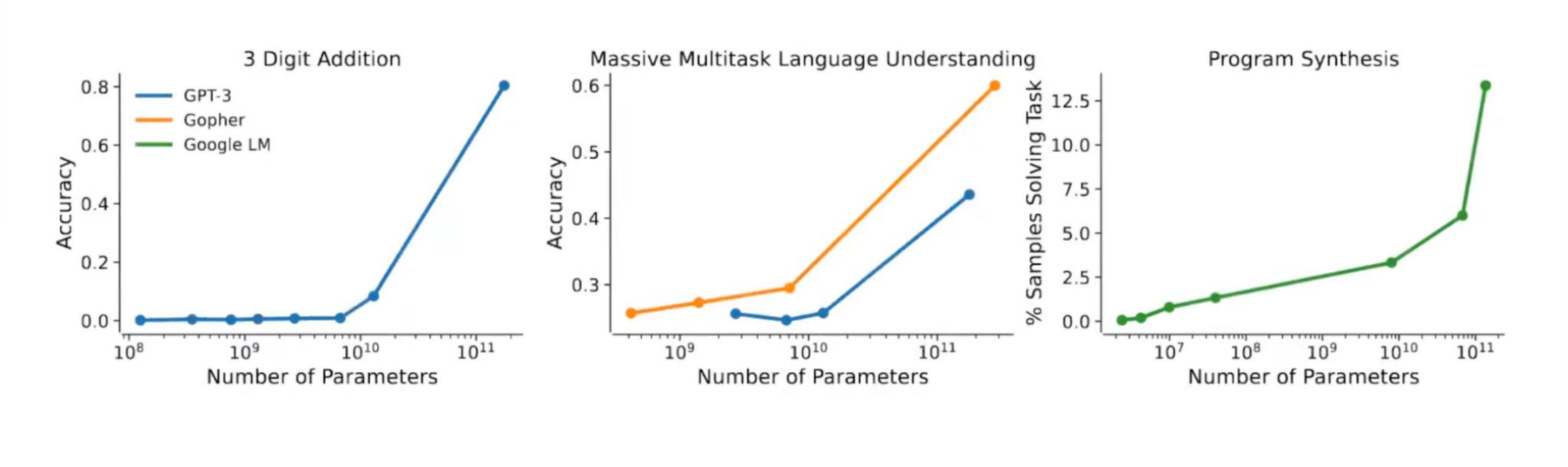


The paper clarifies the definition as:

- “Emergence at scale”
- Not the “emergence of untrained abilities”.
- The reason behind the plots looking this way?

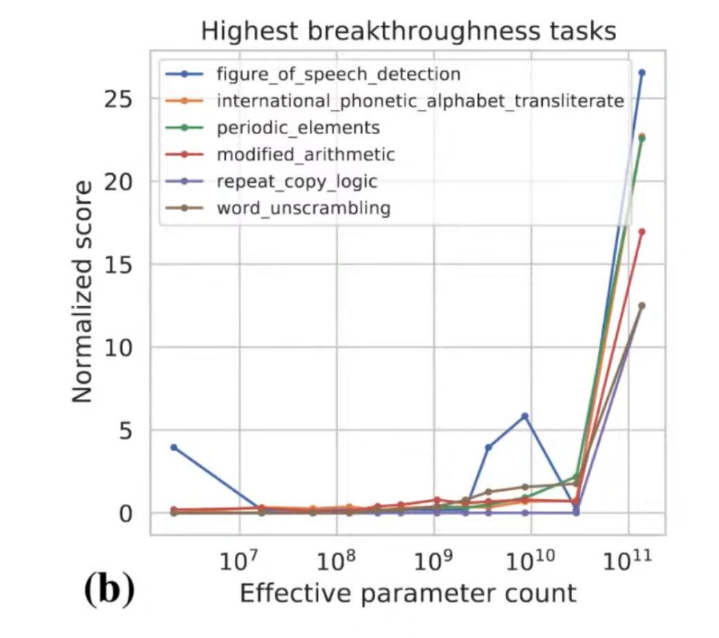
[2005.14165] Language Models are Few-Shot Learners (arxiv.org)

Background: Emergent Abilities in LLM, Some Examples



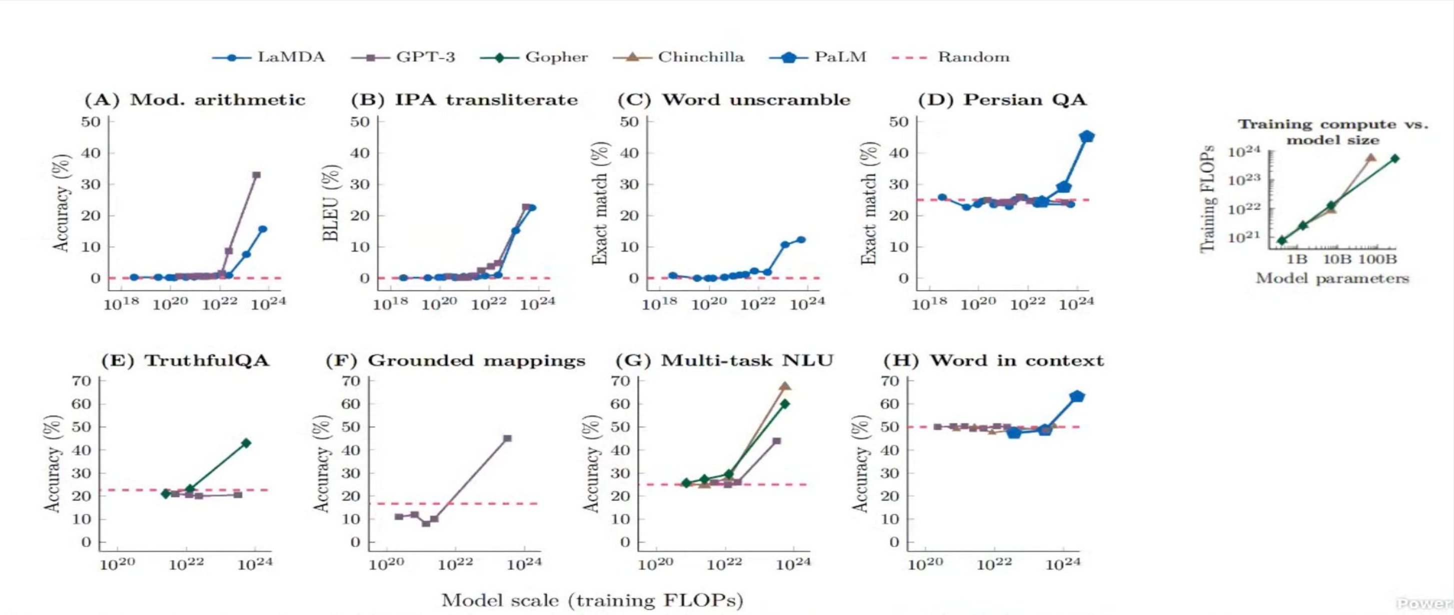
[2202.07785] Predictability and Surprise in Large Generative Models (arxiv.org)

Background: Emergent Abilities in LLM, Some Examples



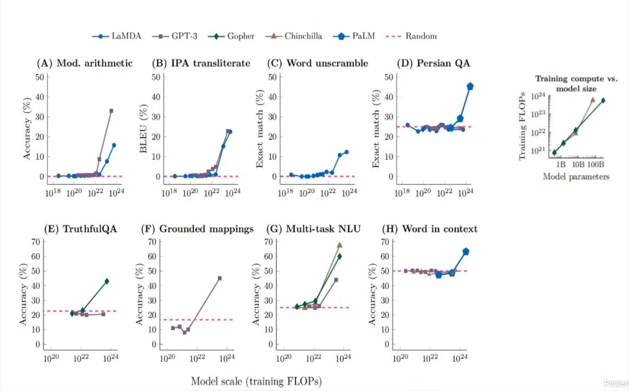
[2206.04615] Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (arxiv.org)

Background: Emergent Abilities in LLM, Some Examples



[2206.07682] Emergent Abilities of Large Language Models (arxiv.org)

Background: Emergent Abilities at scale in LLM



Claims: from previous work

- These abilities are not present in smaller-scale models but are present in large-scale models”
- They cannot be predicted by simply extrapolating the performance improvements on smaller-scale models.

Motivation: What makes emergent abilities intriguing and possibly challenging?

- Sharpness
- Unpredictability

If true, many critical research questions!

- What controls which abilities will emerge?
- What controls when abilities will emerge?
- How can we make desirable abilities emerge faster?
- How can we ensure undesirable abilities never emerge?

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 Google's BIG-Bench occur for 1 of the 2 following metrics:

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

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

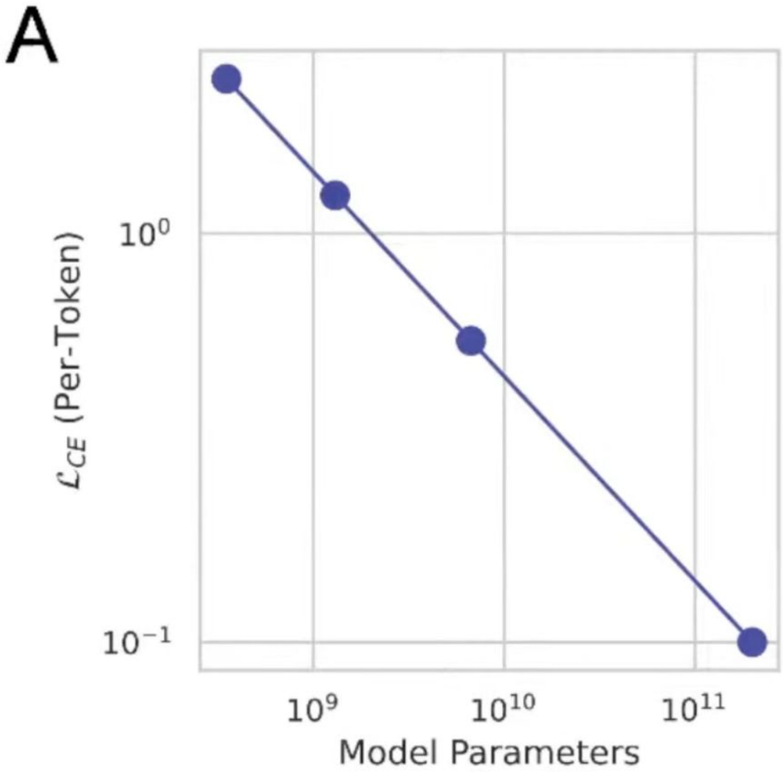
This raises the specter that "emergent" abilities might not be due to fundamental changes in models with scale, but due to our choice of metrics!

Alternative Explanation for Emergent Abilities

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

For concreteness, suppose per-token cross entropy loss follows power law as function of model size (i.e. number of parameters) N

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



Alternative Explanation for Emergent Abilities

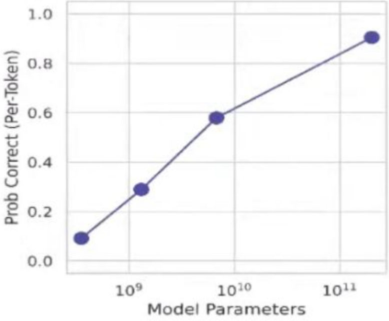
Step 2: Solve for per-token probability of selecting the correct token, as a function of model size

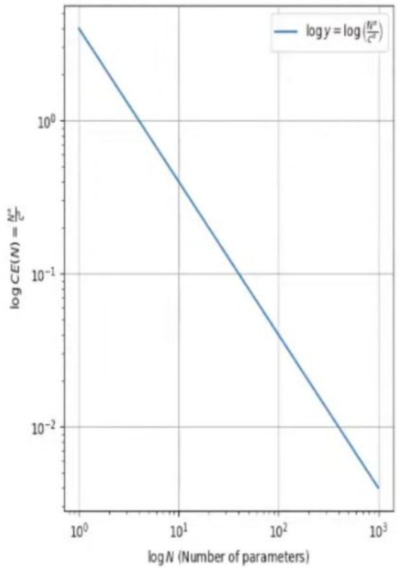
$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha} = -\log \hat{p}_{v^*}(N)$$

$$\hat{p}_{v^*}(N) = \exp\Big(-(N/c)^{\alpha}\Big)$$

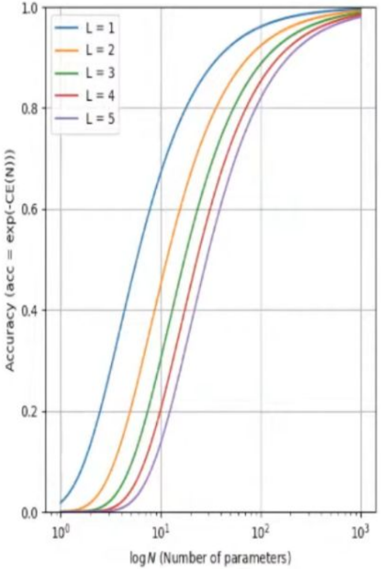
We propose this jump is a **fundamental property of score function** (even for L=1), **not a fundamental property of the model**

Sharp "S" **not unpredictable**!





One only needs scaling laws for this to hold (left)



Sharp "S" shape (Right)

Alternative Explanation for Emergent Abilities

Step 3 (Option A): Choose a metric that **nonlinearly** scales the per-token error rate

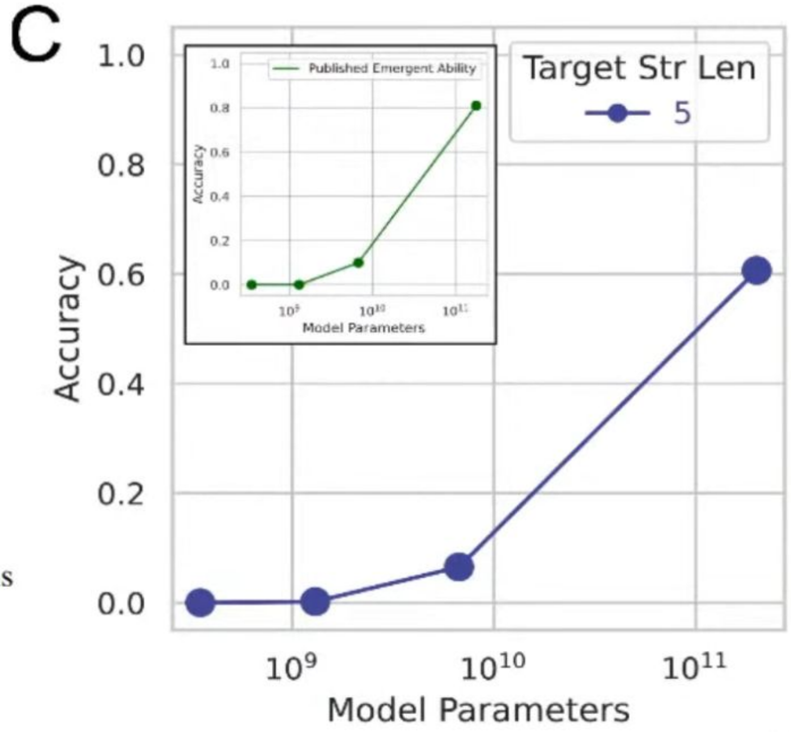
Example: Task is adding 2 K-Digit integers

Measure performance with Exact String Match (Accuracy):

1 if all K+1 digits in model's output are correct

0 otherwise

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



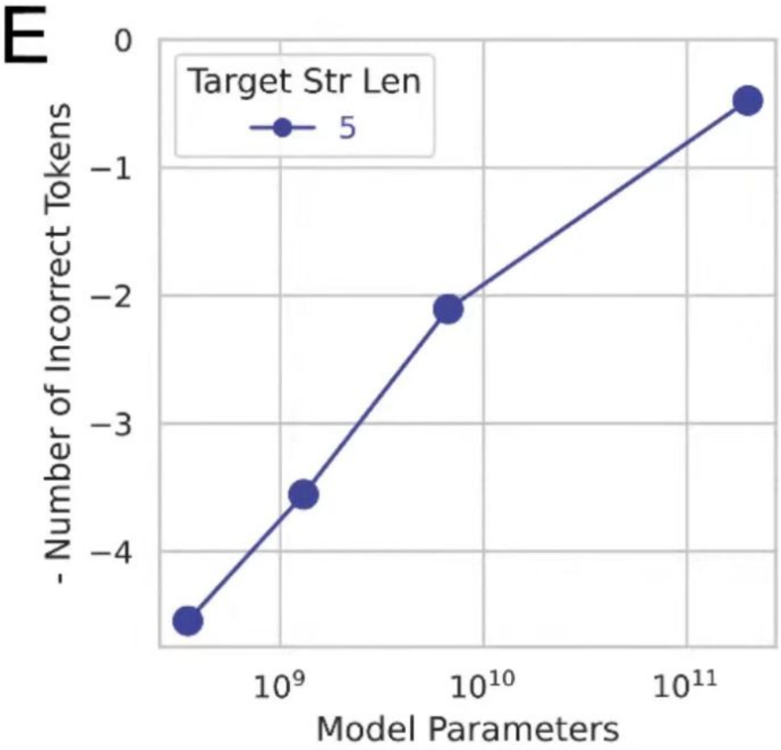
Alternative Explanation: Intervening and Changing the scoring function

Step: Instead choose a metric that **linearly** scales the per-token error rate

Example: Task is adding 2 K-Digit integers

Measure performance with Number of Incorrect Tokens (i.e., Edit Distance)

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



Alternative Explanation for Emergent Abilities

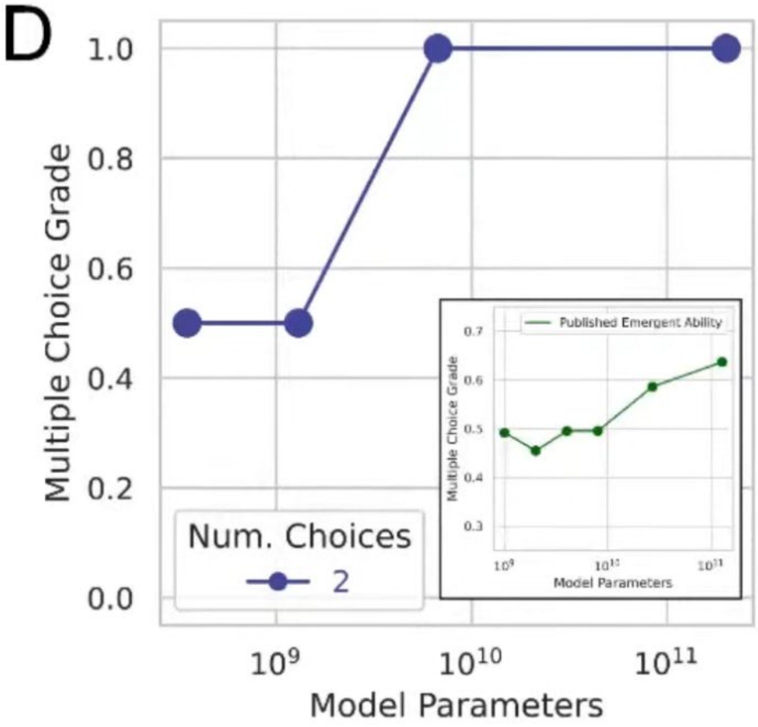
Step 3 (Option B): Choose a metric that **discontinuously** scales the per-token error rate

Example: Task is choosing 1 of 2 multiple choice options

Measure performance with Multiple Choice Grade:

1 if highest probability mass on correct option

0 otherwise



Alternative Explanation for Emergent Abilities

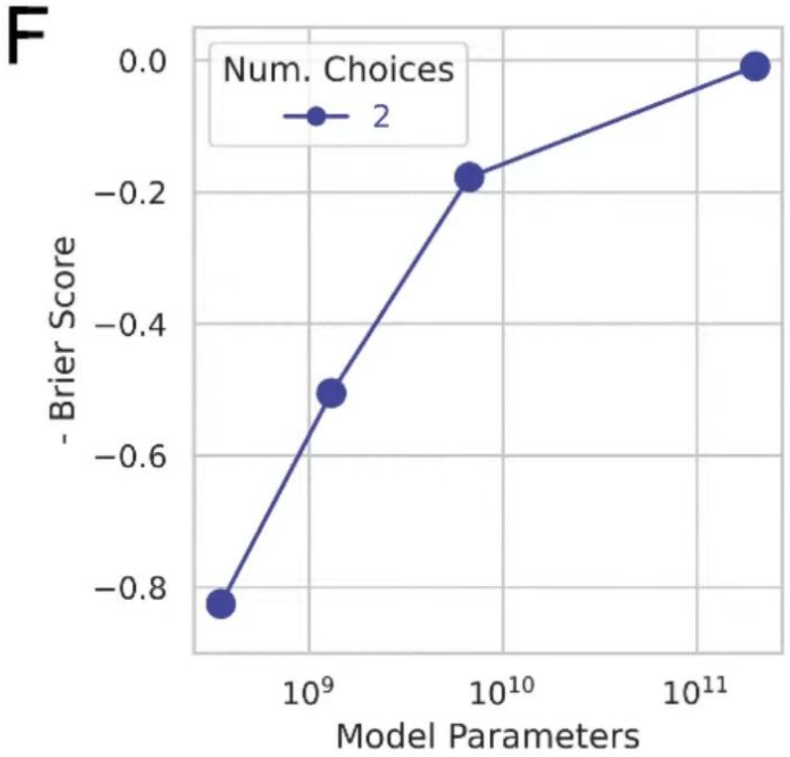
Step: Instead choose a metric that **continuously** scales the per-token error rate

Example: Task is choosing 1 of 2 multiple choice options

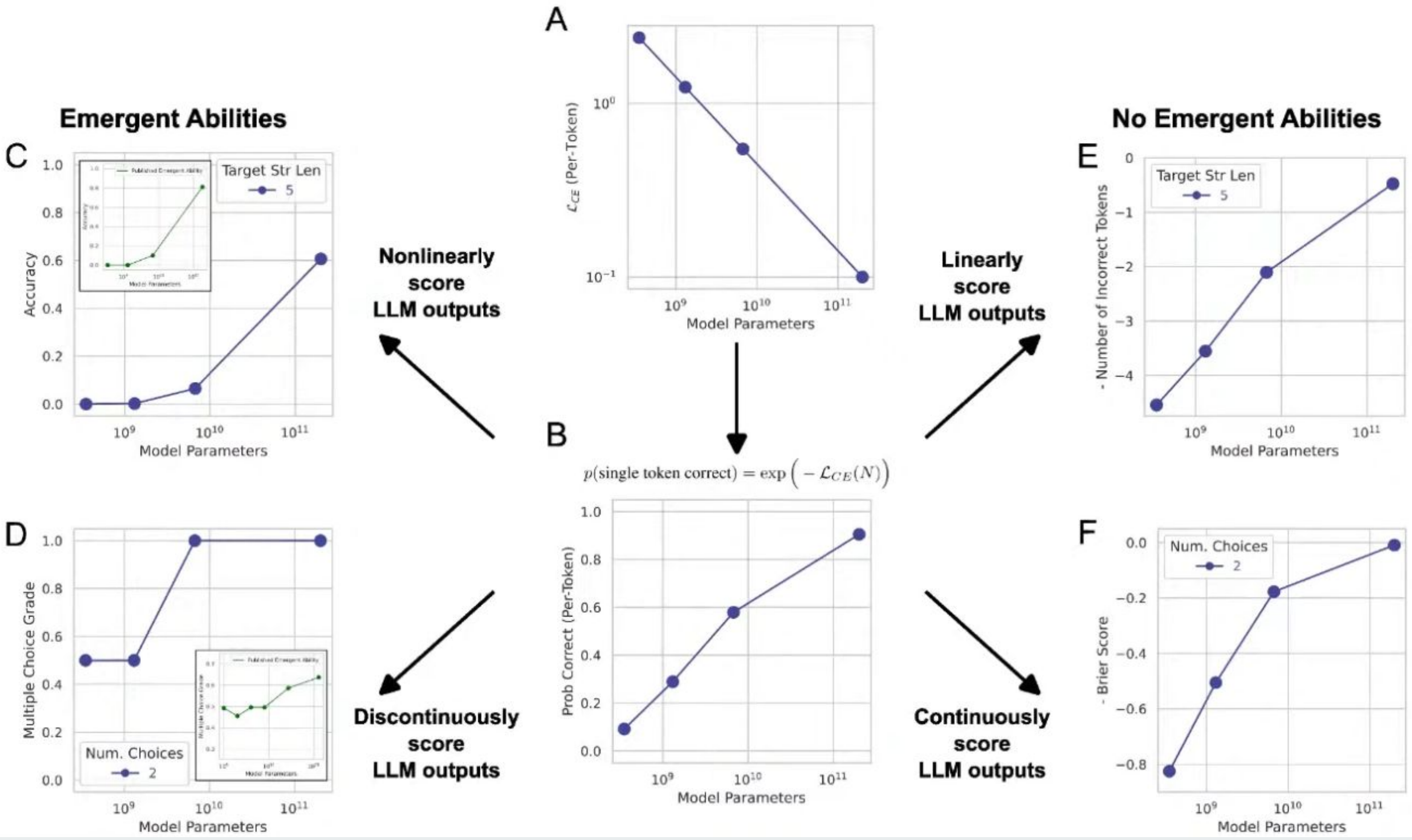
Measure performance with Brier Score

Brier Score = Mean-Squared Error

Brier Score = (1 - probability mass on correct option)^2



Alternative Explanation for Emergent Abilities



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

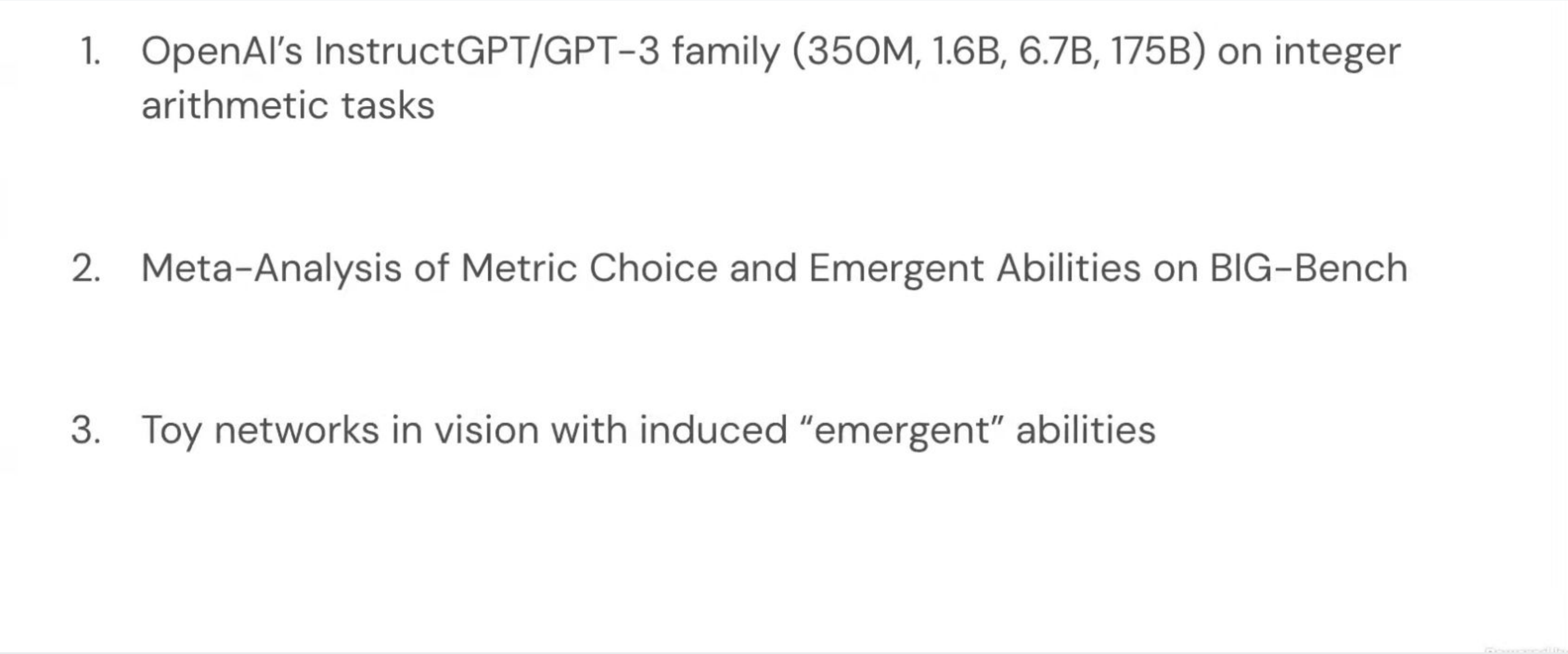
Alternative Explanation for Emergent Abilities

Here, we argue that there are no sharp, unpredictable changes in model capabilities with scale

Hypotheses: Performance across scale is attributable to 3 factors:

1. Using a metric that nonlinearly or discontinuously transforms per-token error rates
2. Using too few test data to accurately estimate the performance of smaller models, thereby causing smaller models to appear wholly unable to perform the task
3. Evaluating too few large-scale models

What are the evidence? Testing on the following models

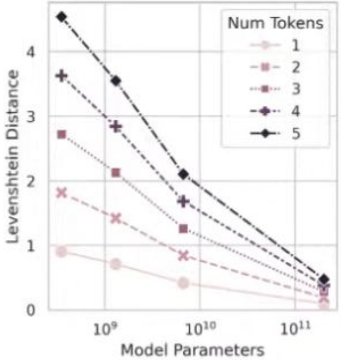


1. OpenAI's InstructGPT/GPT-3 family (350M, 1.6B, 6.7B, 175B) on integer arithmetic tasks

2. Meta-Analysis of Metric Choice and Emergent Abilities on BIG-Bench

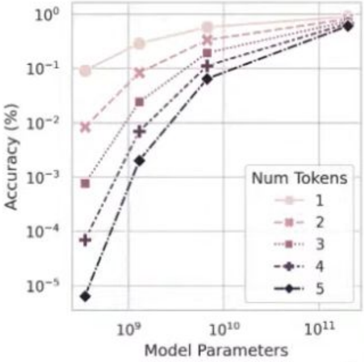
3. Toy networks in vision with induced "emergent" abilities

Predictions for InstructGPT/GPT-3 on Arithmetic Tasks



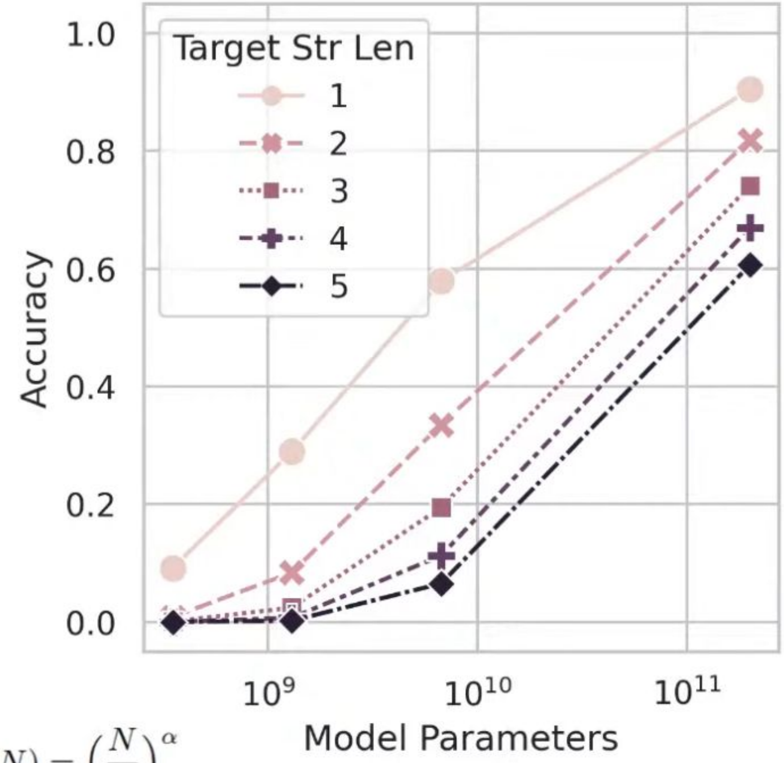
1. **Changing the metric** from a sharp metric (e.g., extact-match, Accuracy-5) to a softer metric (e.g., string edit/Levenshtein distance) should reveal smooth, continuous, predictable performance improvement with model scale

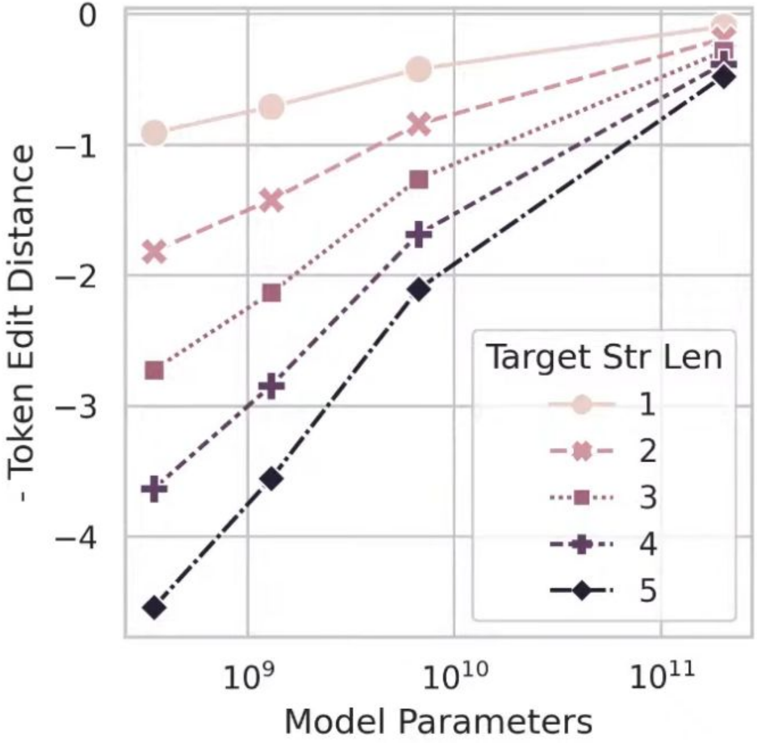
2. **Increasing the resolution** of measured model performance by increasing the test dataset size should reveal above chance performance & smooth, continuous, predictable model improvements on the original sharp metric



3. **Regardless of metric (either sharp or soft), increasing the target string length should affect the model performance** as a function of the length-1 target performance (~geometric for accuracy, ~linear for Levenshtein distance)

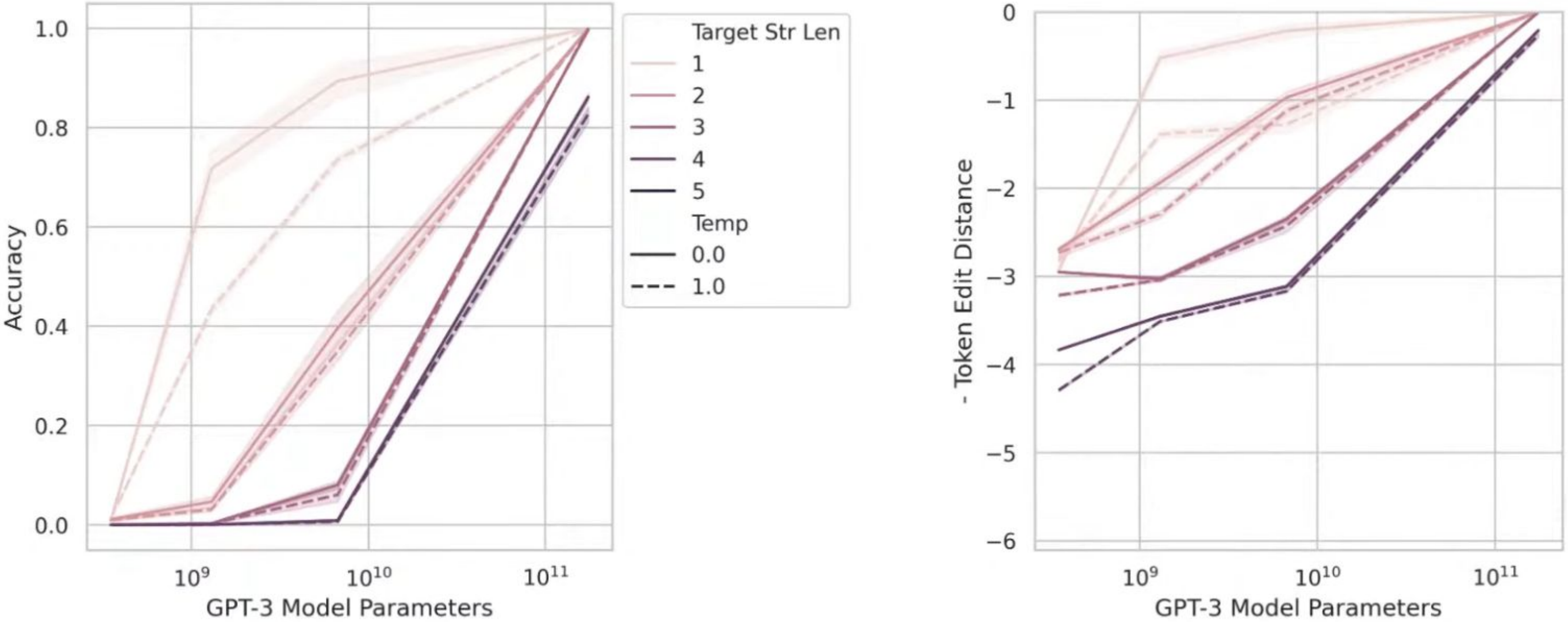
Prediction: Change the Metric, Observe ~Linear Increase



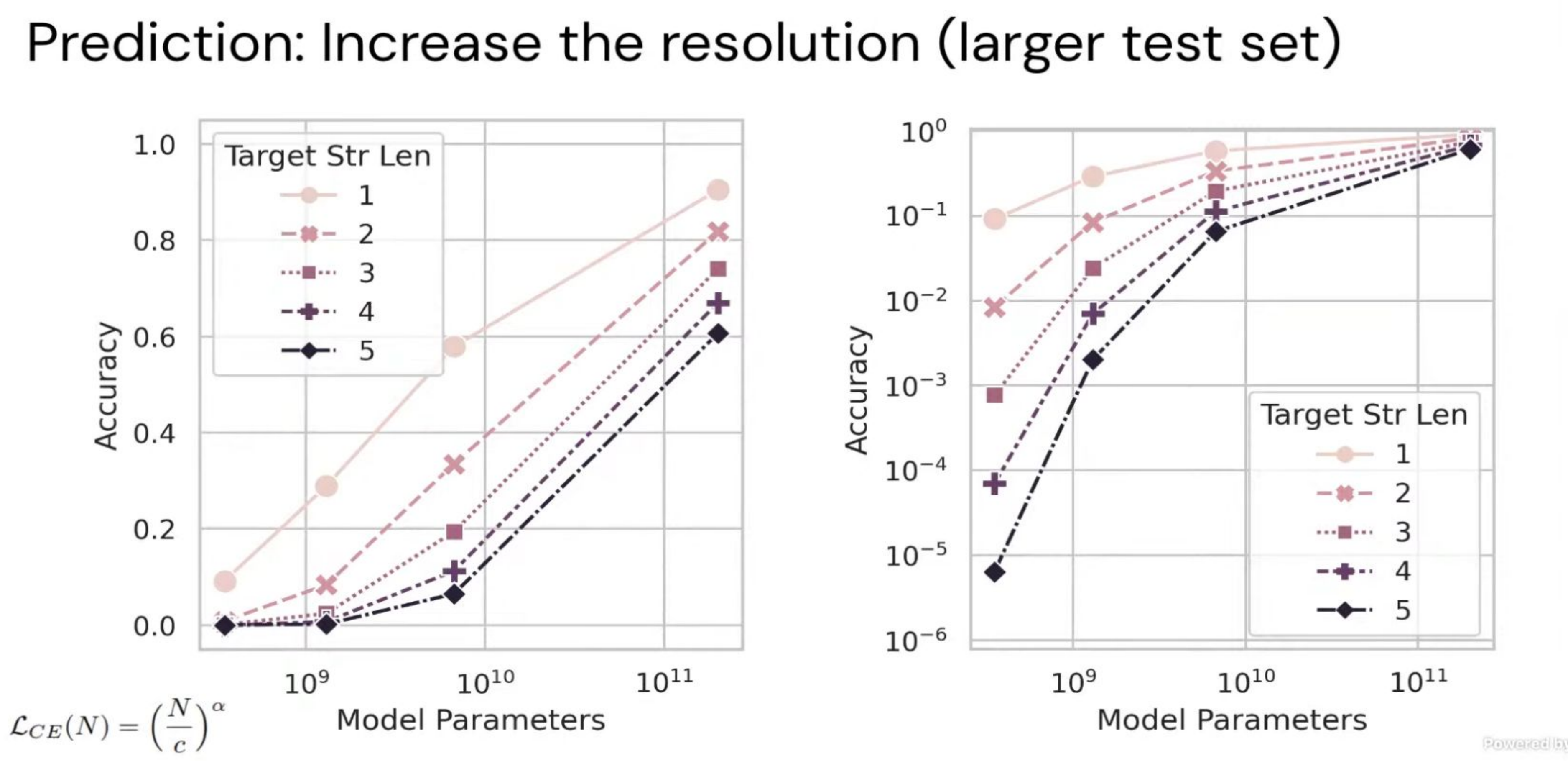


$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha}$$

GPT3.5/3 Two 2-Digit Integer Multiplication



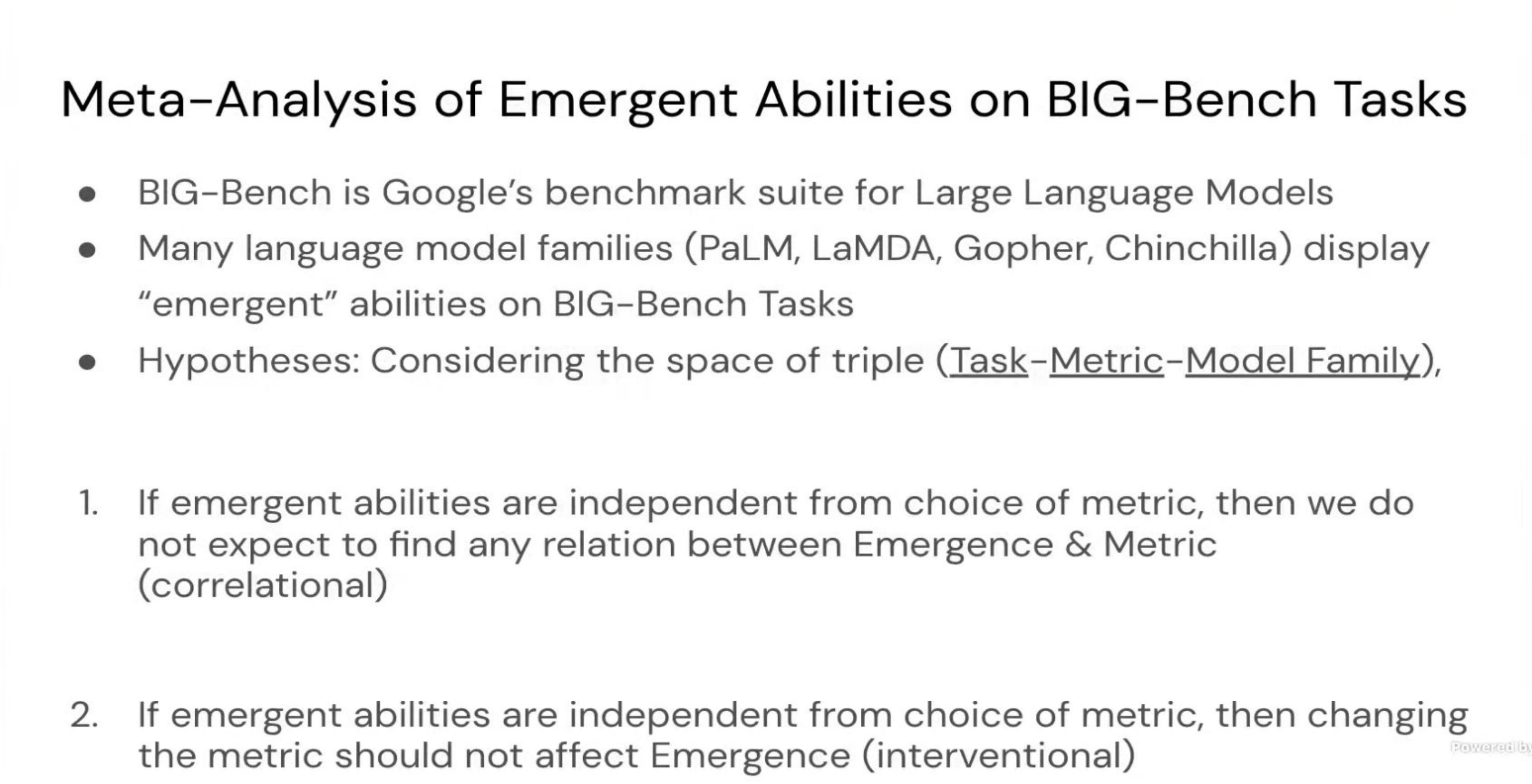
Second Prediction: Increasing the resolution or using Larger Test Set



Prediction: Increase the resolution (larger test set)

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

3rd Prediction: Meta Analysis of Emergent Abilities on BIG-Bench Tasks



Meta-Analysis of Emergent Abilities on BIG-Bench Tasks

- BIG-Bench is Google's benchmark suite for Large Language Models
- Many language model families (PaLM, LaMDA, Gopher, Chinchilla) display "emergent" abilities on BIG-Bench Tasks
- Hypotheses: Considering the space of triple (Task-Metric-Model Family),

1. If emergent abilities are independent from choice of metric, then we do not expect to find any relation between Emergence & Metric (correlational)

2. If emergent abilities are independent from choice of metric, then changing the metric should not affect Emergence (interventional)

Emergence Appears Only Under Specific Metrics

The BIG-Bench paper introduced a quantitative score for whether an ability is emergent

Consider a single Model Family e.g. PaLM

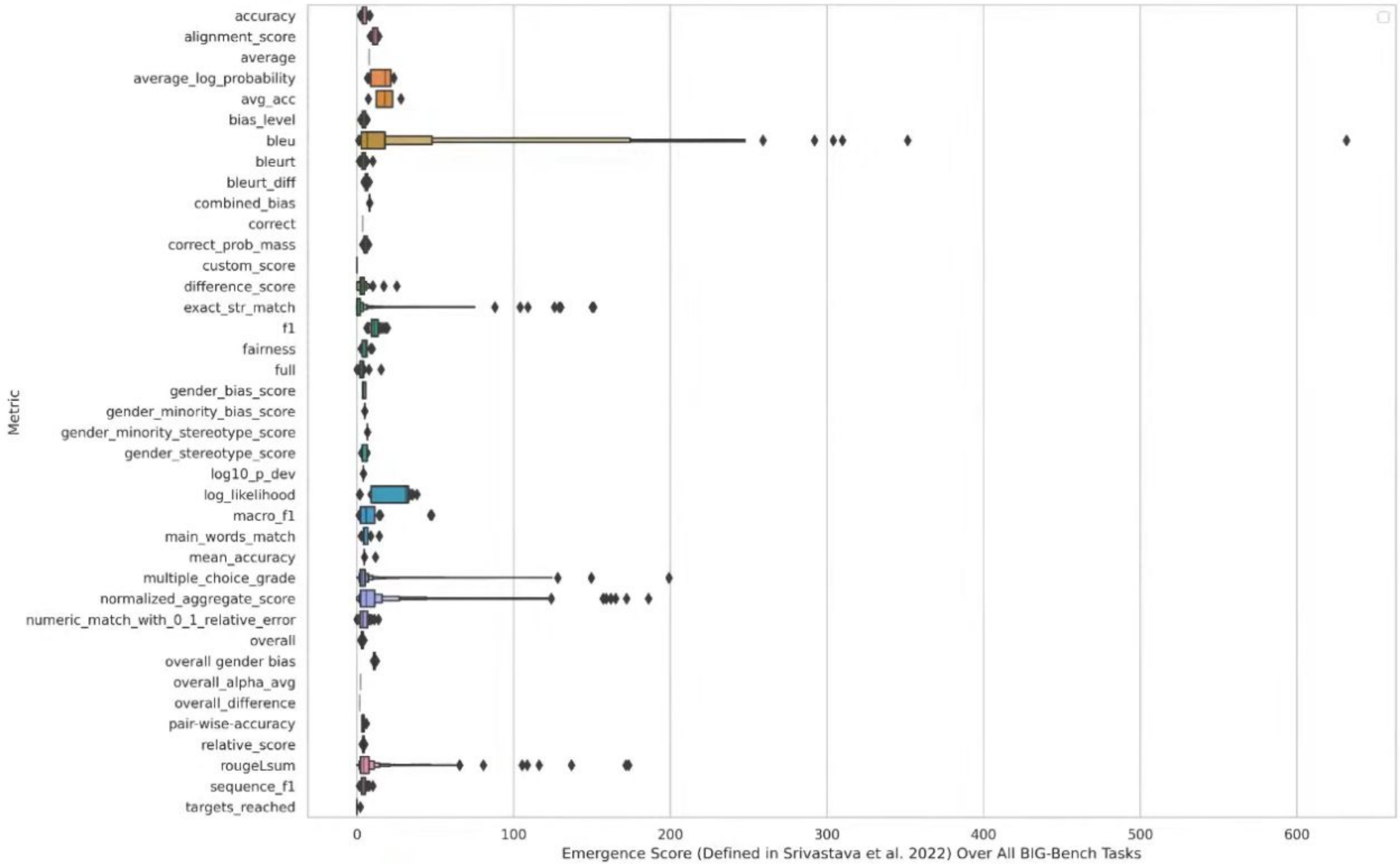
Let x_n be the scale of one family member e.g. PaLM-540B

Let y_n be the family member's score on some Task and Metric

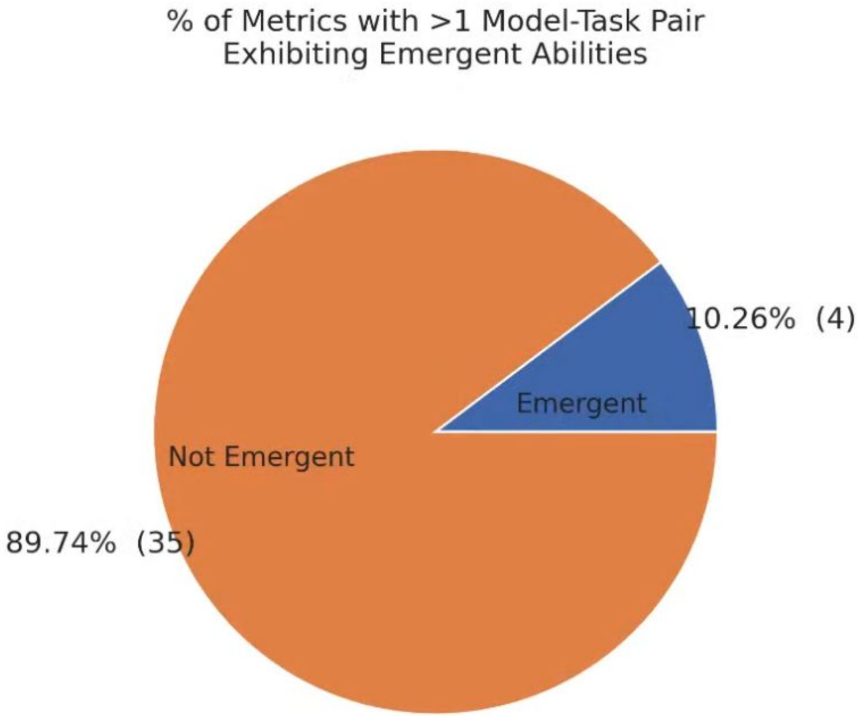
Sort the pairs (x_n, y_n) by model scale x_n, smallest to largest

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

Emergence Appears Only Under Specific Metrics

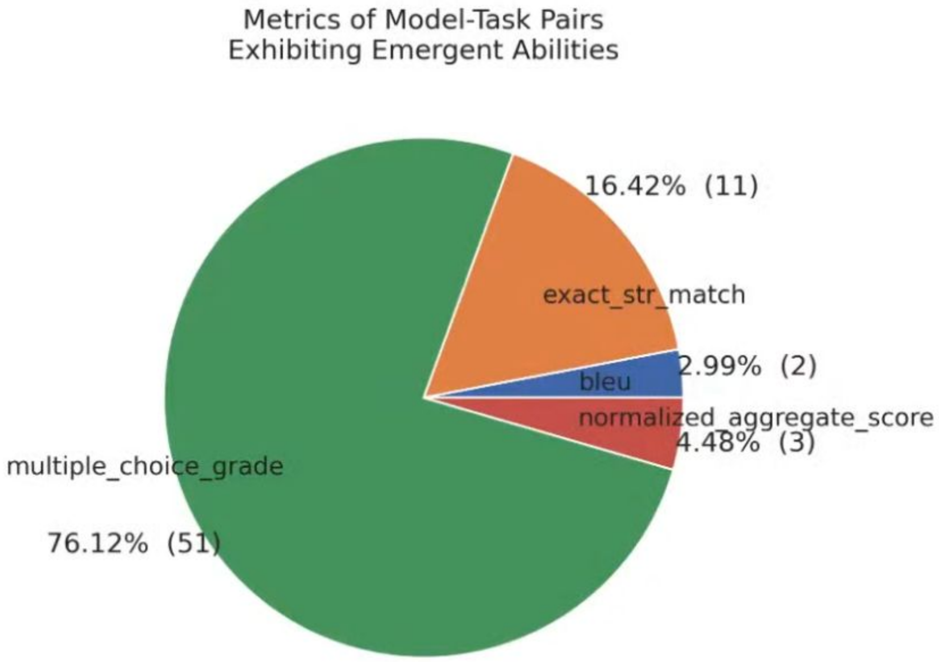


Emergence Appears Only Under Specific Metrics



35 out of 39 BIG–Bench Metrics exhibit no emergent abilities

Emergence Appears Only Under Specific Metrics



Metrics of Model-Task Pairs Exhibiting Emergent Abilities

Of the Tasks where emergent abilities are observed

76% of the tasks' use Metric of Multiple Choice Grade

16.42% of the tasks use Metric of Exact String Match

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

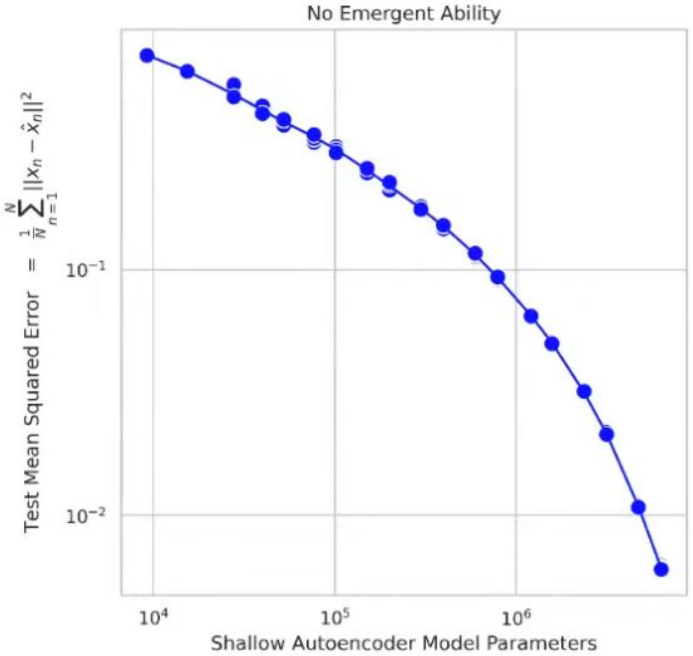
$$\text{Exact String Match} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$$

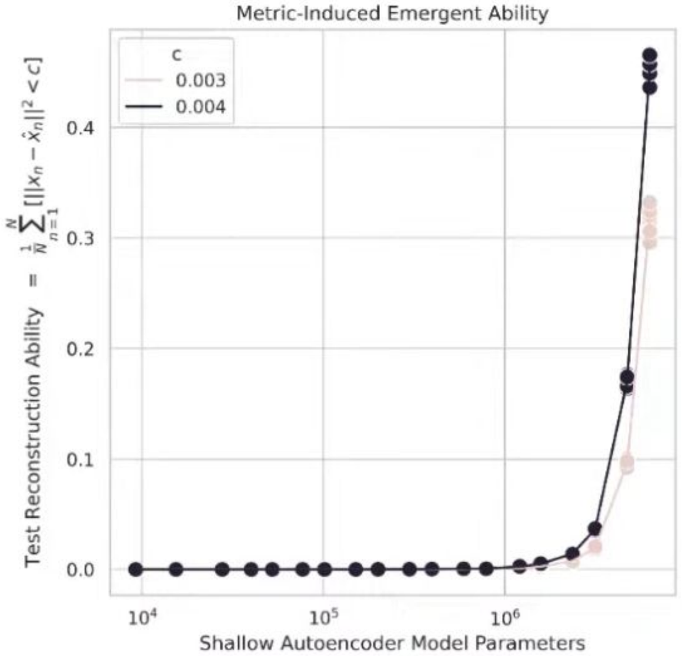
Inducing Emergent Abilities in Vision

- One reason emergent abilities of large language models are so fascinating are that they (to the best of our knowledge) **have never been observed on other ML/AI tasks** (e.g., computer vision tasks)

- To show **how metric choice can produce seemingly emergent abilities**, we induce emergent abilities in diverse architectures on different vision tasks

Emergent Reconstruction of CIFAR100





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

Discussion: Limitations & Open Problems

- **Limitation:** Most leading LLMs are closed. Our analysis is somewhat restricted by ability to query large models (e.g., GPT family provides query access, while PaLM, LaMDA, Gopher, Chinchilla, do/did not)

- independence mathematical model assumption

- Our analysis does not rule out the possibility of emergent abilities, only that current claims of emergence are not sufficiently supported by the evidence.

Conclusion

- (Large Language) Models typically improve with scale, i.e., parameters, data, compute

- Some have argued that some improvements in large models’ capabilities are unpredictable (along with a semi-precise definition of emergence)

- We argue that many claimed emergent capabilities are predictable, as shown by our mathematical model or by using alternative metrics or better statistics

Questions and Discussion