Review: Are Emergent Abilities of Large Language Models a Mirage?

Jeong Han Kyo

January 8, 2024

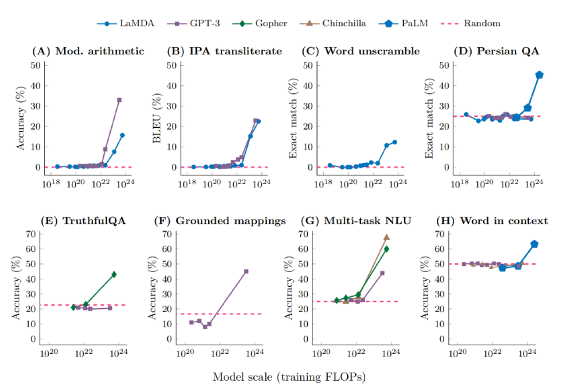
Seoul national university, statistics, IDEA LAB

Table of contents

1. Introduction
    - 1.1. Basic Concept
    - 1.2. Motivation
    - 1.3. Methos and approch
2. Research Methods
    - 2.1. Analyzing InstructGPT/GPT-3's Emergent Arithmetic abilities
    - 2.2. Meta-Analysis of Claimes Emergent Abilities
    - 2.3. Inducing Emergent Abilities in Networks on Vision Tasks
3. Result & limitation
4. Appendix

Basic Concept

- Emergent ability in LLMs
  1. **Sharpness:** transitioning seemingly instantaneously from not present to present
  2. **Unpredictability:** transitioning at seemingly unforeseeable model scales



Motivation

- Recent works claim that emergent ability is fundamental property of LLMs. (Wei et al., Emergent abilities of large language models. 2022)

- Authors present **alternative explanation** for emergent ability
  - Researcher's **choice of metric** rather than due to fundamental changes in model behavior with scale
  - Particularly choice of **a nonlinear or discontinuous metric** can distort the model family's performance to appear sharp and unpredictable.

Methods and approch

1. Verifying and making predictions about the effect of metric choice on tasks showing emergent ability using the InstructGPT/GPT-3 model famliy.

2. Meta-analysis of metric choices for tasks showing emergent ability in the BIG-Bench

3. Demonstrating how to **induce emergent abilities** in vision tasks across deep network architectures by choosing metrics.

Table of contents

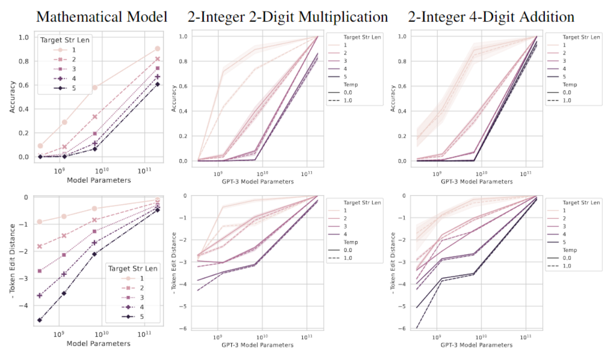
1. Introduction
    1.1. Basic Concept
    1.2. Motivation
    1.3. Methos and approch
2. Research Methods
    2.1. Analyzing InstructGPT/GPT-3's Emergent Arithmetic abilities
    2.2. Meta-Analysis of Claimes Emergent Abilities
    2.3. Inducing Emergent Abilities in Networks on Vision Tasks
3. Result & limitation
4. Appendix

Analyzing InstructGPT/GPT-3's Emergent Arithmetic abilities

1. Changing the metric **from** a nonlinear or discontinuous metric **to a linear or continuous metric** should reveal smooth, predictable performance improvements with model scale.

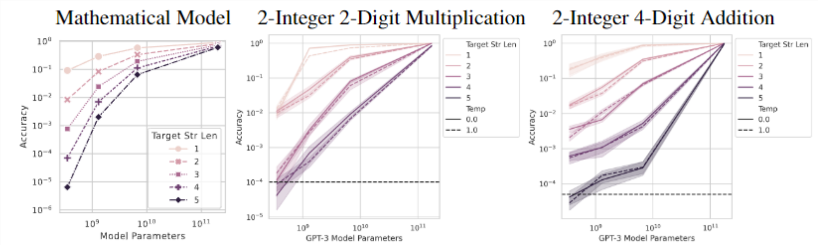
2. For nonlinear metrics, **increasing the resolution** of measured model performance by increasing the test dataset size should reveal smooth, predictable model improvements

3. Regardless of metric, **increasing the target string length** predictable affect the model's performance as a function of the length-1 target performance: approximately geometrically for accuracy and approximately quasilineary for token edit distance.



Claimed emergent abilities evaporate upon changing the metric.

- Top: When performance is measured by a nonlinear metric (e.g., Accuracy), the InstructGPT/GPT-3 familiy'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 performance improvements.



Claimed emergent abilities evaporate upon using better statistics.

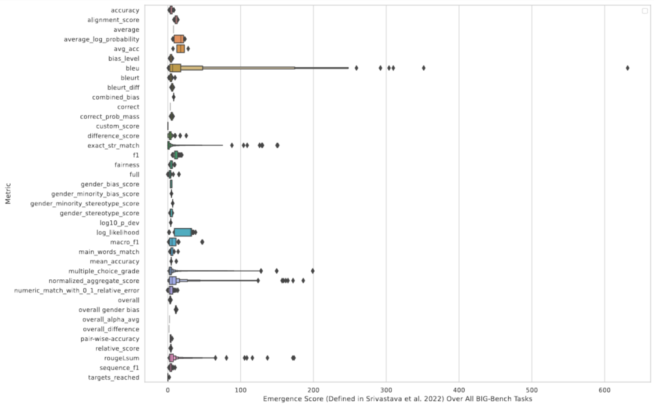
- Generating additional test data increases the resolution and reveals that even on Accuracy, the performance is above chance and improves improves in a smooth, continuous, predictable manner

Meta-Analysis of Claimes Emergent Abilities

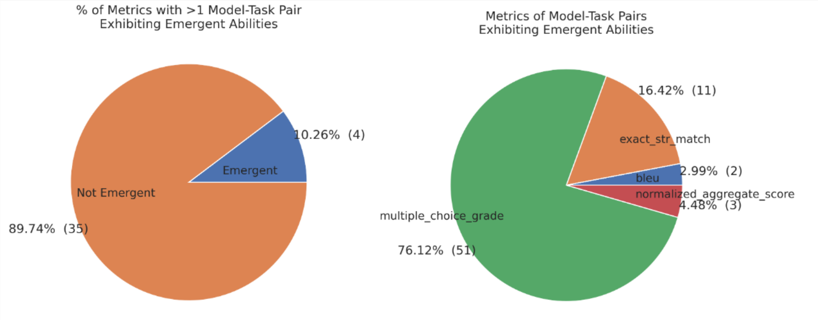
1. At the "population level" of Task-Metric-Model Family triplets, emergent abilities should appear predominantely on **specific metrics**, not task-model family pairs, and **specifically with nonlinear and/or doscontinuous metrics.**

2. On individual Task-Metric-Model Family triplets that display an emergent ability, **changing the metric to a linear and/or continuous metric** should remove the emergent ability.

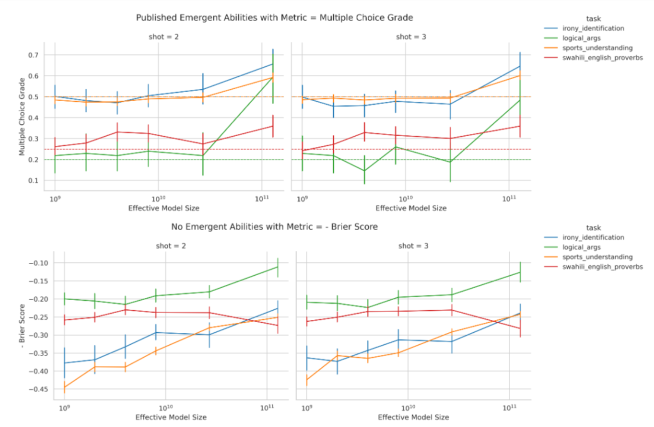
   Using LaMDA family (Thoppilan et al., Language models for dialog applications. 2.22) for its outputs are available through BIG-Bench.



Possible emergent abilities appear with at most **5 out of 39** BIG-Bench metrics using Emergence Score from (Sirvastava et al., Quantifying and extrapolating the capabilities of language models. 2022) ;bleu, exact-str-match, multiple choice grade...



- [Right] because emergence score only suggests emergence, also analyzed hand-annotated task-metric-model family triplets (wet et al., 2022), which revealed emetgent abilities appear with 4/39 metrics.
- [Left] over 92% of emergent abilities appear under one of two metrics: Multiple Choice Grade and Exact String Match

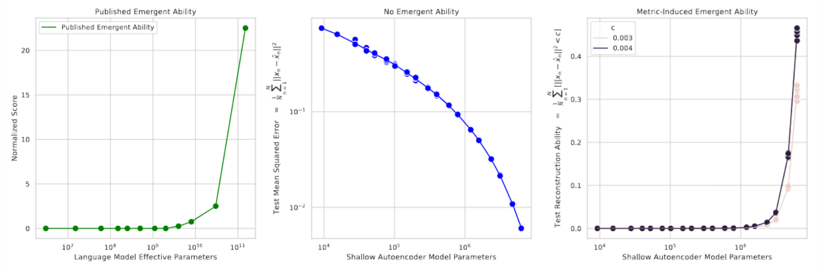


- Changing the metric when evaluating task-model family pairs causes emergent abilities to disappear
- Top: The LaMDA model family displays emergent abilities when measured under the discontinuoous Multiple Choice Grade
- Bottom: The LaMDA model family's emergent abilities disppear when measured under a continuous BIG-Bench metric: Brier Score

Inducing Emergent Abilities in Networks on Vision Tasks

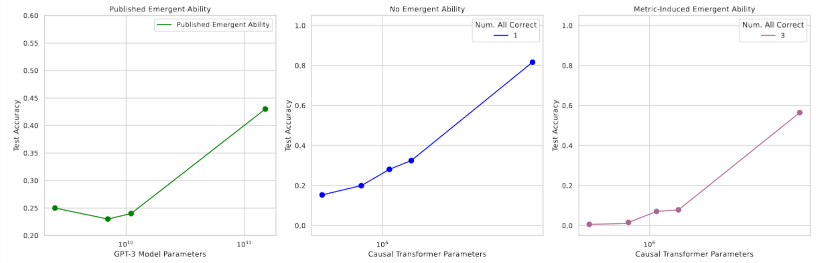
To demonstrate how emergent abilities can be induced by the researcher's choice of metric, authors show **how to produce emergent abilities** in deep networks of various architectures

1. Emergent Reconstruction of CIFAR100 Natural Images by Nonlinear Autoencoders
2. Emergent Classification of Omniglot Characters by Autoregressive Transformers



Induced emergent reconstruction ability in shallow nonlinear autoencoders

- Left: A published emergent ability at the BIG-Bench Periodic Elements task (Srivastava et al., 2022)
- Middle: Shallow nonlinear autoencoders trained on CIFAR100 (Krizhevsky, Learning multiple layers of features from tiny images, 2009) display smoothly decreasing mean squared reconstruction error.
- Right: Using a newly defined Recontruction metric induces an unpredictable change.



Induced emergent classification ability in autoregressive Transformers

- Left: A published emergent ability on the MMLU benchmark (Ganguli et al., Predictability and surprise in large generative model, 2022)
- Middle: Autoregressive transformers trained to classify Omniglot images display increasing accuracy with increasing scale
- Right: When accuracy is redefined as classifying all imgages correctly, a seemingly emergent abiility appears.

Table of contents

1. Introduction
    - 1.1. Basic Concept
    - 1.2. Motivation
    - 1.3. Methos and approch
2. Research Methods
    - 2.1. Analyzing InstructGPT/GPT-3's Emergent Arithmetic abilities
    - 2.2. Meta-Analysis of Claimes Emergent Abilities
    - 2.3. Inducing Emergent Abilities in Networks on Vision Tasks
3. Result & limitation
4. Appendix

Result

- Emergent abilities may be creations of the researcher’s choices, not a fundamental property of the model family on the specific task.

Limitation

1. Authors's experiments and analyses are limited because some LLMs with claimed emergent abilities (e.g., PaLM, Gopher, Chinchiila) are private and not queryable at the time of their analysis.

2. The best metrics arguably depends on human preferences, which may exhibit qualitatively different behavior;

Table of contents

1. Introduction
    - 1.1. Basic Concept
    - 1.2. Motivation
    - 1.3. Methos and approch
2. Research Methods
    - 2.1. Analyzing InstructGPT/GPT-3's Emergent Arithmetic abilities
    - 2.2. Meta-Analysis of Claimes Emergent Abilities
    - 2.3. Inducing Emergent Abilities in Networks on Vision Tasks
3. Result & limitation
4. Appendix

Appendix.A

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

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

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

$$\text{Token Edit Distance}(N) \approx L\left(1 - p_N(\text{ single token correct })\right) = L\left(1 - \exp\left(-(N/c)^{\alpha}\right)\right)$$

Appendix.B

$y_i \in \mathbb{R}$:model performance at model scales $x_i \in \mathbb{R}$, sorted such that $x_i < x_{i+1}$

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

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