Are Emergent Abilities of Large Language Models a Mirage?

Author: Schaeffer et al.

created by paper2slides

2023-04-28

https://arxiv.org/abs/2304.15004

Executive Summary

- Recent work claims that large language models (LLMs) exhibit emergent abilities, which are sudden, unpredictable changes in capabilities at specific scales.
- This paper argues that emergent abilities are artifacts of the chosen evaluation metric rather than innate properties of model scaling.
- Through a mathematical model and empirical analysis, the paper demonstrates that changing the metric can remove the appearance of emergent abilities.
- Broad implications include considerations for metric choice in evaluations and the importance of using proper controls in experiments involving LLMs.

Introduction

- Emergent properties are phenomena where increased system scale leads to new, unforeseen behaviors.
- In large language models, emergent abilities have been reported, drawing significant attention due to their abruptness and unpredictability.
- Key studies: GPT-3, PaLM, LaMDA, and Gopher models have shown such behaviors.
- Motivating Questions:
  - What controls the emergence of specific abilities?
  - How can desirable abilities be made to emerge more quickly?
  - How can undesirable abilities be prevented?

Problems with the Emergence Claim

- Current emergence claims rely on metrics that may nonlinearly or discontinuously scale model error rates.
- For example, metrics like Accuracy and Multiple Choice Grade cause small changes in model output to appear as large jumps.
- These metrics may induce the appearance of emergent abilities even when model improvements are smooth and continuous.

Alternative Explanation: Mathematical Model

- Suppose a model family's per-token cross entropy loss $\mathcal{L}_{CE}$ decreases smoothly with the number of parameters N:

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

- The per-token probability of selecting the correct token:

$$p(\text{correct token}) = \exp\Big(- (N/c)^{\alpha}\Big)$$

- Nonlinear metrics, such as Accuracy, amplify small losses into apparent sharp changes:

$$\text{Accuracy}(N) \approx \exp\Big(- L(N/c)^{\alpha}\Big)$$

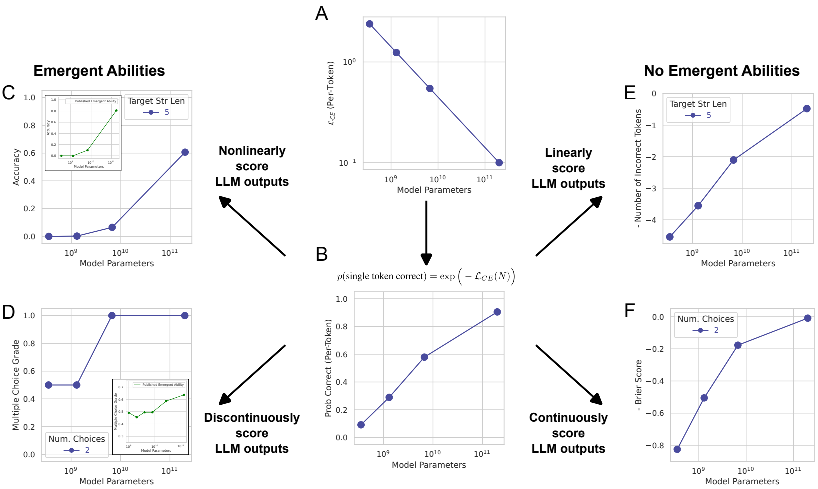

Figure: Metrics impact on perceived model performance.

Empirical Analysis: InstructGPT/GPT-3

- Tasks: 2-digit multiplication and 4-digit addition.
- Metrics: Accuracy vs. Token Edit Distance.

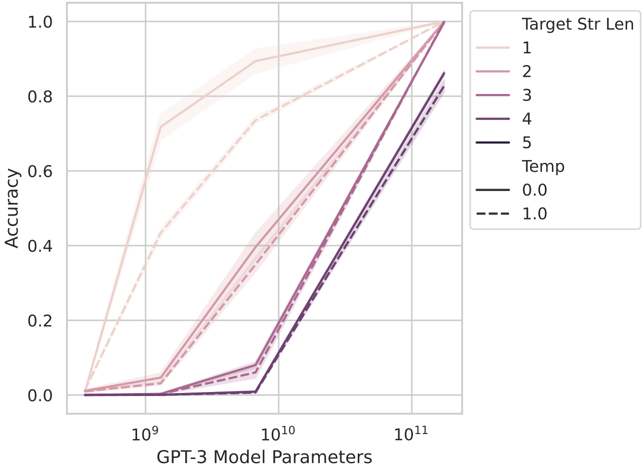

Figure: Accuracy reveals sharp changes in 2-digit multiplication.

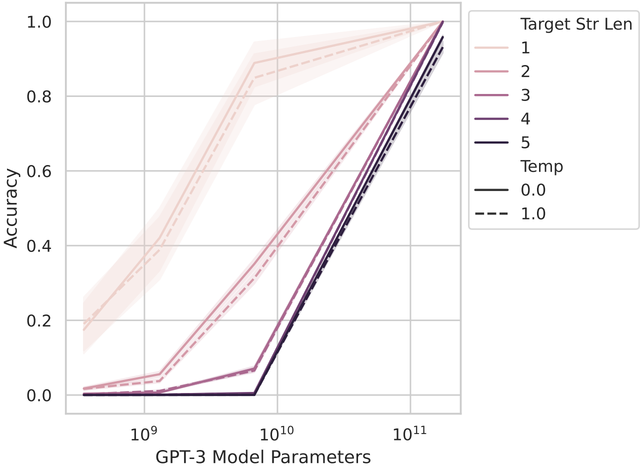

Figure: Accuracy reveals sharp changes in 4-digit addition.

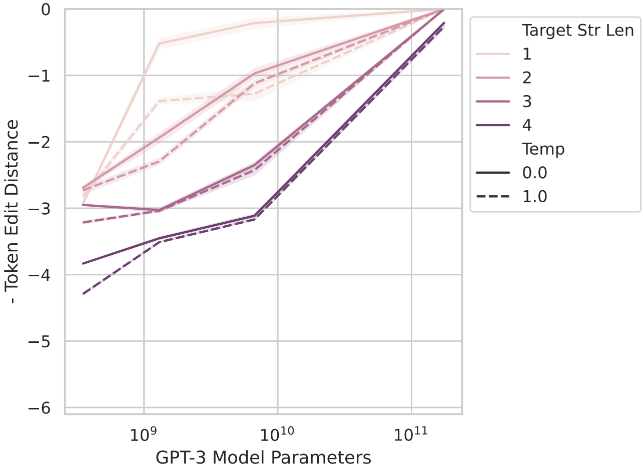

Figure: Token Edit Distance shows smooth improvements in 2-digit multiplication.

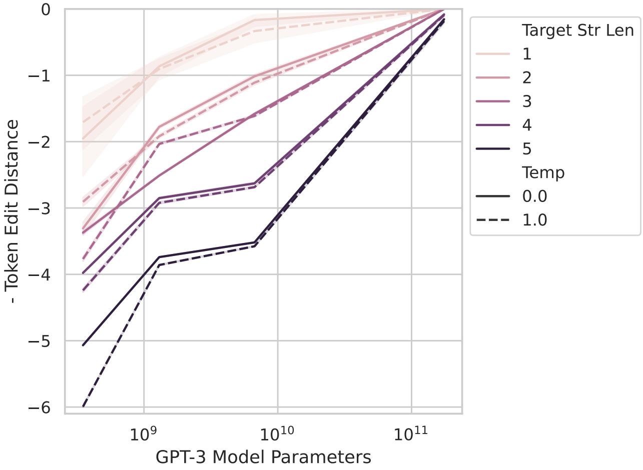

Figure: Token Edit Distance shows smooth improvements in 4-digit addition.

Better Statistics Reduce Emergence

- Increasing the resolution by using more test data helps reveal continuous improvements.
- Analysis showed all models have above-chance accuracy when better statistics are employed.

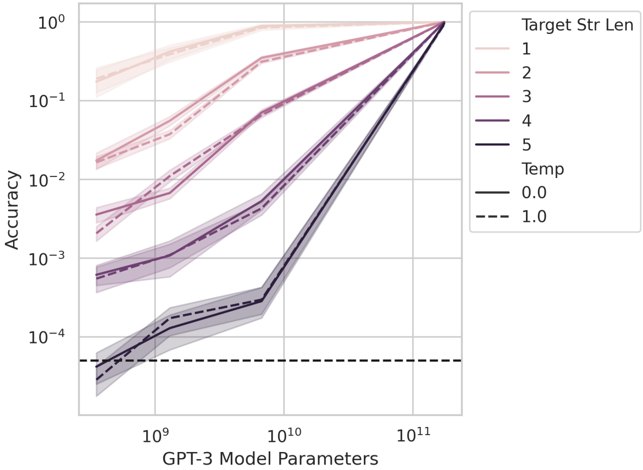

Figure: Increased resolution reveals performance more accurately in 4-digit addition.

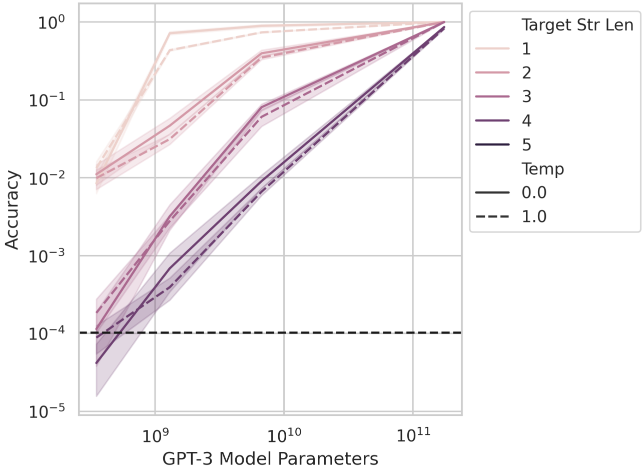

Figure: Increased resolution reveals performance more accurately in 2-digit multiplication.

Meta-Analysis: BIG-Bench

- Analysis of BIG-Bench tasks revealed 92
- Prediction: Changing these metrics should reduce or remove the appearance of emergent abilities.
- Metrics: Multiple Choice Grade vs. Brier Score in the LaMDA model.

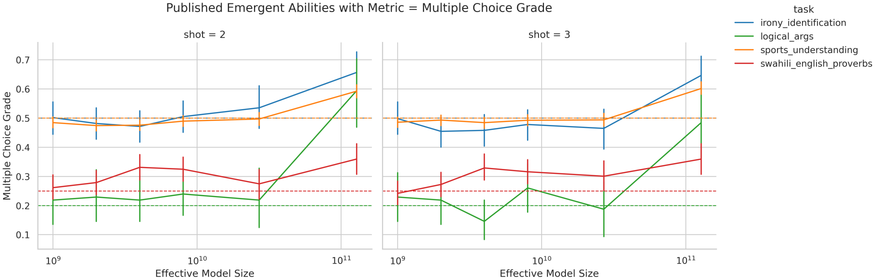

Figure: LaMDA shows emergent abilities under Multiple Choice Grade.

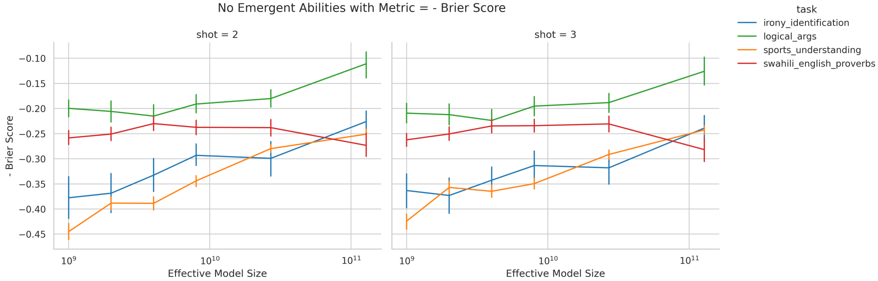

Figure: The same abilities are smooth under Brier Score.

Inducing Emergence in Vision Tasks

- Demonstrated how emergent abilities can be induced in networks across different architectures and tasks by changing metrics.
- Example tasks: CIFAR100 reconstruction with autoencoders, MNIST classification with CNNs.
- Emphasizes that the choice of metric can create emergent abilities where none innately exist.

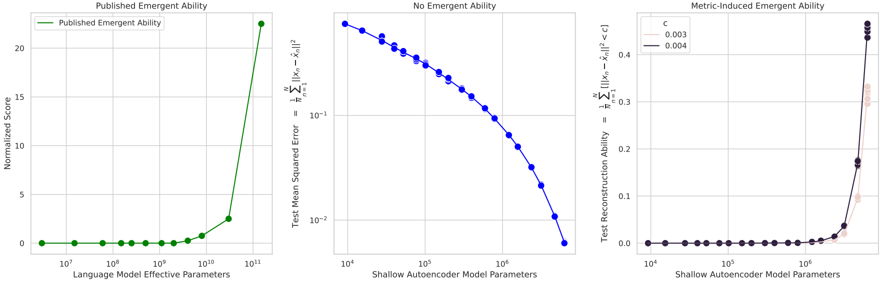

Figure: Induced emergent reconstruction ability in shallow nonlinear autoencoders for CIFAR100.

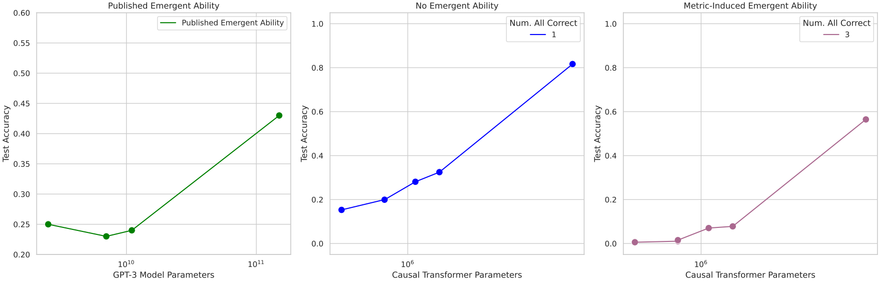

Figure: Induced emergent classification ability in autoregressive Transformers for Omniglot.

Conclusion and Implications

- Emergent abilities in large language models are likely a mirage created by choice of evaluation metrics.
- Important implications for:
  - Benchmark creation
  - Metric selection
  - Interpreting model abilities
- Future work should ensure proper controls and better statistical practices to accurately assess model improvements.