Are Emergent Abilities of Large Language Models A Mirage?

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

[1]Computer Science, Stanford University

Summary & Contributions

- Recent work claims large language models display *emergent abilities* [4, 2, 3, 1].
- What makes emergent abilities intriguing is two-fold: *sharpness* and *unpredictability*.
- We posit an alternative hypothesis: For a task & model family, when analyzing *fixed* model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in model behavior with scale.
- Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance.
- We test and confirm our alternative explanation in 3 complementary ways.

Background: Emergent Abilities of Large Language Models

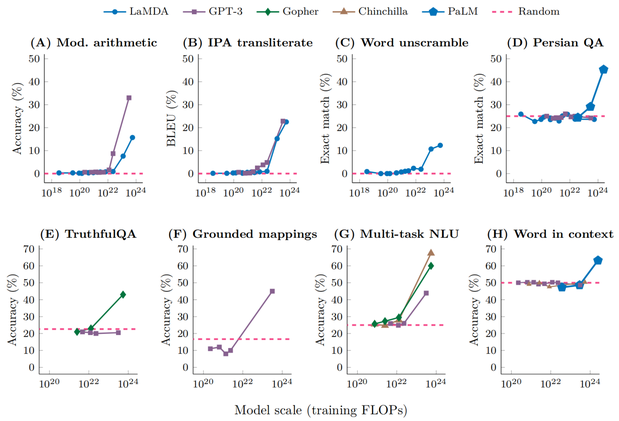
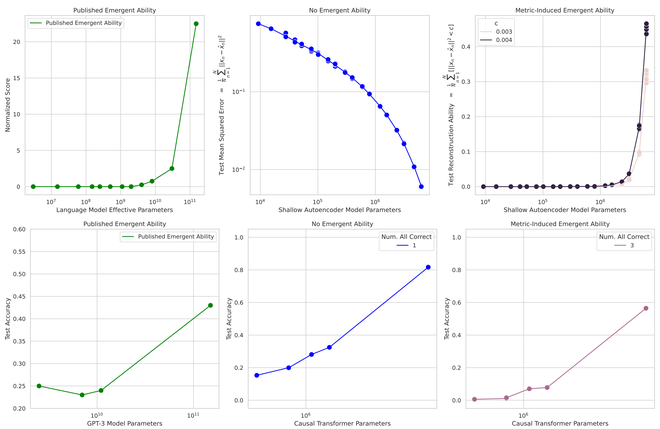
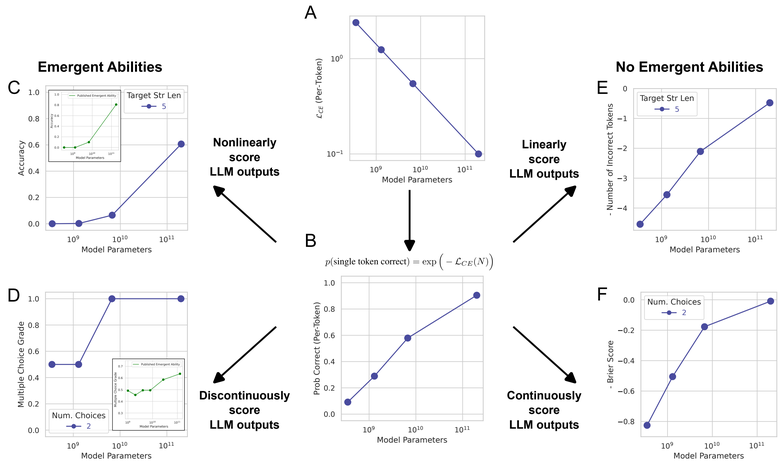


Figure 1. Emergent abilities of large language models. Figure from [4].

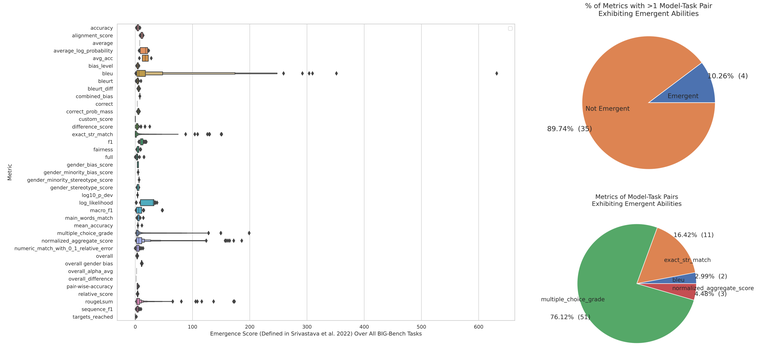
Emergent Abilities Can Be Induced on Vision Tasks via Metric Choice



Alternative Explanation for Origin of Emergent Abilities



BIG-Bench Emergent Abilities Only Appear Under Specific Metrics



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

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

References

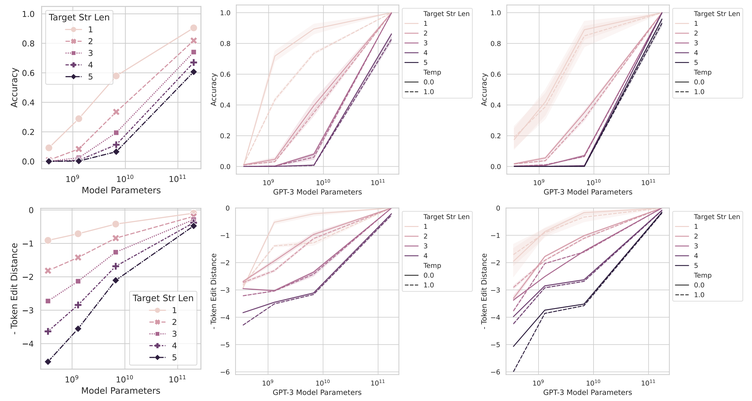
[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. Language models are few-shot learners.

[2] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, et al. Predictability and surprise in large generative models.

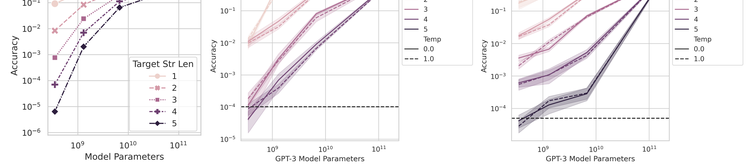
[3] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, et al. Emergent abilities of large language models.

GPT3/3.5 Emergent Abilities Vanish After Changing Metrics



GPT3/3.5 Emergent Abilities Vanish With Better Statistics



BIG-Bench Emergent Abilities Vanish After Changing Metrics