Are Emergent Abilities of Large Language Models a Mirage?

Denis Sapozhnikov, AMI 202

Disclaimer

While what follows will largely go in opposition to the paper "Emergent Capabilities of Large Language Models", it is important to emphasize the significance, relevance and purity of the experiments in this paper. It performed a large meta-analysis combining several families of models and several metrics and indeed confirmed the fact that there are emergent abilities for some problems that only arise on large models.

Feel like a real review and scientist

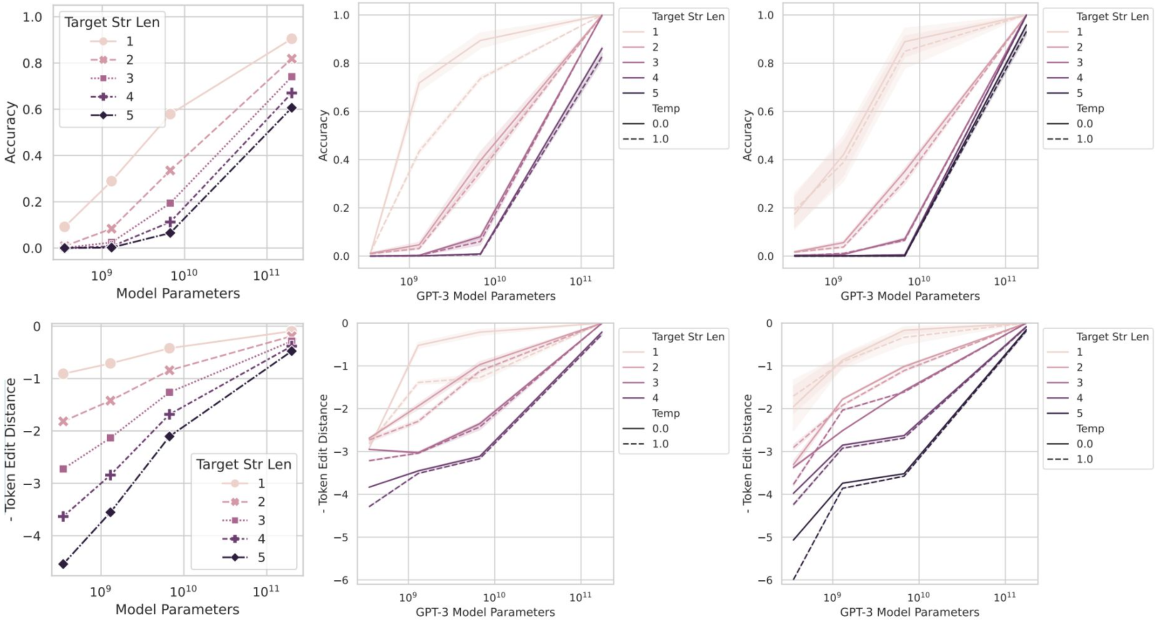
- What weaknesses do you see in the work described?
- What additional experiments would you do?
- Do you have any ideas why Emergent Abilities might occur?

Why Emergent Abilities might occur?

- If a multi-step reasoning task requires L steps of sequential computation, this might require a model with a depth of at least O(L) layers
- Better memorization
- Bad Evaluation metrics for multi-step problems (Exact String Matching, for instance)

Multiplication and Addition tasks

- Discrete metric of correctness
- Nonlinear metric
- Suggested approach
    - ○ Token Edit Distance
    - ○ Split task by length of numbers



Math “proof”

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

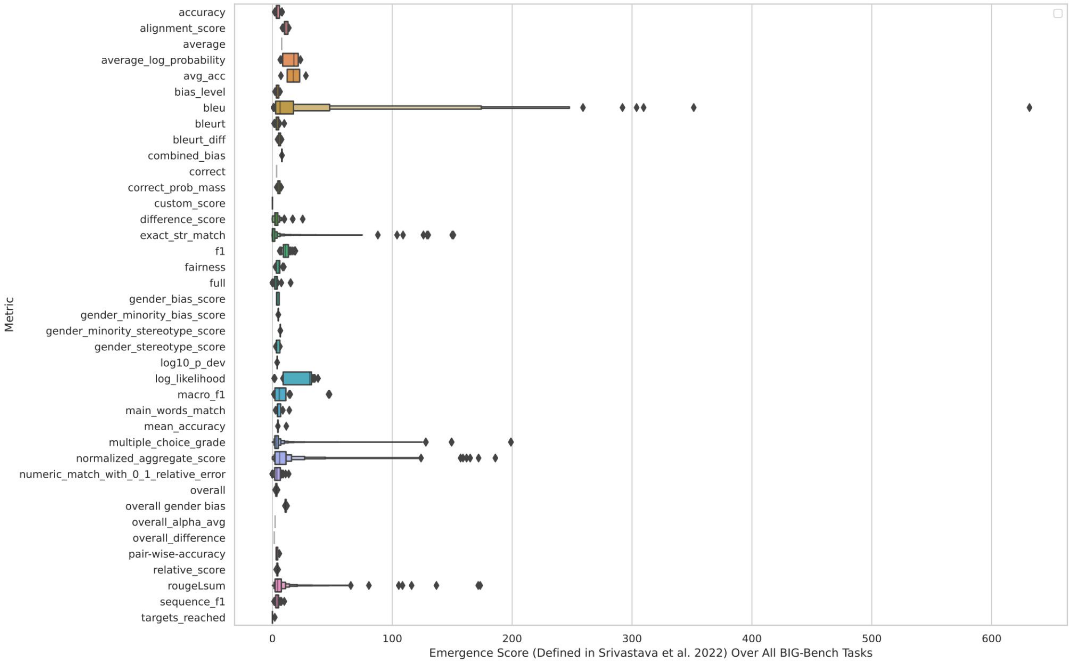
$$p(\text{single token correct}) = \exp\Big(- \mathcal{L}_{CE}(N)\Big) = \exp\Big(- (N/c)^{\alpha}\Big)$$

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

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

Are emergency is a task-model property?

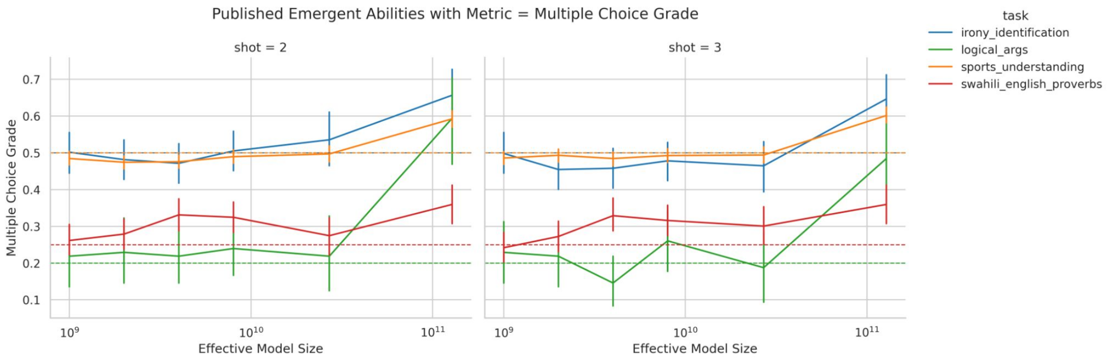
$$\text{Emergence Score}\left(\left\{(x_n, y_n)\right\}_{n=1}^{N}\right) \overset{\text{def}}{=} \frac{\text{sign}(\arg\max_i y_i - \arg\min_i y_i)(\max_i y_i - \min_i y_i)}{\sqrt{\text{Median}(\{(y_i - y_{i-1})^2\}_i)}} \tag{1}$$

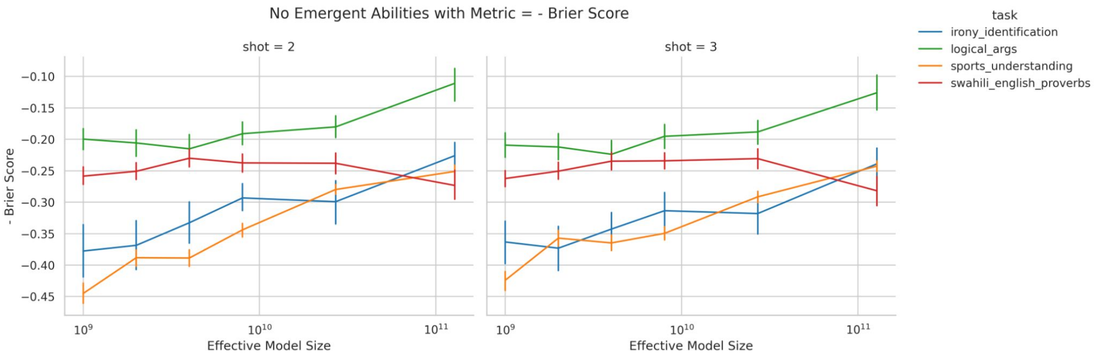


Are emergency is a task-model-metric property?

Brier score:

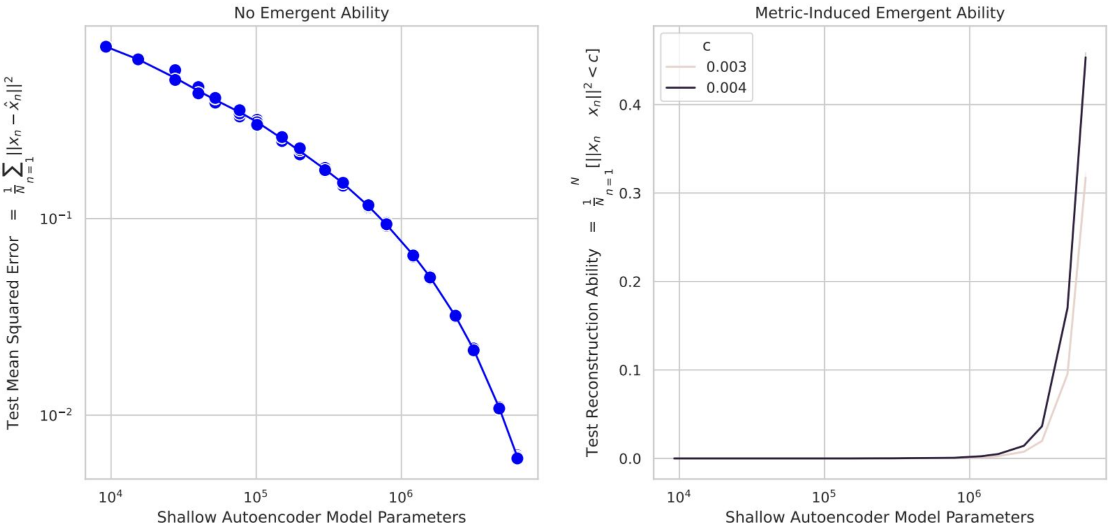
$$\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti}-o_{ti})^2$$



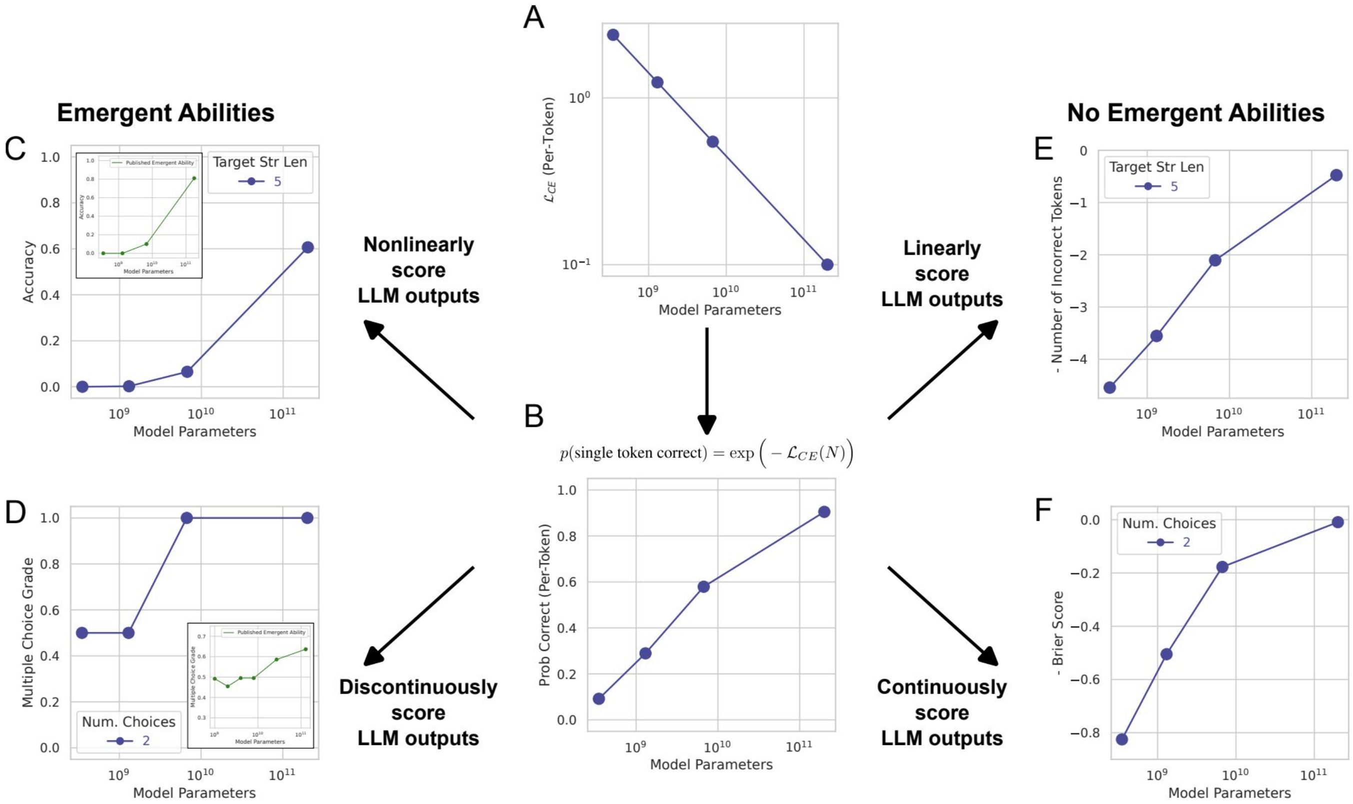


Inducing Emergent Abilities in Networks on Vision Tasks

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



TLDR



Emergent Abilities

No Emergent Abilities

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

Nonlinearly score LLM outputs

Linearly score LLM outputs

Discontinuously score LLM outputs

Continuously score LLM outputs

Discussion section