Are Emergent Abilities of Large Language Models a Mirage?

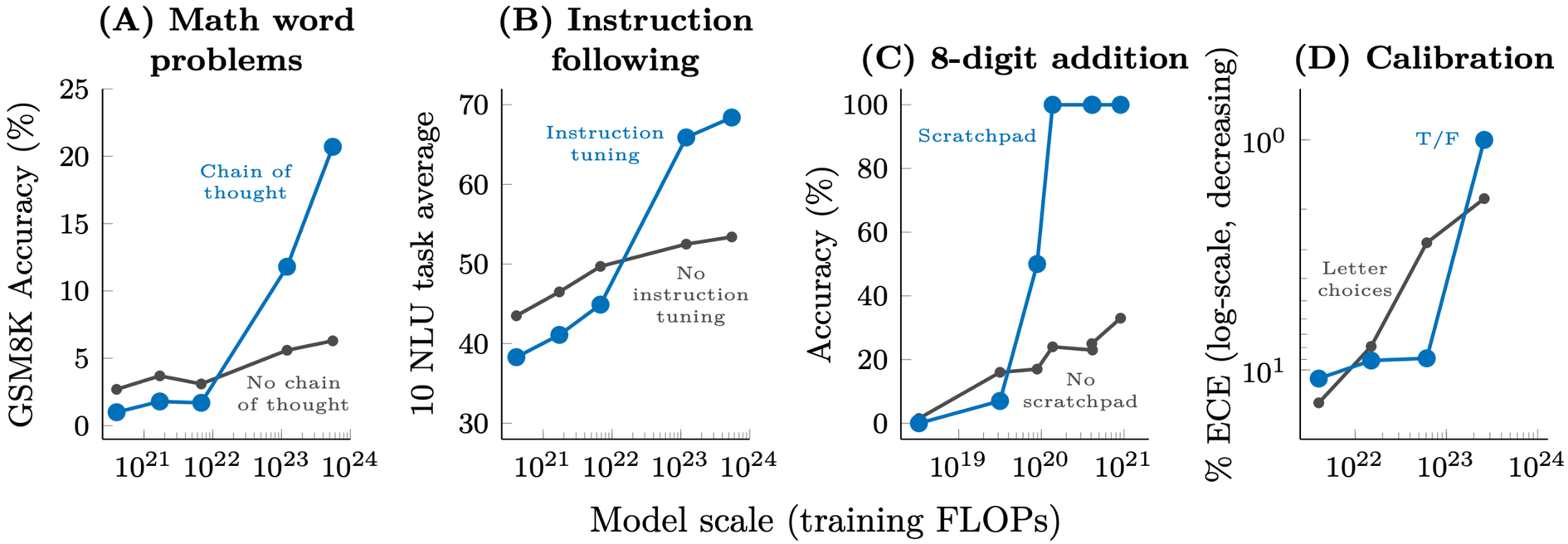
ReFORM Reading Group — 3/5/26

Background

- Wei et al. (2022) — language models exhibit **emergent abilities**
- These abilities are present in large models but not smaller ones, and hence cannot “[be] directly predicted by extrapolating a scaling law”

Background

- Wei et al. (2022) — language models exhibit **emergent abilities**
- These abilities are present in large models but not smaller ones, and hence cannot “[be] directly predicted by extrapolating a scaling law”



Is this real?

- Schaeffer, Miranda, Koyejo (NeurIPS 2023) — **no**
- Main argument — emergent abilities are a consequence of **discontinuous metrics**
    - Underlying quantity is per-token correctness probability
    - Metrics which are smooth in this do not exhibit emergence
    - We can induce emergence by picking new, discontinuous metrics for familiar tasks

A Simple Model

- To start, assume a smooth scaling law for test loss, e.g.

$$L_{CE}(N) = \left(\frac{N}{c}\right)^{\alpha}, \; L_{CE}(N) = -\log \hat{p}_N(v^*)$$

- By rearrangement, the probability of a single correct token is

$$\exp\left(-(N/c)^{\alpha}\right)$$

- So, if tokens are independent, the probability of L correct tokens is

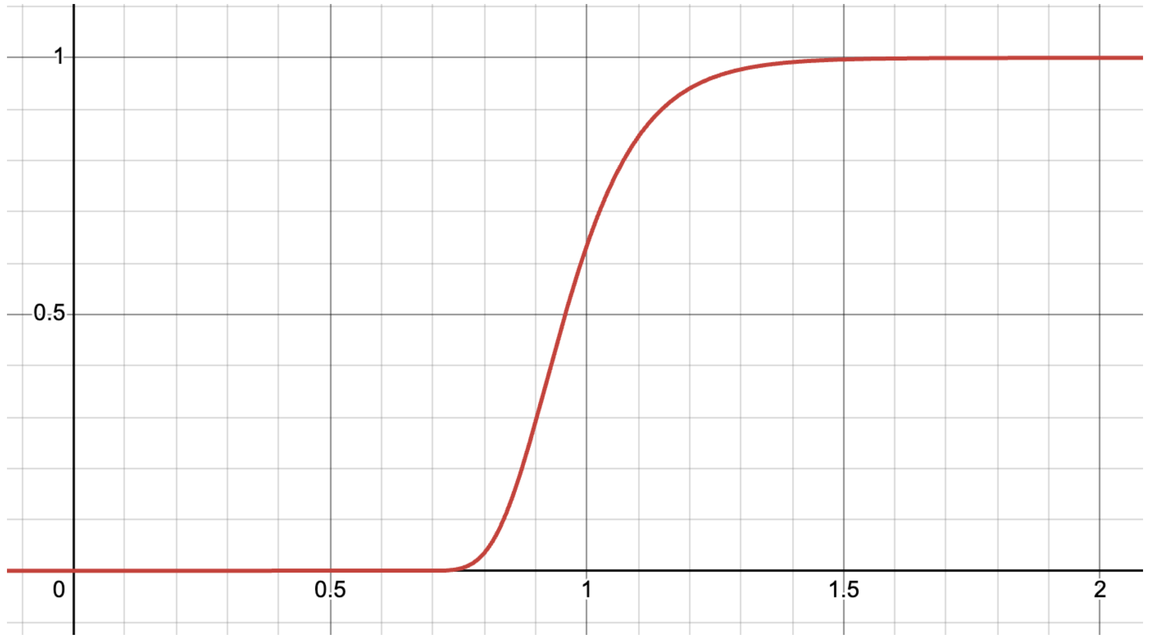
$$\exp\left(-L(N/c)^{\alpha}\right)$$

A Simple Model cont.

- On a plot with x-axis as log(params) and y-axis as accuracy:

$$y = \exp(-Ae^{-\alpha x})$$

- This can be very discontinuous (A = 10000, α = 10):

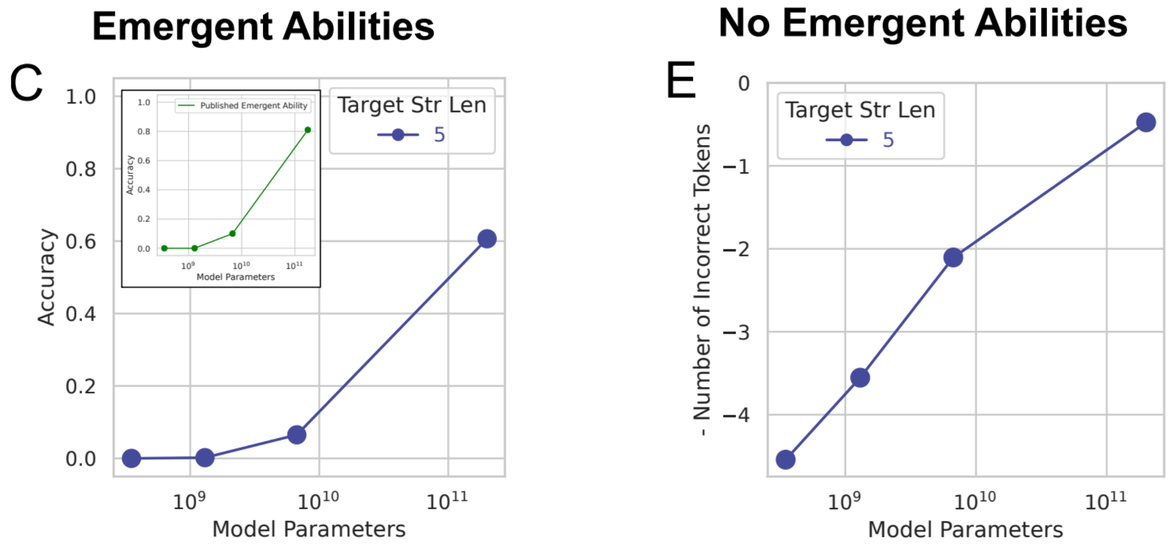


A Simple Model cont.

- Not all metrics scale geometrically — for token edit distance (number of incorrect tokens), we instead get

$$TED(N) \approx L(1 - \exp(-(N/c)^{\alpha}))$$

- This is not linear in $\log N$, but is often smoother



Empirics

- Metric importance is validated using OpenAI API models — 350M, 1.3B, 6.7B, 175B models
- Main tasks of interest — integer addition and multiplication (famous examples, also precisely rely on token accuracy)
- Main prediction — emergence will be visible with accuracy as a metric, but not with token edit distance

Empirics

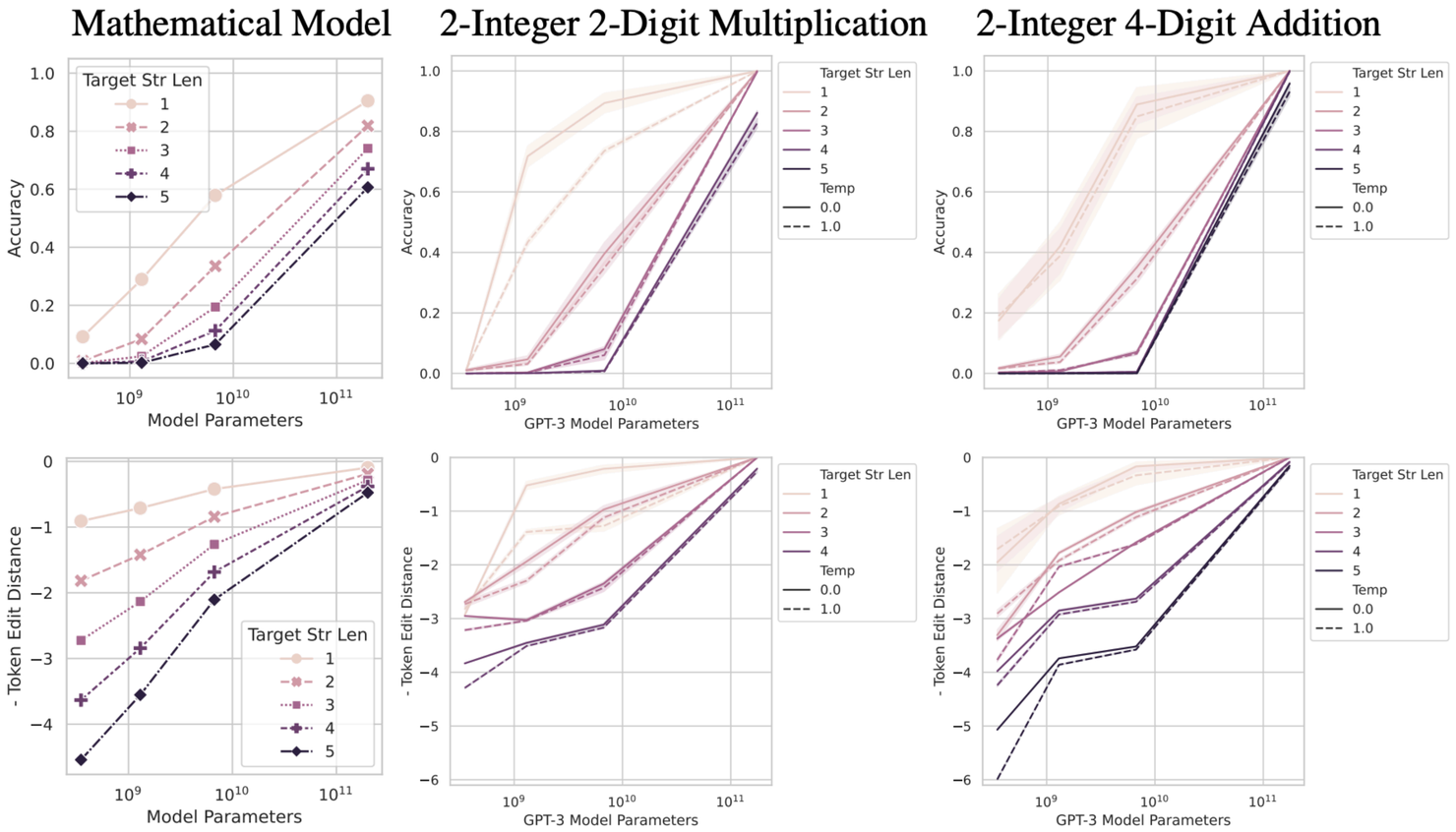


Figure 3: **Claimed emergent abilities evaporate upon changing the metric.** Top: When performance is measured by a nonlinear metric (e.g., Accuracy), the InstructGPT/GPT-3 [4, 27] family's performance appears sharp and unpredictable on longer target lengths. Bottom: When performance is instead measured by a linear metric (e.g., Token Edit Distance), the family exhibits smooth, predictable performance improvements.

Empirics

- Secondary experiment — inducing unobserved emergent abilities
- Task: reconstructing CIFAR100 images using nonlinear autoencoders
    - Usually, accuracy is measured by MSE
    - This paper — use an intentionally discontinuous loss

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

Empirics

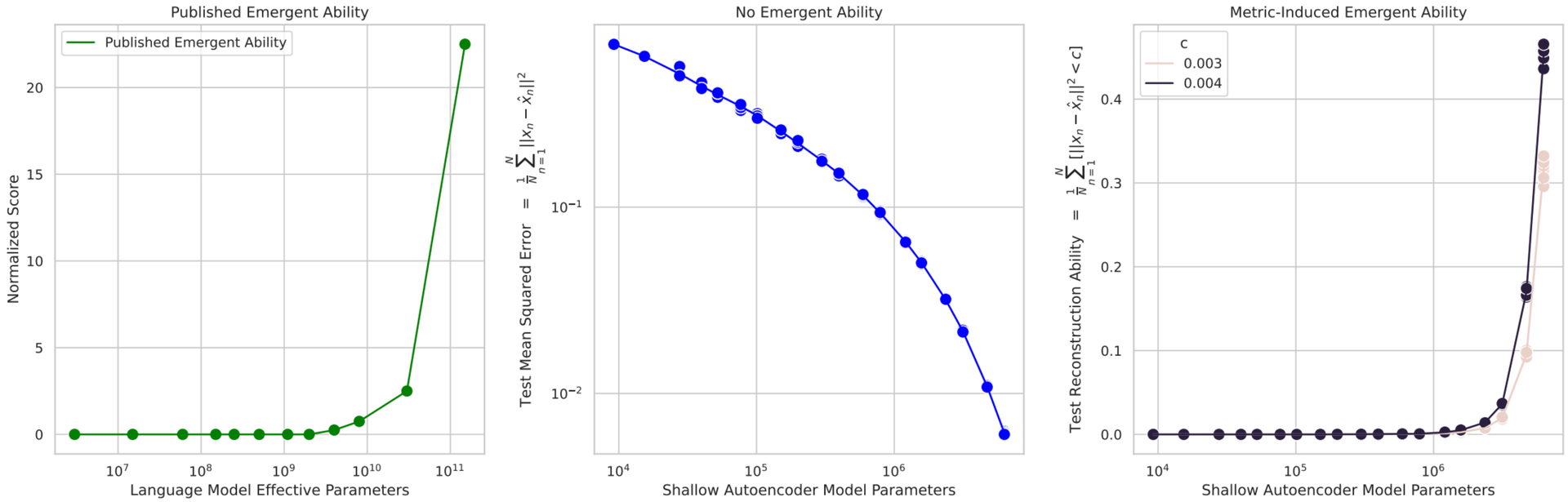


Figure 7: **Induced emergent reconstruction ability in shallow nonlinear autoencoders.** (A) A published emergent ability at the BIG-Bench Periodic Elements task [33]. (B) Shallow nonlinear autoencoders trained on CIFAR100 [21] display smoothly decreasing mean squared reconstruction error. (C) Using a newly defined Reconstruction$_c$ metric (Eqn. 1) induces an unpredictable change.

Should we care?

- Unclear that many capabilities we naturally care about are smooth (or nice) functions of per-token probabilities
    - Some benchmarks really matter
    - RL (?)
- Also, phase transitions are very common in math / stats / TCS / etc!
    - Example — existence of cliques in random graphs
- Also, all preceding discussion requires an underlying scaling law