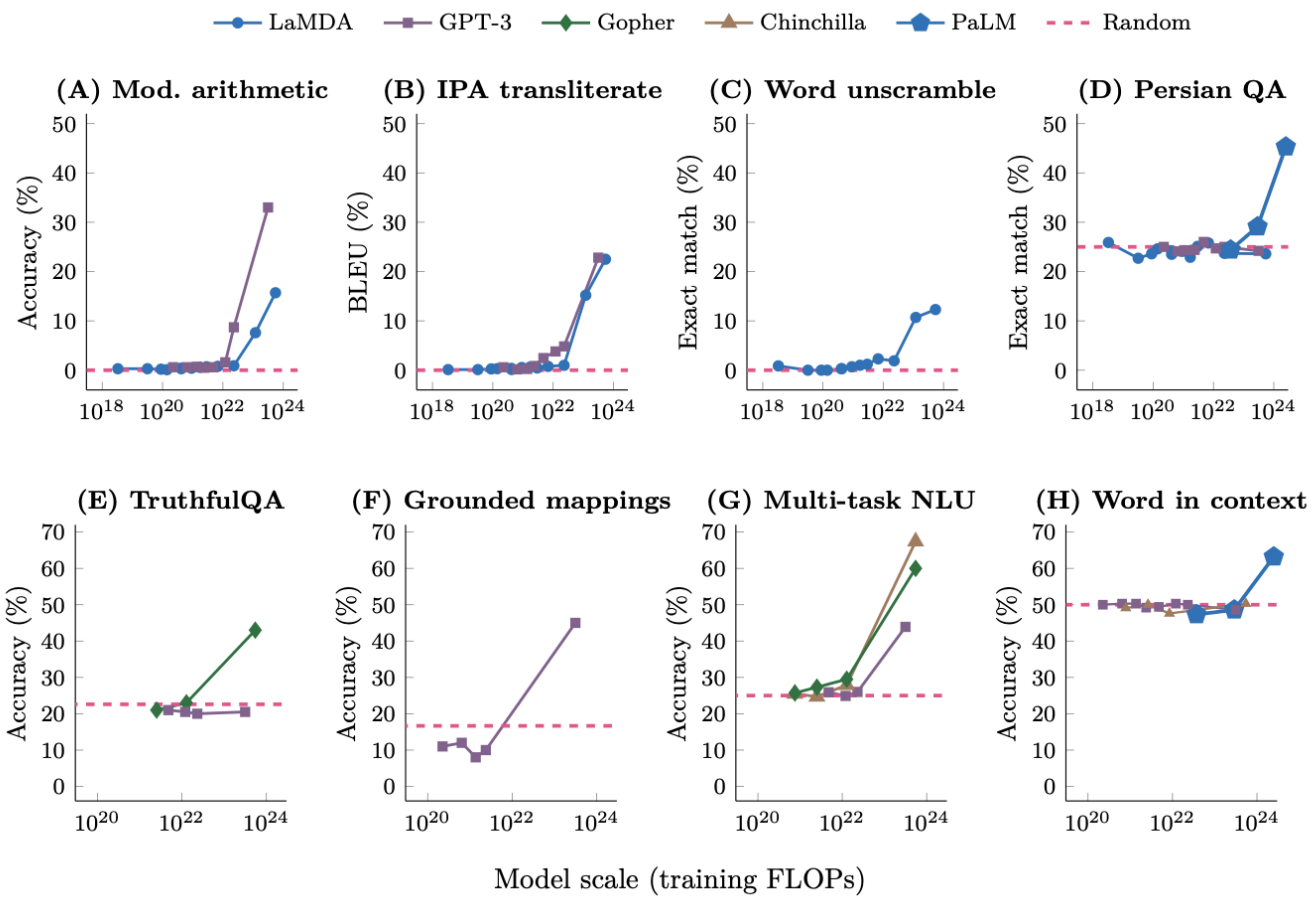
Are emergent abilities of LLMs a mirage?

Dmitry Ligay

Higher School of Economics

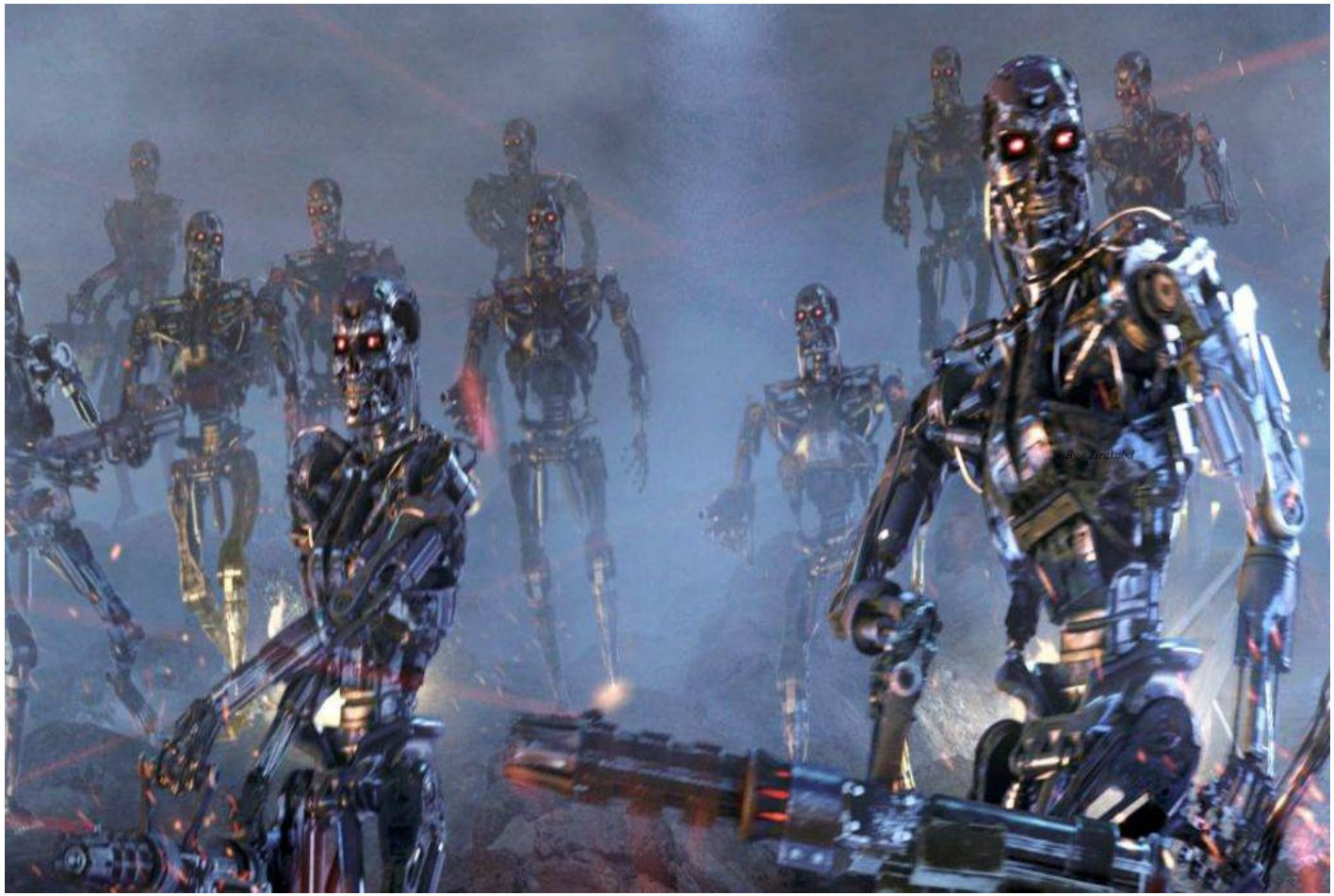
What is emergence?

Abilities not present in smaller models which appear as we increase the model.



Motivation

Emergent abilities are sharp and unpredictable.



Are these abilities emergent?

Emergent abilities appear only on metrics which are nonlinar/discontinuous over per-token error rate.

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

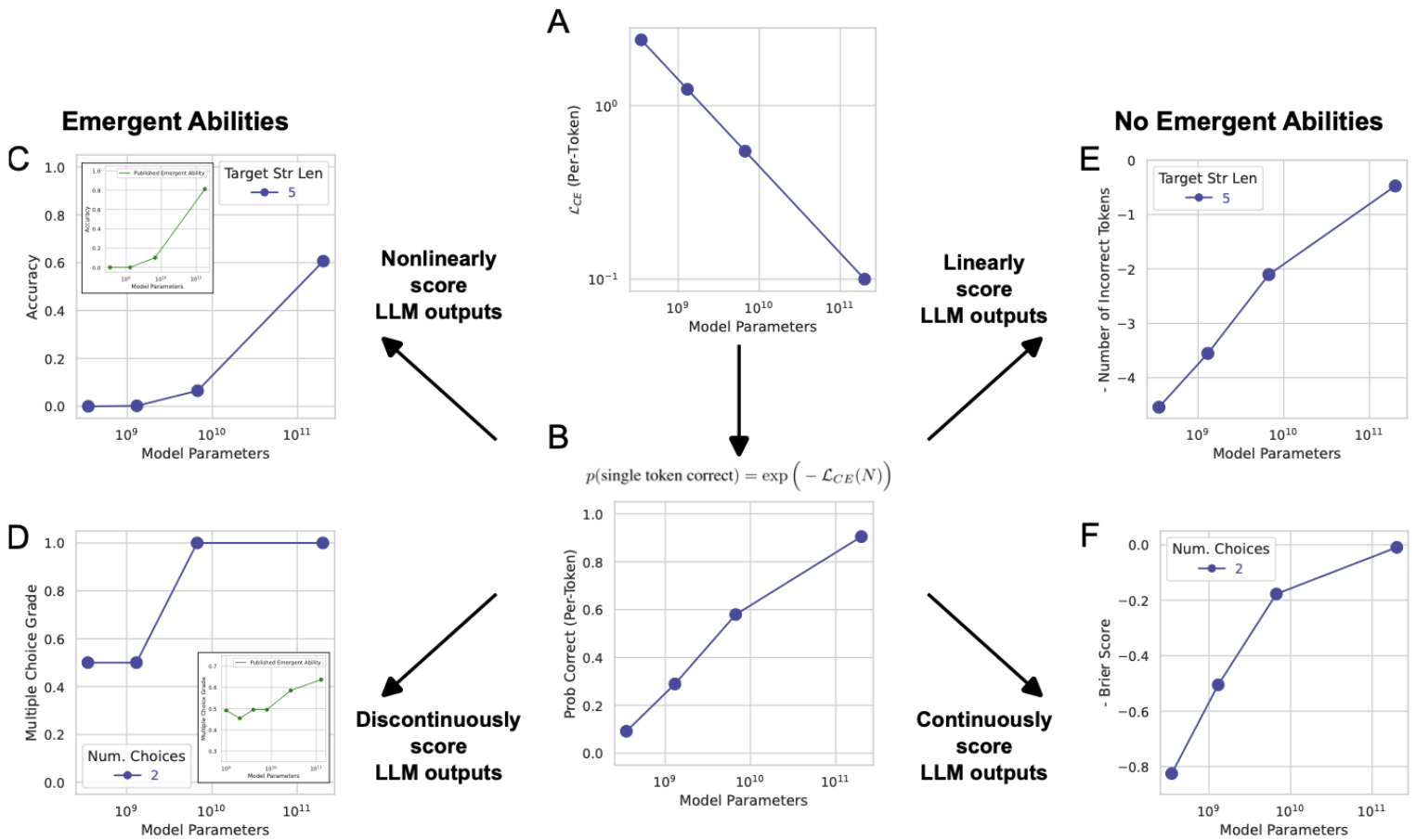
$$\text{Exact String Match} \overset{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$$

Hypothesis

Emergent abilities might be induced because of:

- choice of metric
- too few test data to estimate the performance of smaller models

Choice of metric



Fixing nonlinearity

Suppose that our model guesses the token correctly with probability p. Then (L is the target string length):

$$\text{Accuracy}(L) \approx p^L, \text{Token edit distance}(L) \approx L(1-p)$$

First metric is nonlinear over target sequence length, second doesn't have this issue.

Fixing nonlinearity

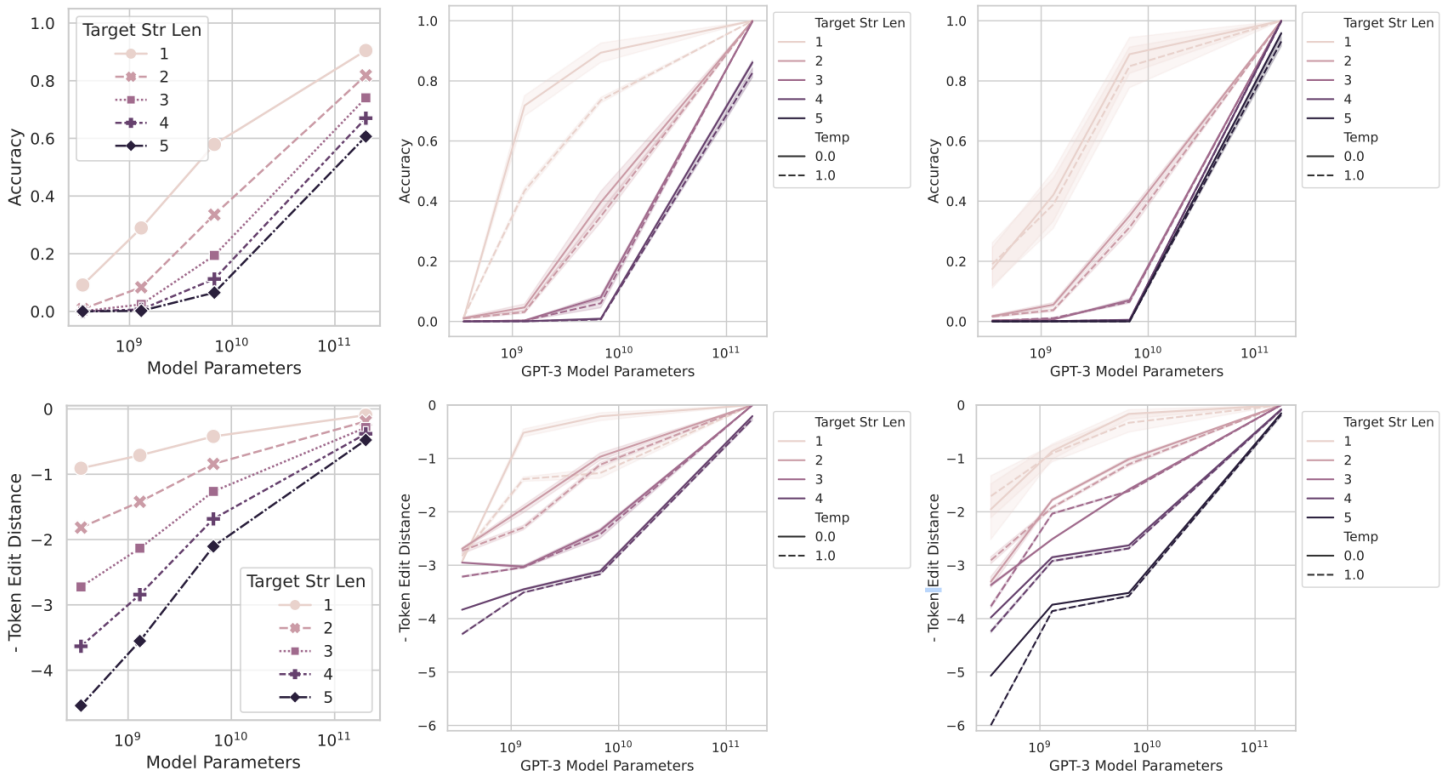


Figure 3: **Claimed emergent abilities evaporate upon changing the metric.** Left to Right: Mathematical Model, 2-Integer 2-Digit Multiplication Task, 2-Integer 4-Digit Addition Task. Top: When performance is measured by a nonlinear metric (e.g., Accuracy), the InstructGPT/GPT-3 [3, 24] family's performance 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.

Increasing test data size

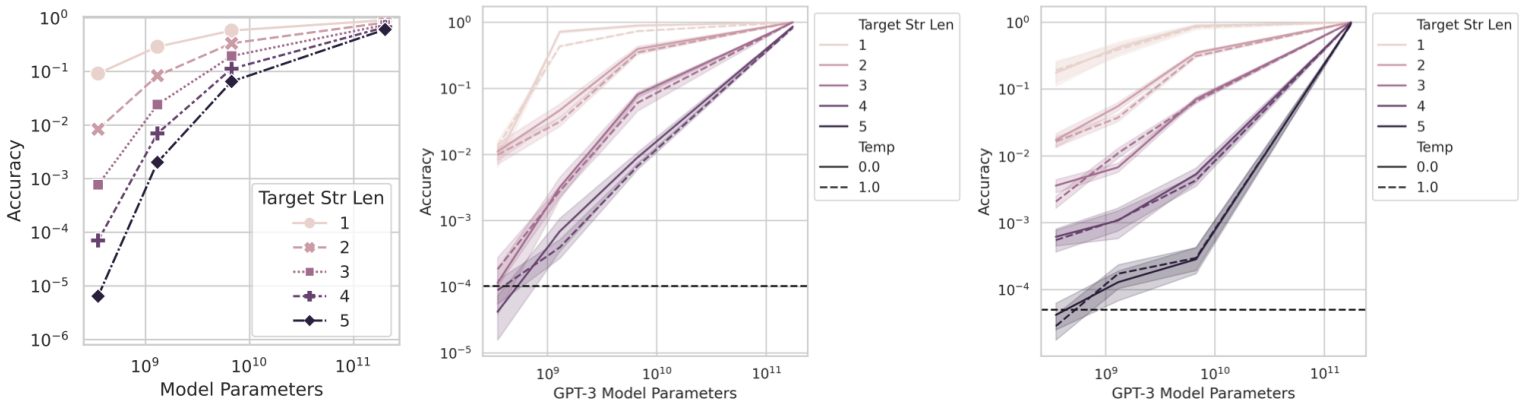


Figure 4: **Claimed emergent abilities evaporate upon using better statistics.** Left to Right: Mathematical Model, 2-Integer 2-Digit Multiplication Task, 2-Integer 4-Digit Addition Task. Based on the predictable effect Accuracy has on performance, measuring performance requires high resolution. Generating additional test data increases the resolution and reveals that even on Accuracy, the InstructGPT/GPT-3 family's [3, 24] performance is above chance and improves in a smooth, continuous, predictable manner that qualitatively matches the mathematical model.

Emergence score

Let x be the model scale, y the model performance judging by the metric. Then:

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

Are there ”emergent” tasks?

Authors have found that there are no tasks (BIG-Bench) which exhibit emergence abilities.
However, there are 4 out of 39 *metrics* which show emergence.

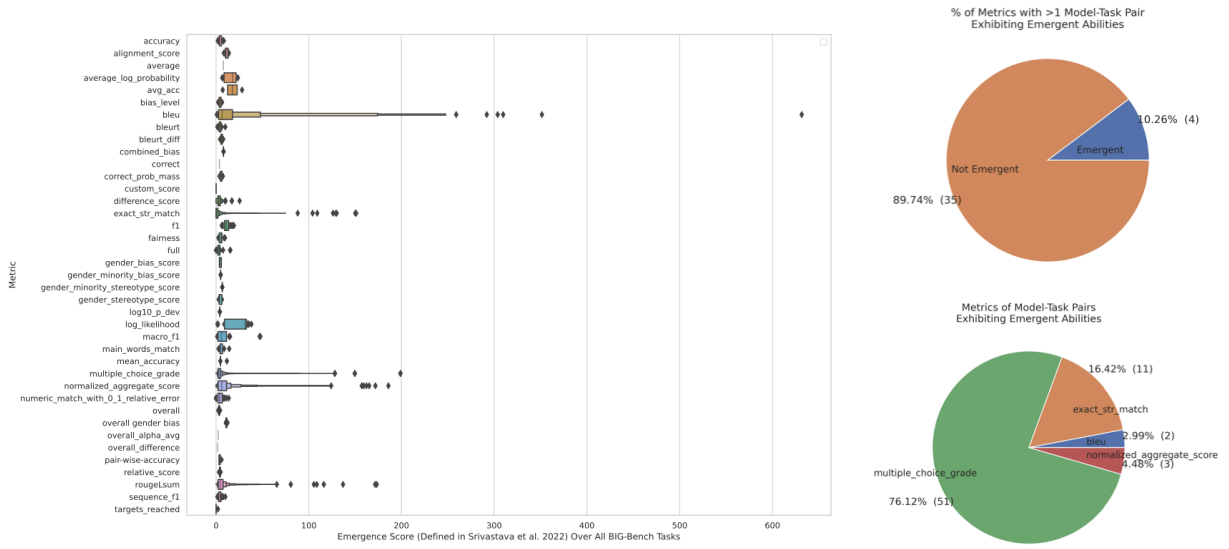


Figure 5: **Emergent abilities appear only for specific metrics, not task-model families.** (A) *Possible* emergent abilities appear with *at most* 5 out of 39 BIG-Bench metrics. (B) Hand-annotated data by [32] reveals emergent abilities appear only under 4 preferred metrics. (C) $> 92\%$ of emergent abilities appear under one of two metrics: Multiple Choice Grade and Exact String Match.

Inducing emergence on computer vision tasks

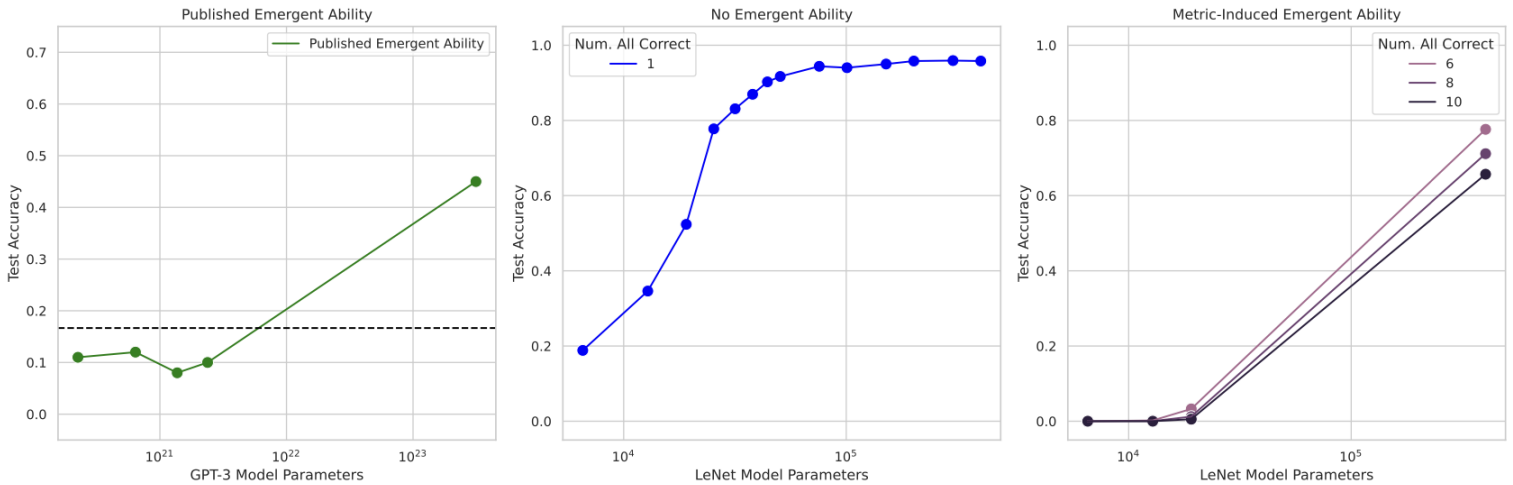


Figure 10: **Induced emergent MNIST classification ability in convolutional networks.** (A) A published emergent ability from the BIG-Bench Grounded Mappings task [33]. (B) LeNet trained on MNIST [21] displays a predictable, commonplace sigmoidal increase in test accuracy as model parameters increase. (C) When accuracy is redefined as correctly classifying K out of K independent test data, this newly defined metric induces a seemingly unpredictable change.

Main takeaways

Currently claimed emergent abilities are not model properties, but metric properties.
That removes a bit of LLMs mystery, and responds to some safety concerns.