Are Emergent Abilities of Large Language Models a Mirage?

Yiran Zhong, Niyati Bafna

What are “emergent” capabilities?

- Defined by the following things:
  - Sharpness: these abilities simply “appear” without warning
  - Unpredictability: we don’t know when they will appear
  - Mysticism: they are part of the LLM dark arts, a miracle of scale
  - Hype: they indicate that the LLM is *fundamentally* different from the LM.

(Wei et al, 2022)

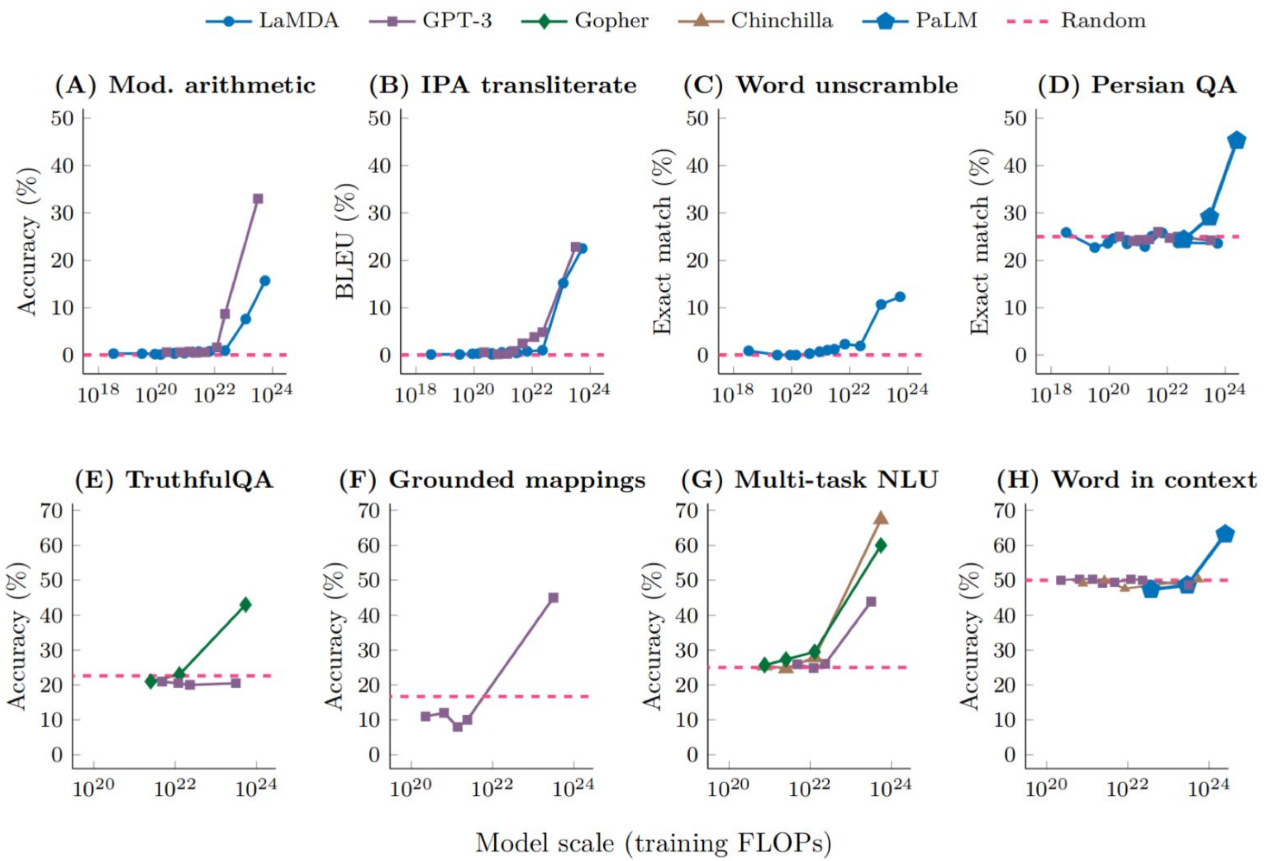


Figure 1: **Emergent abilities of large language models**. Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [33].

But…

- These abilities only show up on tasks measured by metrics with certain properties

→ Claim of paper: emergent properties are a mirage ←

Implying

The LLM is still a familiar creature…

…which we can understand from the little LM. (*phew*)

Culprit properties of evaluation

- “**Non-linear** or **discontinous** metric”

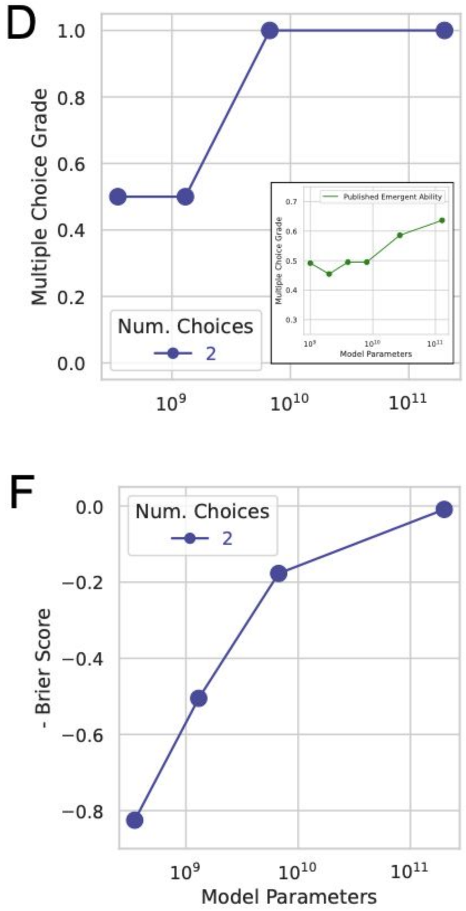
- Bad statistics
- Test dataset **resolution**
- **Insufficient sampling**

Discontinuity

Intuition #1: All-or-nothing evaluation doesn't let you see in-between progress

- Example: Multiple-choice
- Instead, using Brier Score (MSE b/w probs)

- (Note that you can still might argue that a dataset-level metric like a sum of accuracies is continuous...)



Statistics: Resolution and Sampling

Intuition #2: a dataset has an inherent granularity that it lets you evaluate any model at. If the resolution is too high, you will miss things.

- Resolution: 1/size
  - Resolution of n coin flips: (½)^n
- If resolution is too high, existing model performance will be missed

Intuition #3: If the dataset covers a small range of difficulty, you will see all-or-nothing performance given a model

- Sample over a good enough range!
  - And LLMs will show expected behaviour in performance degradation

Non-linearity (IMO: explained by previous insights)

(Stated) Intuition #4: We know that per-token cross-entropy behaves smoothly. If measured metric is non-linear function of length, then long sequences become very hard…

(1) Non-linear: each token needs to be right (p^L)

(2) Linear: number of tokens we got right (L.f(p))

- Why is (2) better than (1)?
  - This can be understood as a **resolution/discontinuity** problem per sample
  - (2) lets us measure in-between progress of getting some tokens right

So…what's the big deal with "linearity in target length"?

- Unclear to me
  - We are not measuring performance along target length, so this seems irrelevant
  - Instead, it seems that this is another way of saying that each test sample should not be too hard, or you will see jumps in performance
  - The authors say that decreasing resolution was a way of fixing this problem, indicating that the real problem here is resolution

  - If somebody understood the argument, please explain it to me!

A little deeper into the linearity argument (1)

- Let's think about
  - $A = p^L$, $B = f(p).L$, where p is the model probability of the right token
  - A, B: probability of scoring 1 given metric (Accuracy, Token edit distance)

- We know that loss scales as a power law of N
  - and that $L_x = - \log p$

- Now, imagine that for small N, p is small, and grows smoothly with N
  - There will come a sudden time when p is sufficient,
  - For B: We see a little jump every time some token p becomes sufficient
  - For A: We'll see a *single, dramatic*, jump when all L p's become sufficient

A little deeper into the linearity argument (2)

- Note that
  - Dependence on p (rather than L) seems more relevant, since p varies with N
  - Both A and B are *smooth* in p and therefore in N
  - But A and B are just prob(metric = 1) , *not* metric = 1
  - Both Accuracy and TED are *unsmooth* in p and therefore in N
  - This is because we do thresholding/maxing per token

- Seems like: if we switched to Brier Score for (non-linear) Accuracy, we would end up with a final smooth metric wrt N despite it being non-linear

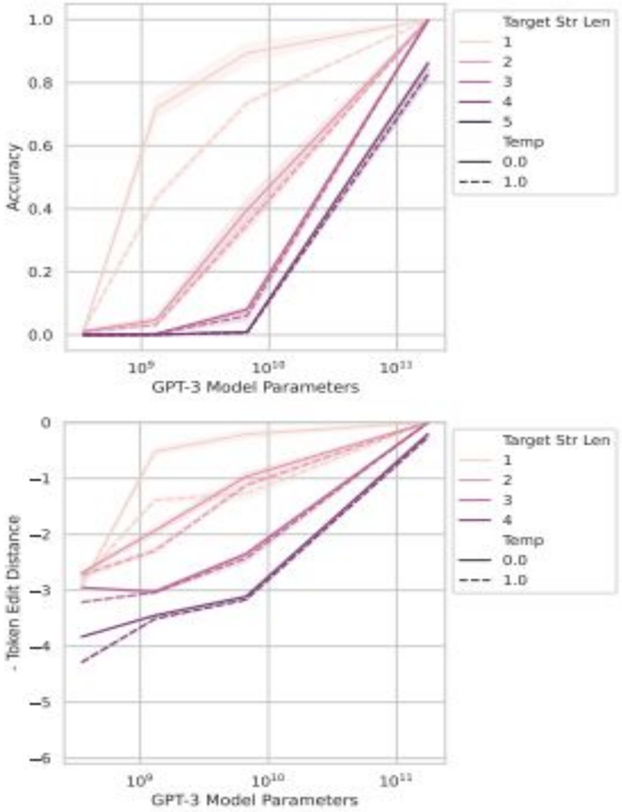
- Non-smoothness is coming from **discontinuity** (which can be thought of as per-sample high **resolution**), not from non-linearity (either in p or L)
  - But when metric is non-linear, non-smoothness is more dramatic

InstructGPT/GPT3's emergent abilities

Changing from a nonlinear/discontinuous metric to a linear/continuous metric reveal smooth improvements in model performance

– Nonlinear metrics, like Accuracy making improvements look sharp and unpredictable

– linear metrics like Token Edit Distance make improvements look more gradual and continuous



InstructGPT/GPT3's emergent abilities (2)

Increasing the resolution of measured model performance by using a larger test dataset should also reveal smooth, continuous improvements

– For nonlinear metrics, a smaller dataset can make the model look like it's suddenly getting better

Longer sequences of input data should lead to predictable changes in model performance

– For accuracy, performance should degrade sharply for longer sequences (geometrically).

– For token edit distance, performance should degrade more smoothly (quasilinearly).

Prediction 1

When switching from a nonlinear to a linear metric, the sudden appearance of abilities vanishes, and performance improvement becomes gradual and predictable.

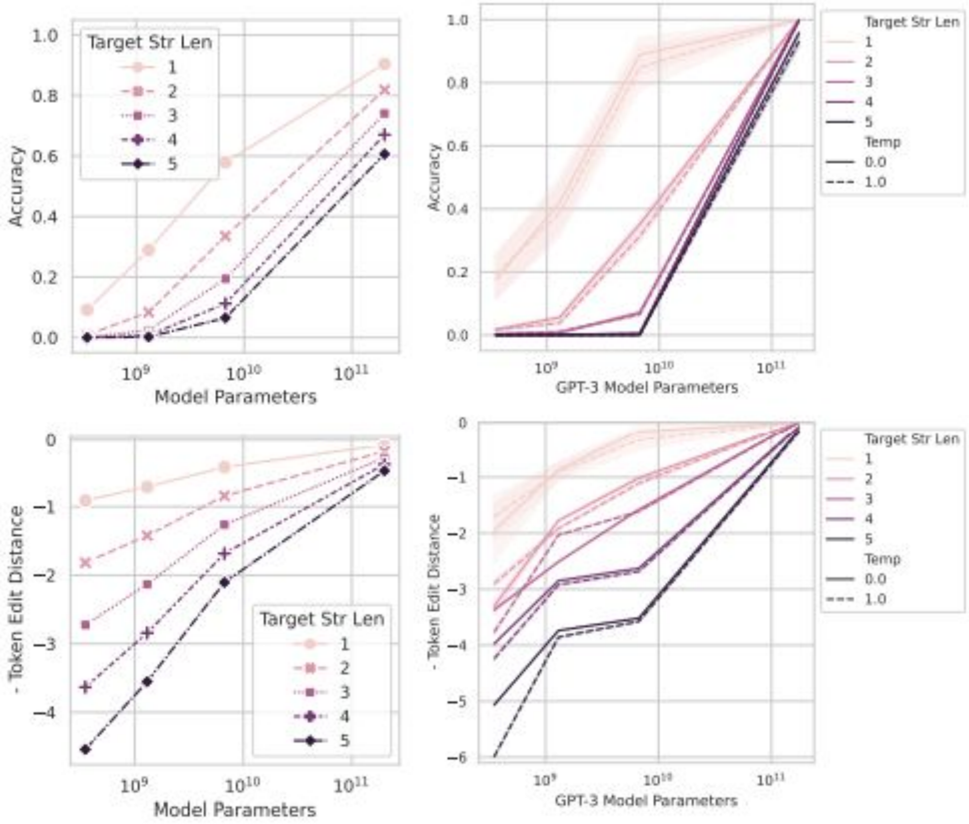
– Top rows shows the model's performance using Accuracy, a nonlinear metric

(performance appears to suddenly improve when dealing with longer input sequences)

– Bottom row shows the model’s performance using Token Edit Distance, a linear metric

(performance improvement is much smoother and more gradual)

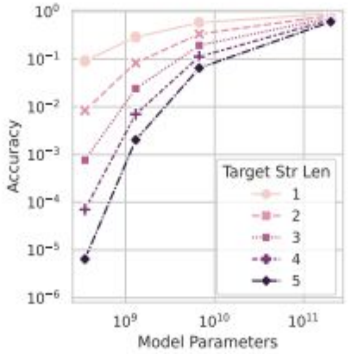
emergent abilities can be an illusion caused by using a sharp, nonlinear metric

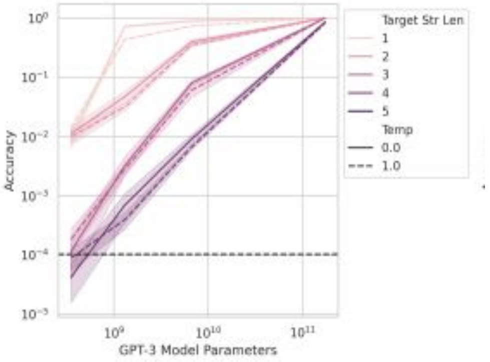


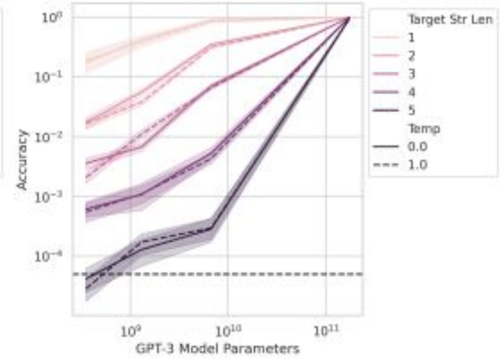
Prediction 2

using more test data, the emergent abilities smooth out, showing that smaller models are not entirely incapable of solving tasks—they just appear to be due to low-resolution measurements.

– further testing with more data (resolution), confirming that, even with accuracy as a metric, the model's improvements follow a smoother, more predictable path.







Meta-Analysis

Use **BIG-Bench**, a collection of benchmark tasks used to evaluate language models.

Key prediction:

Emergent abilities should appear predominantly on specific metrics, not task-model family pairs.

Changing the metric from a nonlinear/discontinuous metric to a linear/continuous one should remove the emergent ability.

Emerging score

Y is model performance at scale x

Numerator: the difference between the best and worst performance scores.

Denominator: how gradual the performance improvements are over model scale.

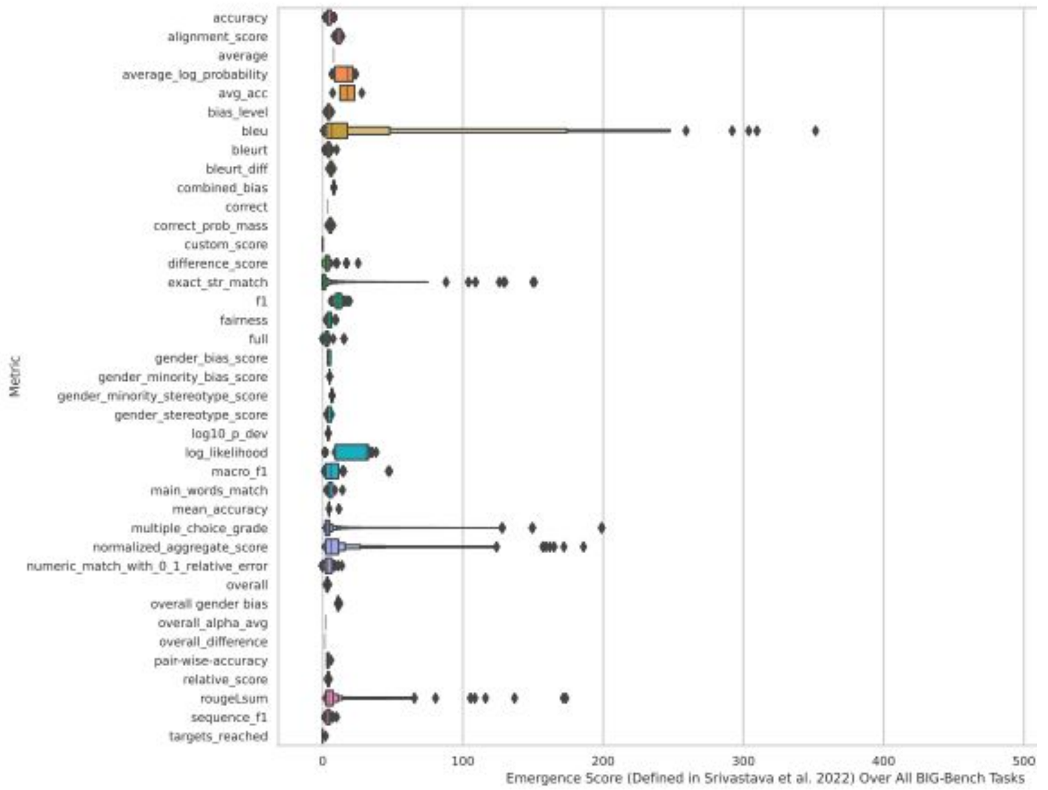
higher score indicates sharper, less gradual changes in performance, suggesting the presence of an emergent ability

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

Emergent abilities appear only under specific metrics

Of the 39 preferred metrics in BIG-Bench, c
5 showed any evidence of emergent abilitie

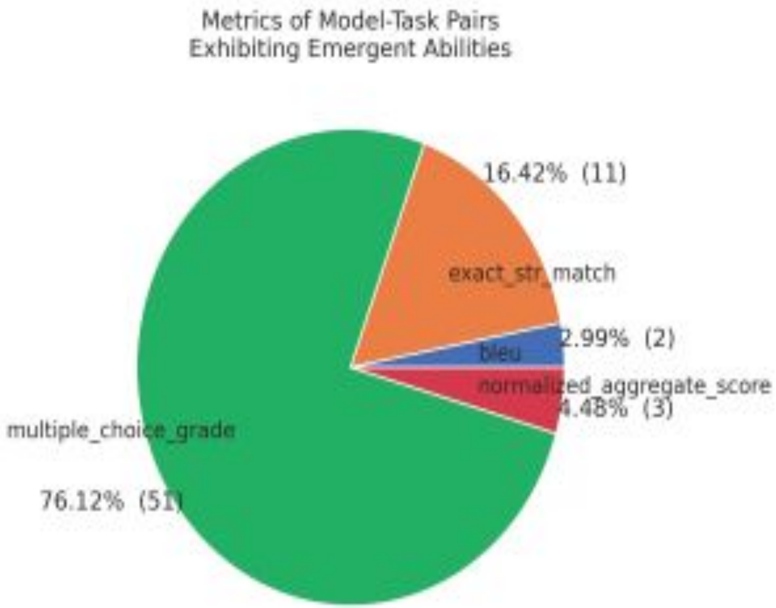
5 metrics that did show emergent abilities were primarily nonlinear and/or discontinuous metrics



Two metrics account for over 92% of emergent abilities

Multiple Choice Grade (a discontinuous metric)

Exact String Match (a nonlinear metric)



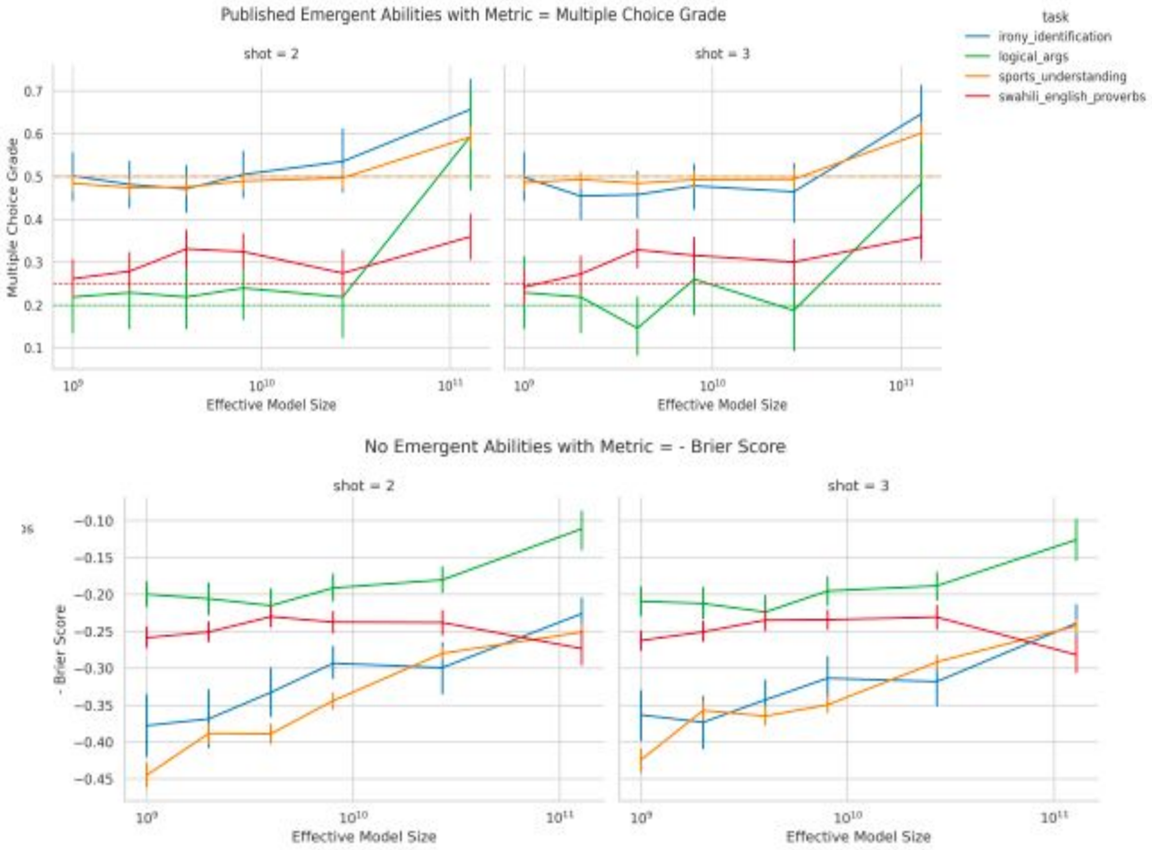
Changing metric

focusing on the **LaMDA model** family

exhibited emergent abilities under the **Multiple Choice Grade**

evaluation metric was switched to a continuous one—**Brier Score**

Finding: emergent abilities are more smooth when switching to a continuous metric



Emergent ability on vision task

emergent abilities can be **artificially induced** in neural networks for vision tasks by manipulating the metric

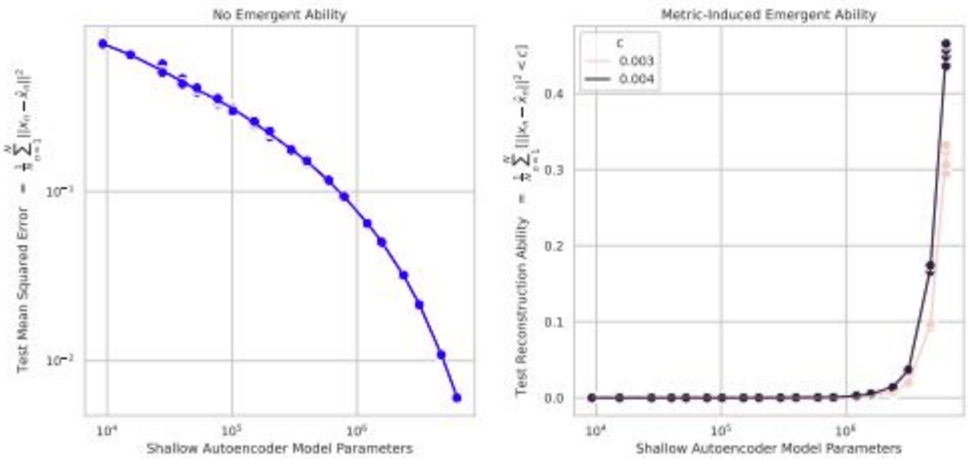
1: reconstruct emergent in autoencoder using CIFAR100 dataset

2: induce emergent abilities in autoregressive transformers trained to classify handwritten characters from the Omniglot dataset.

Experiment 1

Left plot: under the traditional continuous metric (mean squared error), the autoencoder's reconstruction error decrease smoothly as the model size increases.

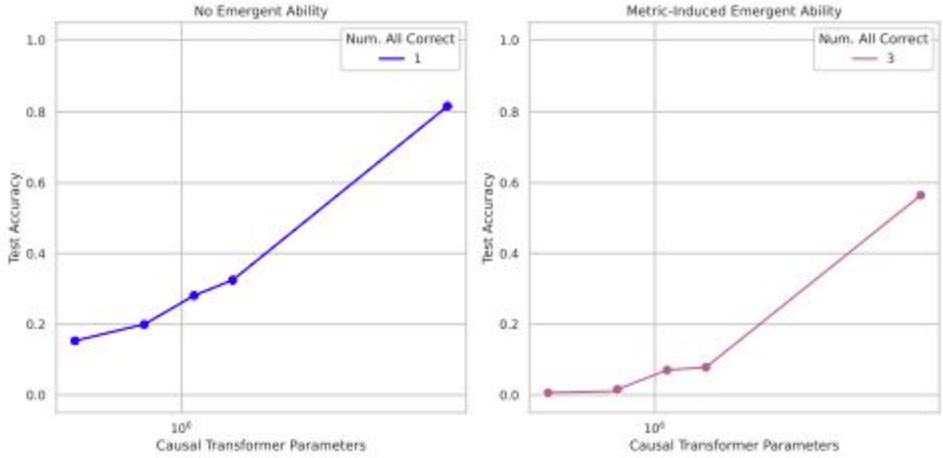
Right plot: When newly defined Reconstruction metric is applied, a sharp and seemingly unpredictable emergent ability is induced



Experiment 2

Left plot: When Autoregressive transformers trained to classify Omniglot images,accuracy improves smoothly as the model size increases.

Right plot: When accuracy is redefined as classifying all images correctly (subset accuracy), a seemingly emergent ability appears, with sharp improvements as model size increases. This further demonstrates that emergent abilities can be induced by using stricter metrics.



Artifacts or Abduction: How Do LLMs Answer MCQs without the Question?

- Yes this is a thing
- LLMs are able to do well on MCQ benchmarks without the question

- Paper looks at
- Memorization: show that the models haven’t simply memorized the benchmark
- Priors: Correct answer text is not inherently more probable than others
- Choice dynamics and question inference: This is largely what’s happening

Takeaway: LLM evaluation on MCQA benchmarks needs to be further investigated - what is it actually learning?