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”
  - 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.

Emergent Abilities of Large Language Models (Wei et al, 2022)

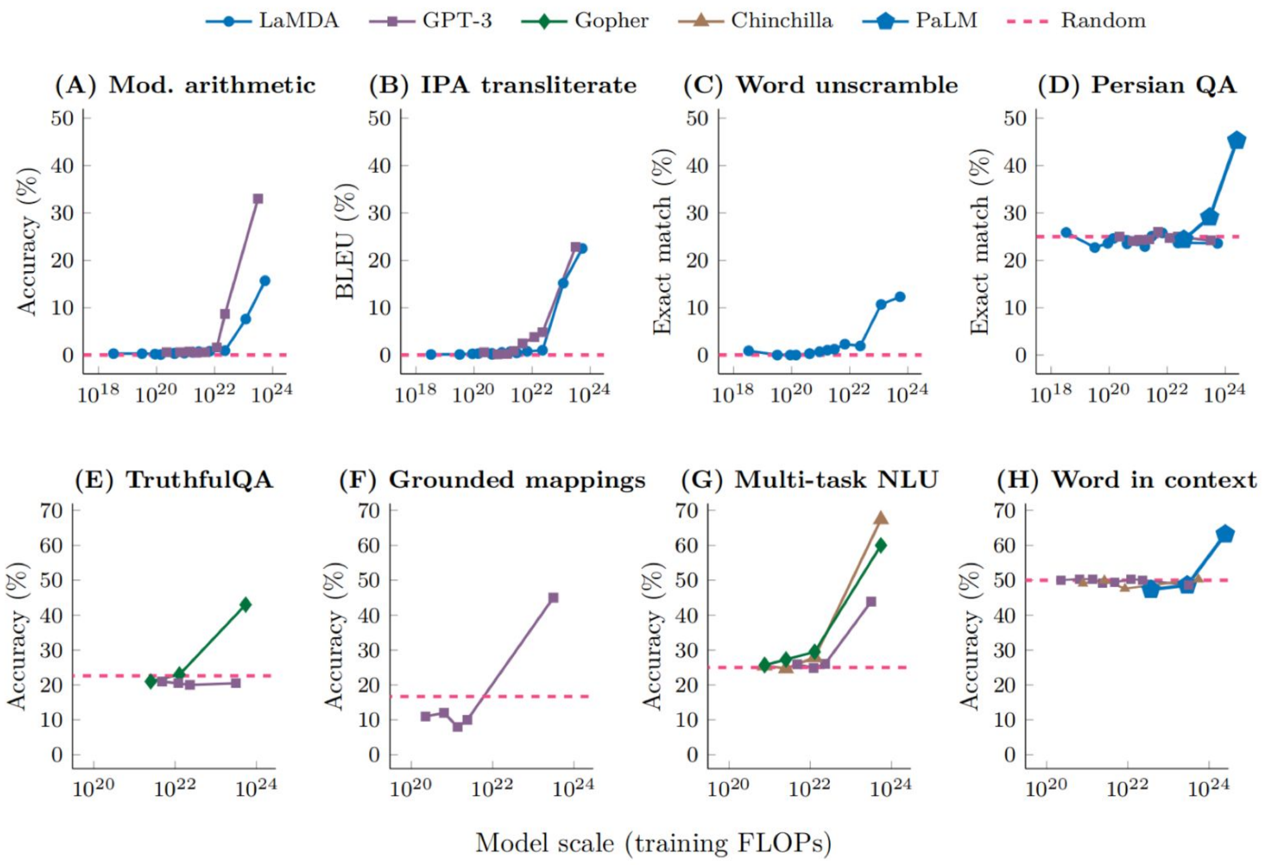


Figure 1: **Emergent abilities of large language models**. Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [33].

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 by extrapolating from little LMs. (*phew*)

Issues with Evaluation

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

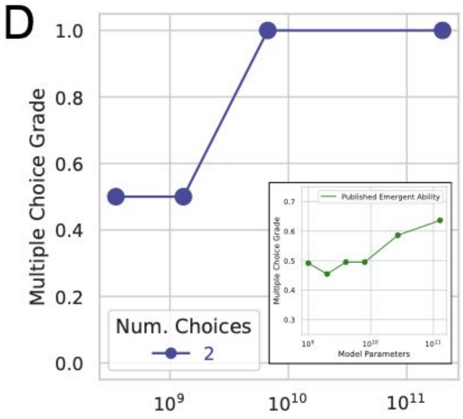
- Bad statistics
  - Test dataset **resolution**
  - **Insufficient sampling**

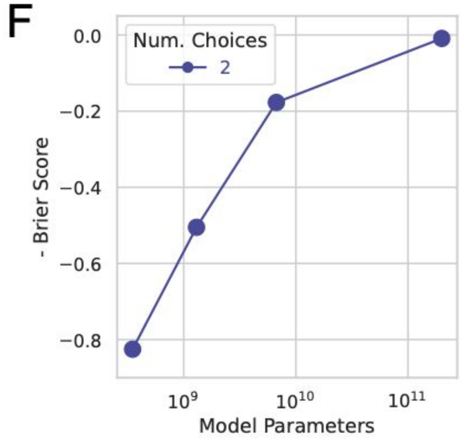
Discontinuity of Metrics

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)

- Our note: you 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

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...

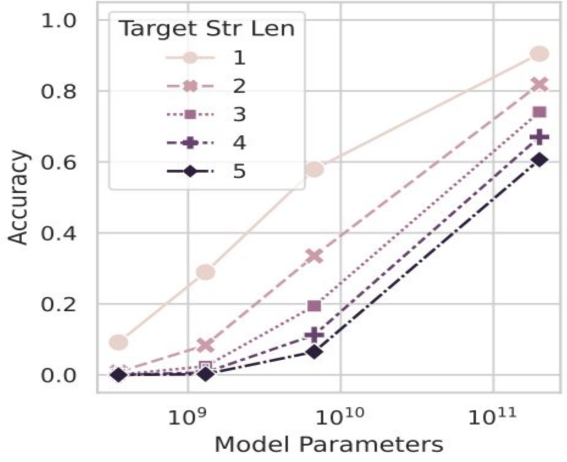
(1) Non-linear: each token needs to be right (p^L)

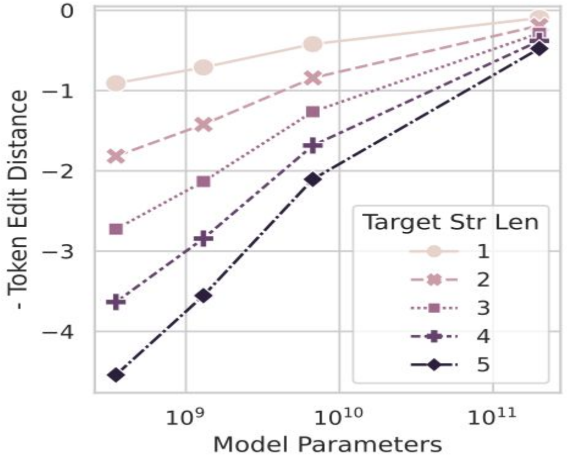
(2) Linear: number of tokens we got right ($L.f(p)$) where f is a linear function

- 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

Linearity in target length

- Note: We are not measuring performance along target length
    - But: when L is high, and you don't have good sampling along the curve, the (actually smooth) curve will appear unsmooth
    - The authors say that decreasing **resolution** was a way of fixing this problem, indicating that this is only a problem with bad resolution





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)

- 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
  - Something like C = L.p^10 is as bad as B

- 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

- 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

InstructGPT/GPT3's emergent abilities

Testing whether the "emergent abilities" of GPT-3 models in arithmetic tasks are real or just an illusion caused by the metrics used.

GPT3: publicly queryable model

Task:

multiplication between two 2-digit integers (e.g., 45 * 67)

addition between two 4-digit integers (e.g., 1234 + 5678)

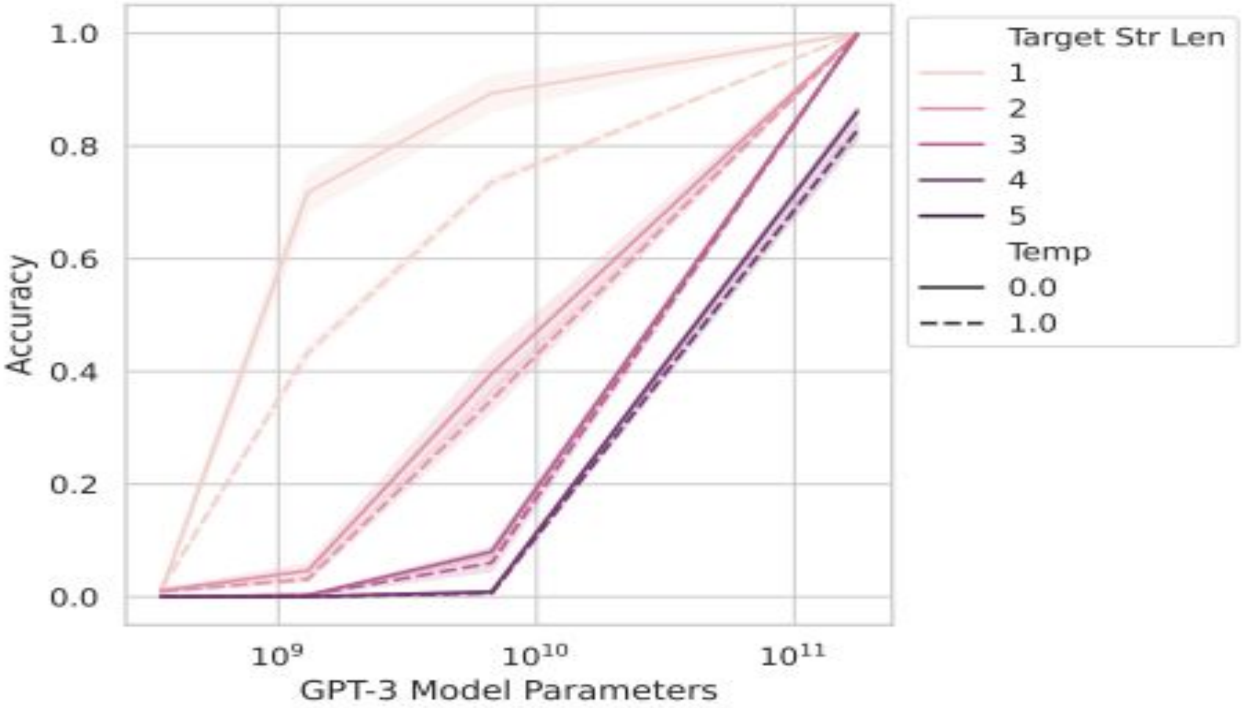
Prediction 1

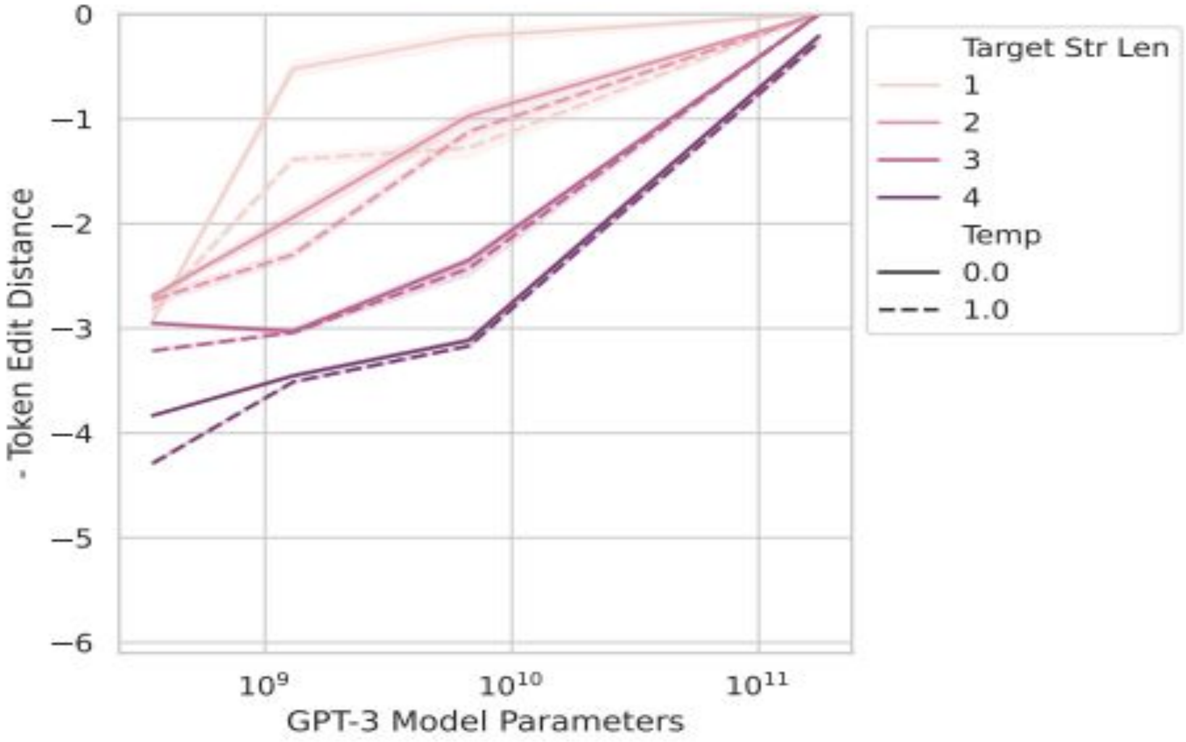
Changing from a nonlinear/discontinuous metric to a linear/continuous metric reveal smooth improvements in model performance

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

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

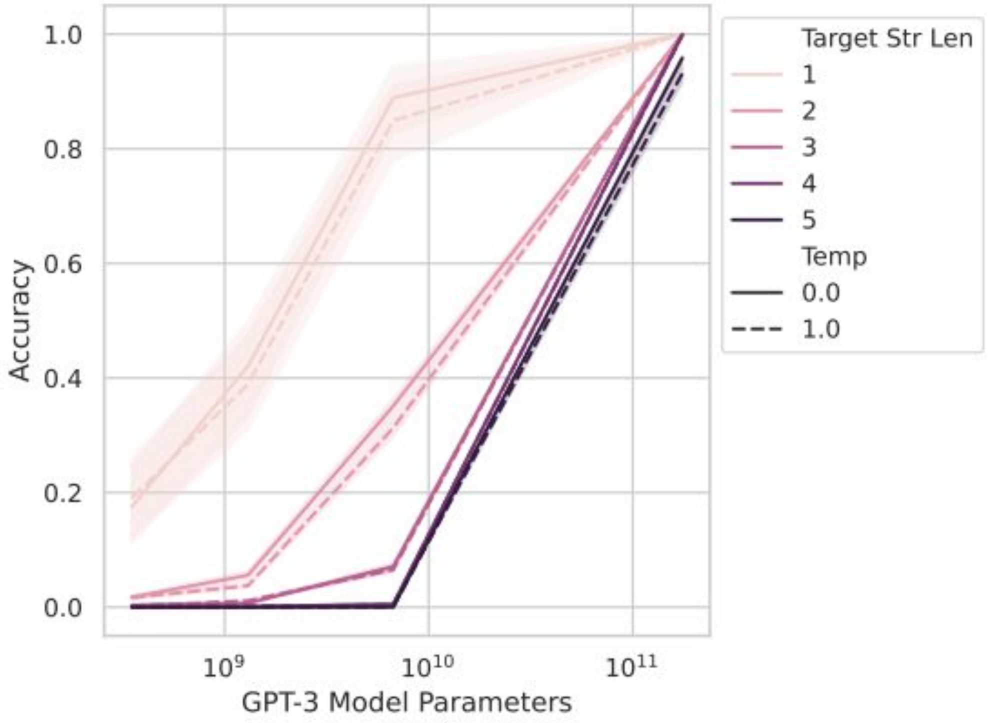
2-digit multiplication

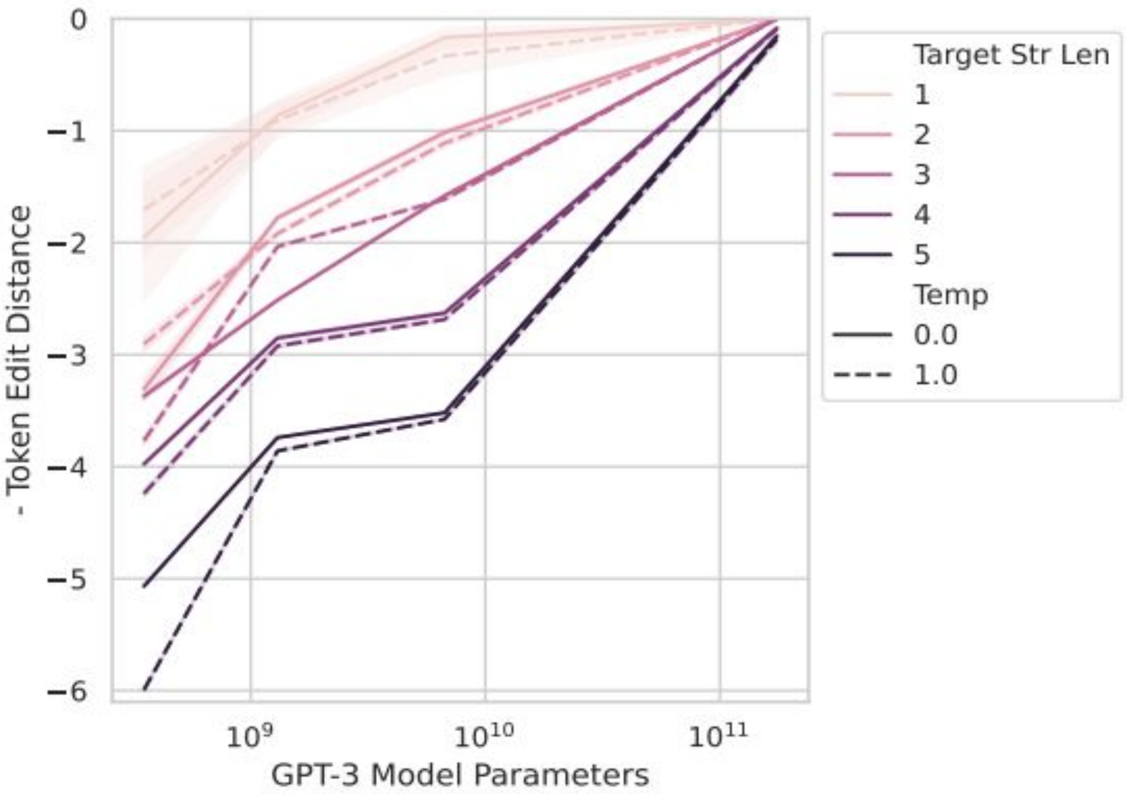




0
−1
−2
−3
−4
−5
−6
- Token Edit Distance
10^9
10^{10}
10^{11}
GPT-3 Model Parameters
Target Str Len
1
2
3
4
Temp
0.0
1.0

4 digit addition task





0
−1
−2
−3
−4
−5
−6
- Token Edit Distance
10^9
10^{10}
10^{11}
GPT-3 Model Parameters
Target Str Len
1
2
3
4
5
Temp
0.0
1.0

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

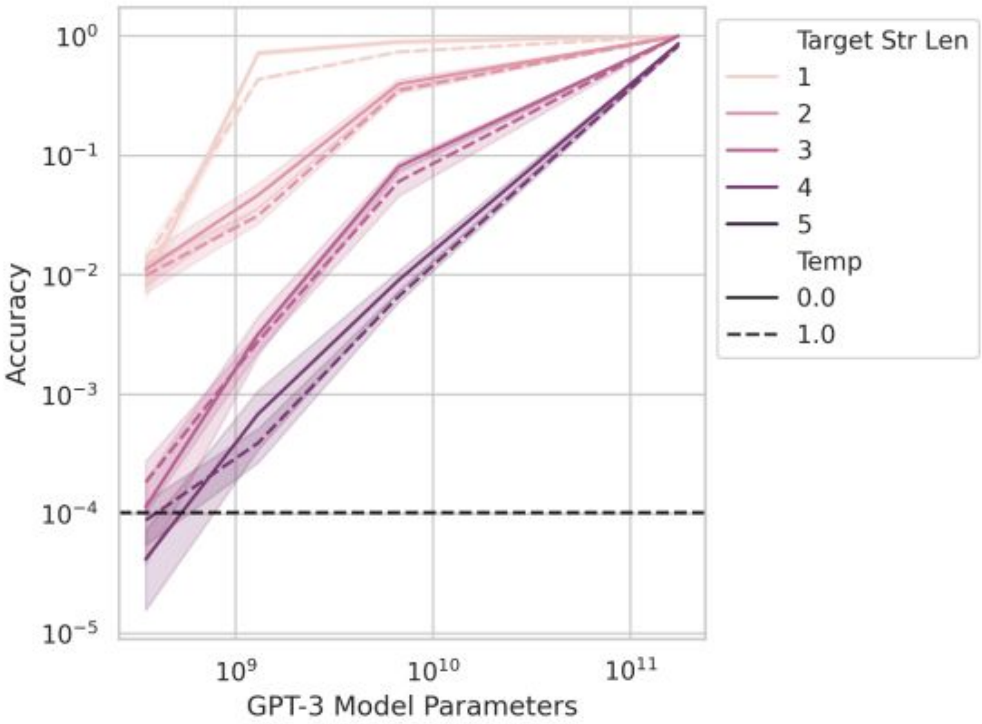
– For nonlinear metrics, a smaller dataset can make the model look like it's suddenly getting better

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

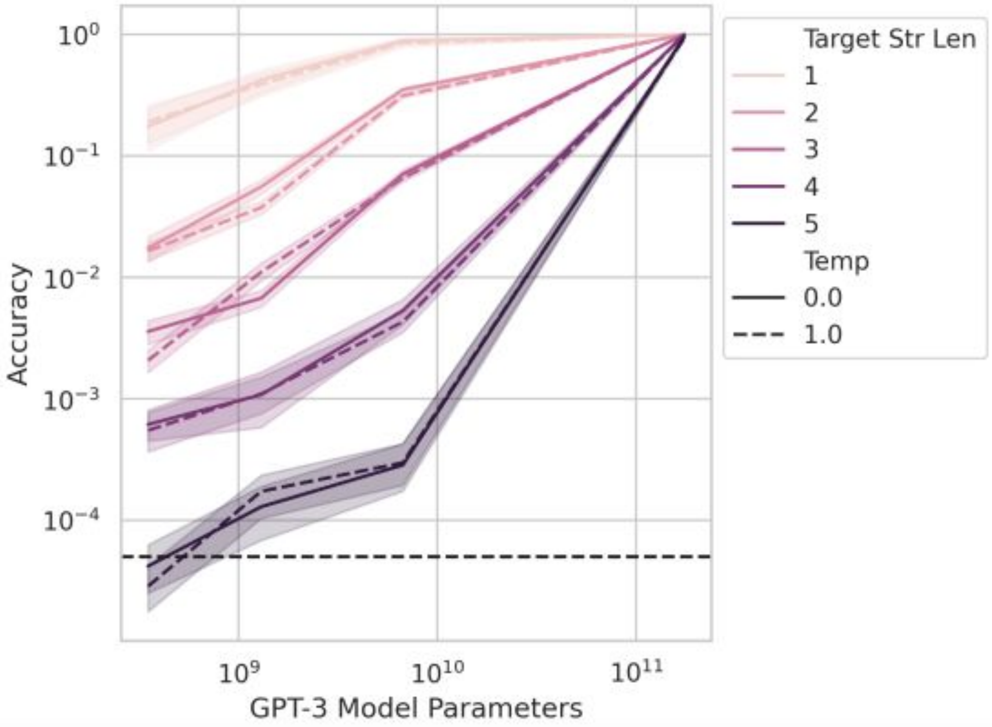
– For accuracy, performance should degrade sharply for longer sequences (geometrically).

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

2-digit multiplication with more test data



4-digit addition with more test data



Meta-Analysis

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

predictions are extended to a broader range of models and tasks. Further studying on whether emergent abilities are dependent on specific metrics, not on task-model family pairs

Emerging score

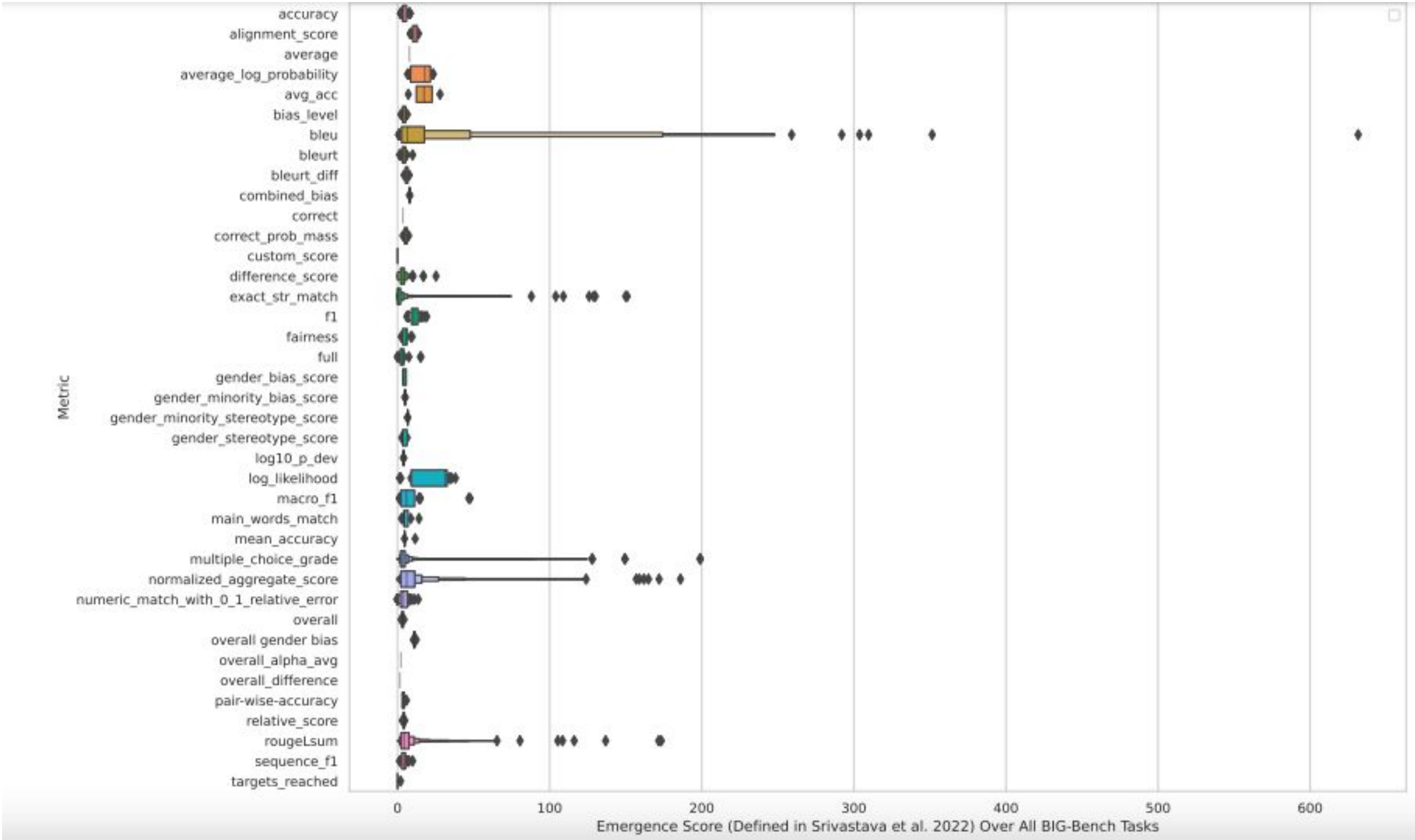
Y is model performance at scale x

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

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


accuracy
alignment_score
average
average_log_probability
avg_acc
bias_level
bleu
bleurt
bleurt_diff
combined_bias
correct
correct_prob_mass
custom_score
difference_score
exact_str_match
f1
fairness
full
gender_bias_score
gender_minority_bias_score
gender_minority_stereotype_score
gender_stereotype_score
log10_p_dev
log_likelihood
macro_f1
main_words_match
mean_accuracy
multiple_choice_grade
normalized_aggregate_score
numeric_match_with_0_1_relative_error
overall
overall gender bias
overall_alpha_avg
overall_difference
pair-wise-accuracy
relative_score
rougeLsum
sequence_f1
targets_reached
Metric
0
100
200
300
400
500
600
Emergence Score (Defined in Srivastava et al. 2022) Over All BIG-Bench Tasks

Emergent abilities appear only under specific metrics

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

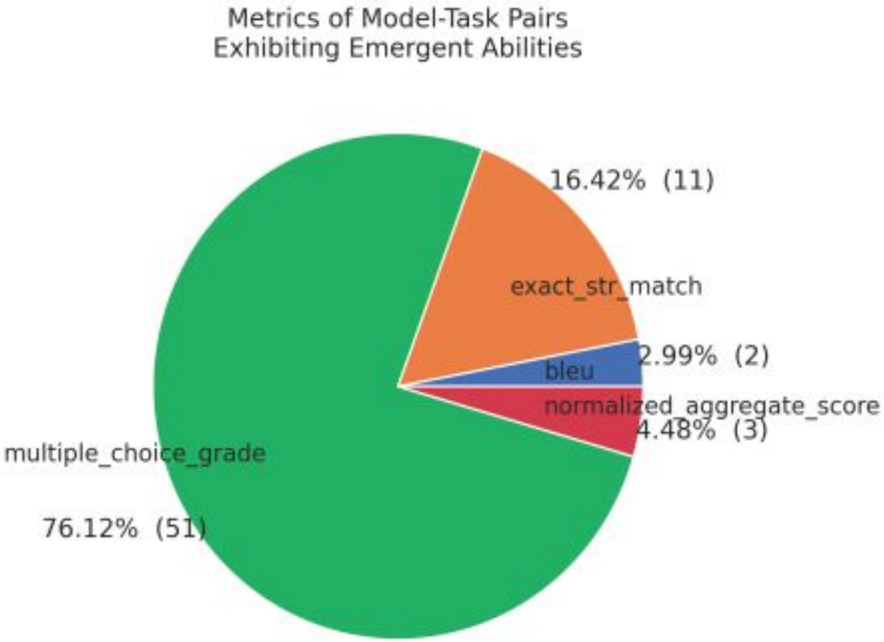
5 metrics that did show emergent abilities were primarily nonlinear and/or discontinuous metrics

– Multiple Choice Grade and Exact String Match. Multiple Choice Grade is discontinuous, and Exact String Match is nonlinear

Two metrics account for over 92% of emergent abilities

Multiple Choice Grade (a discontinuous metric)

Exact String Match (a nonlinear metric)



Try different model and metrics

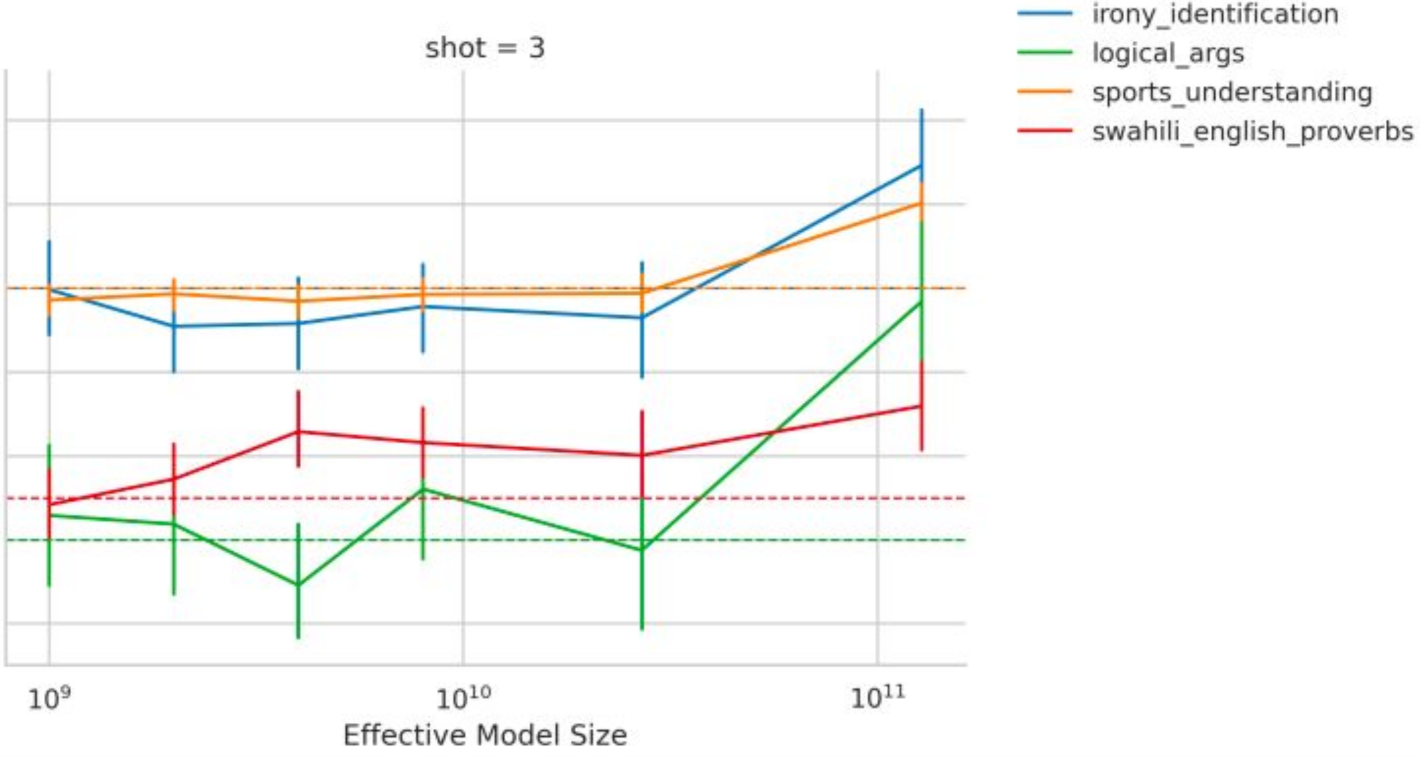
focusing on the **LaMDA model** family

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

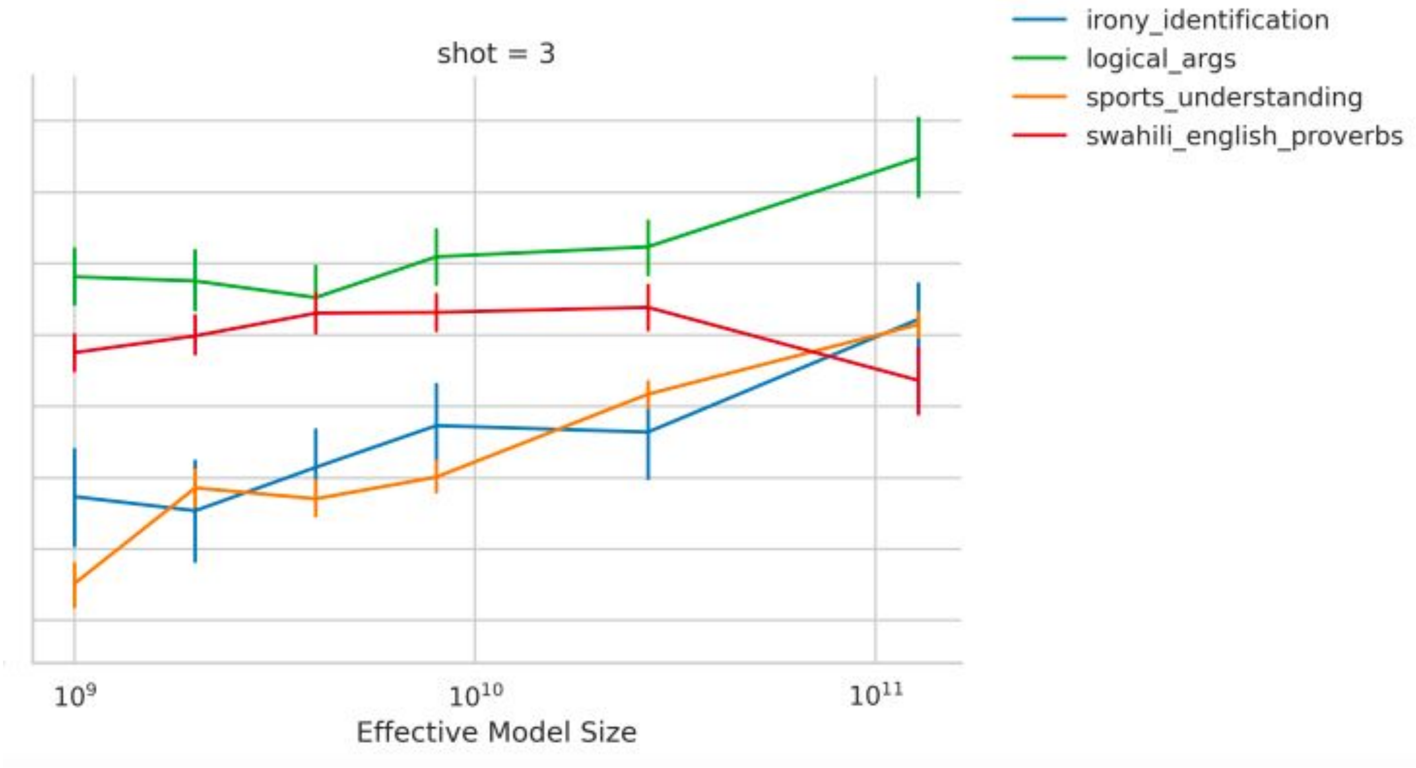
evaluation metric was switched to a continuous one—**Brier Score**

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

Multiple choice grade



Brier score


shot = 3
irony_identification
logical_args
sports_understanding
swahili_english_proverbs
10^9
10^{10}
10^{11}
Effective Model Size

For the Brier score the value is negative, I’m not sure why

Emergent ability on vision task

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

1: define a discontinuous metric that measures a network's ability to reconstruct CIFAR10 images

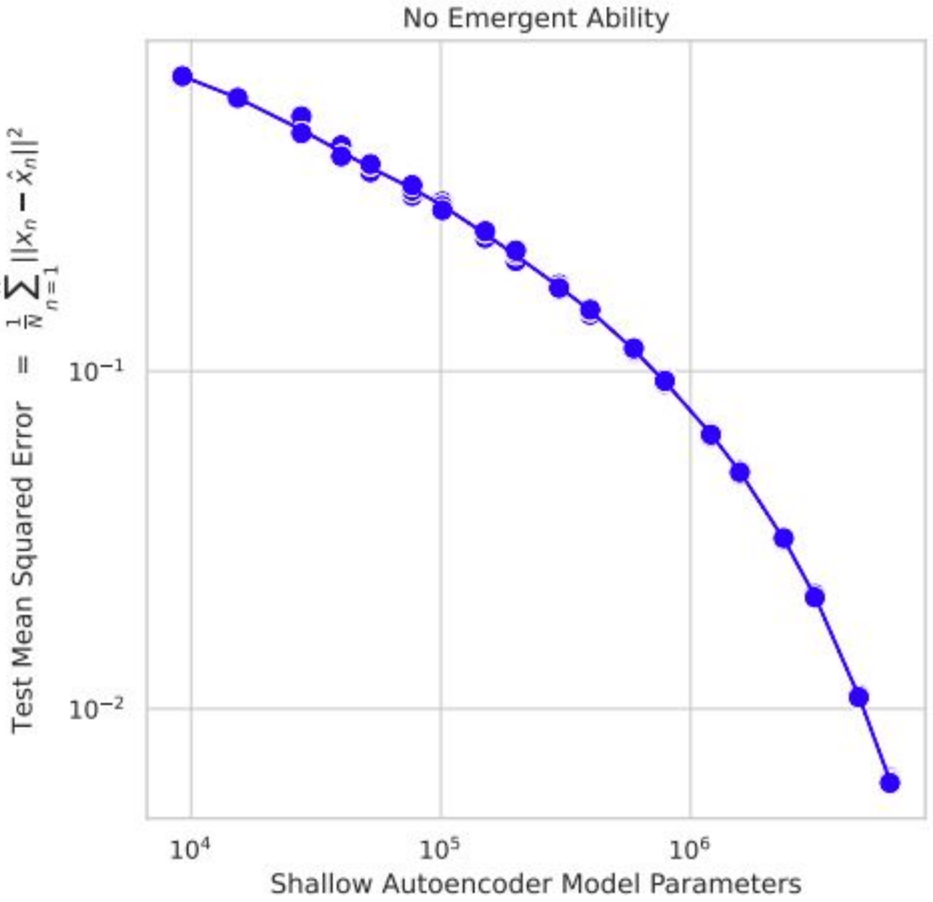
2: induce emergent abilities in autoregressive transformers trained to classify handwritten characters from the Omniglot dataset.

Deep network reconstruct image

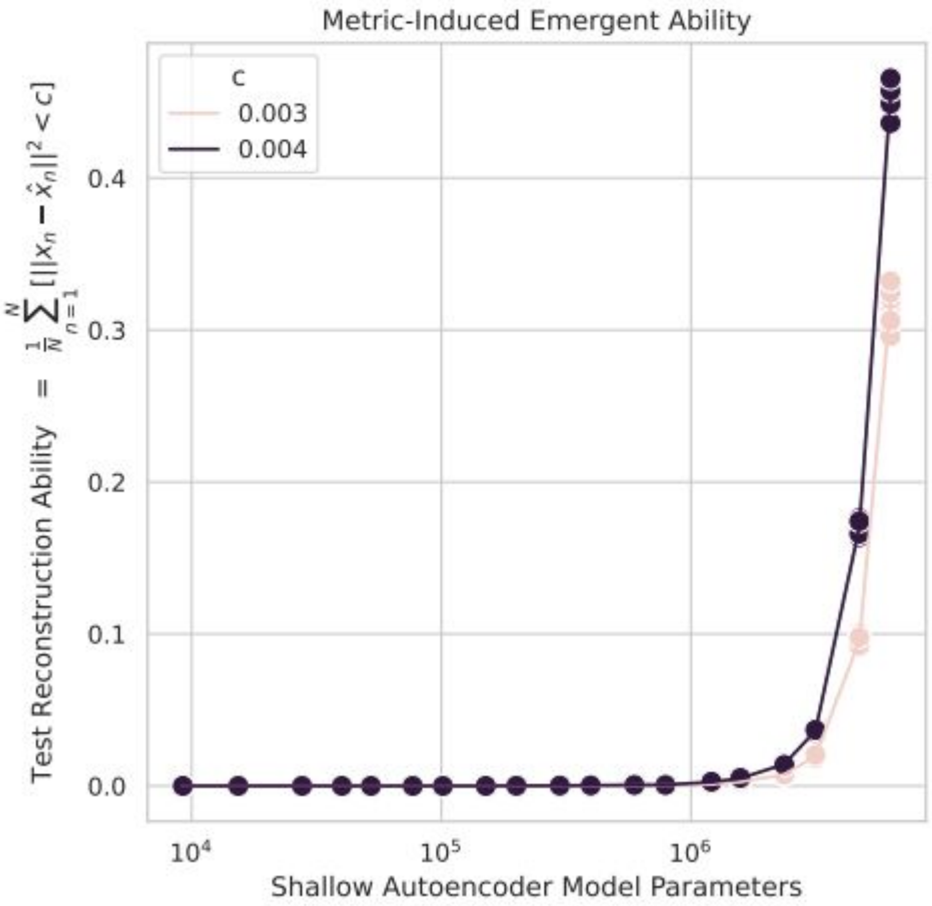
$$\text{Reconstruction}_c\left(\{x_n\}_{n=1}^{N}\right) \stackrel{\text{def}}{=} \frac{1}{N}\sum_n \mathbb{I}\left[||x_n - \hat{x}_n||^2 < c\right]$$

Reconstruction metric:

X_n is origin image, x^hat is reconstructed image, c is threshold value introduce the emergent ability by setting boundary between acceptable reconstruction and unacceptable one.



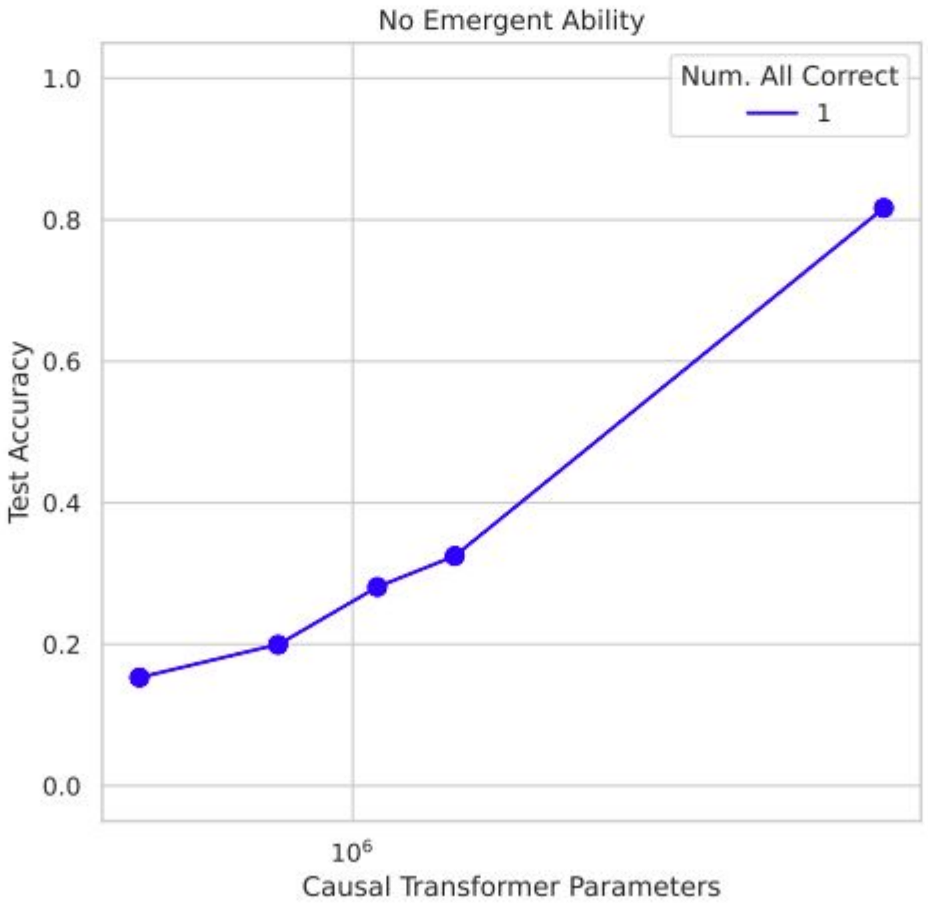
No Emergent Ability
Test Mean Squared Error $= \frac{1}{N}\sum_{n=1}^{N} ||x_n - \hat{x}_n||^2$
10^{-1}
10^{-2}
10^4
10^5
10^6
Shallow Autoencoder Model Parameters

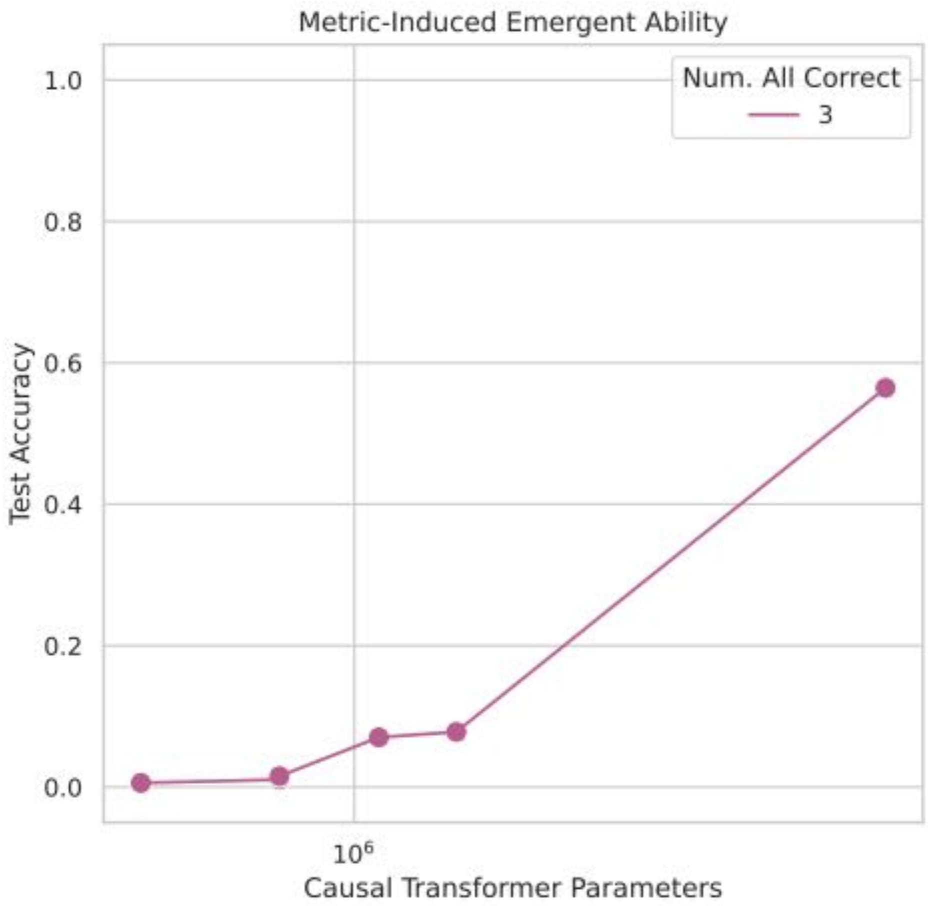


Metric-Induced Emergent Ability
c
0.003
0.004
Test Reconstruction Ability $= \frac{1}{N}\sum_{n=1}^{N} [||x_n - \hat{x}_n||^2 < c]$
0.0
0.1
0.2
0.3
0.4
10^4
10^5
10^6
Shallow Autoencoder Model Parameters

Classify Omniglot handwriting characters

Metrics: Accuracy (gives the model a score of 1 only if it correctly classifies all characters in a sequence, If the model makes even one mistake, it gets 0.)


No Emergent Ability
Num. All Correct
1
1.0
0.8
0.6
0.4
0.2
0.0
Test Accuracy
10^6
Causal Transformer Parameters


Metric-Induced Emergent Ability
Num. All Correct
3
1.0
0.8
0.6
0.4
0.2
0.0
Test Accuracy
10^6
Causal Transformer Parameters

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?