Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Computer Science, Stanford University

Abstract

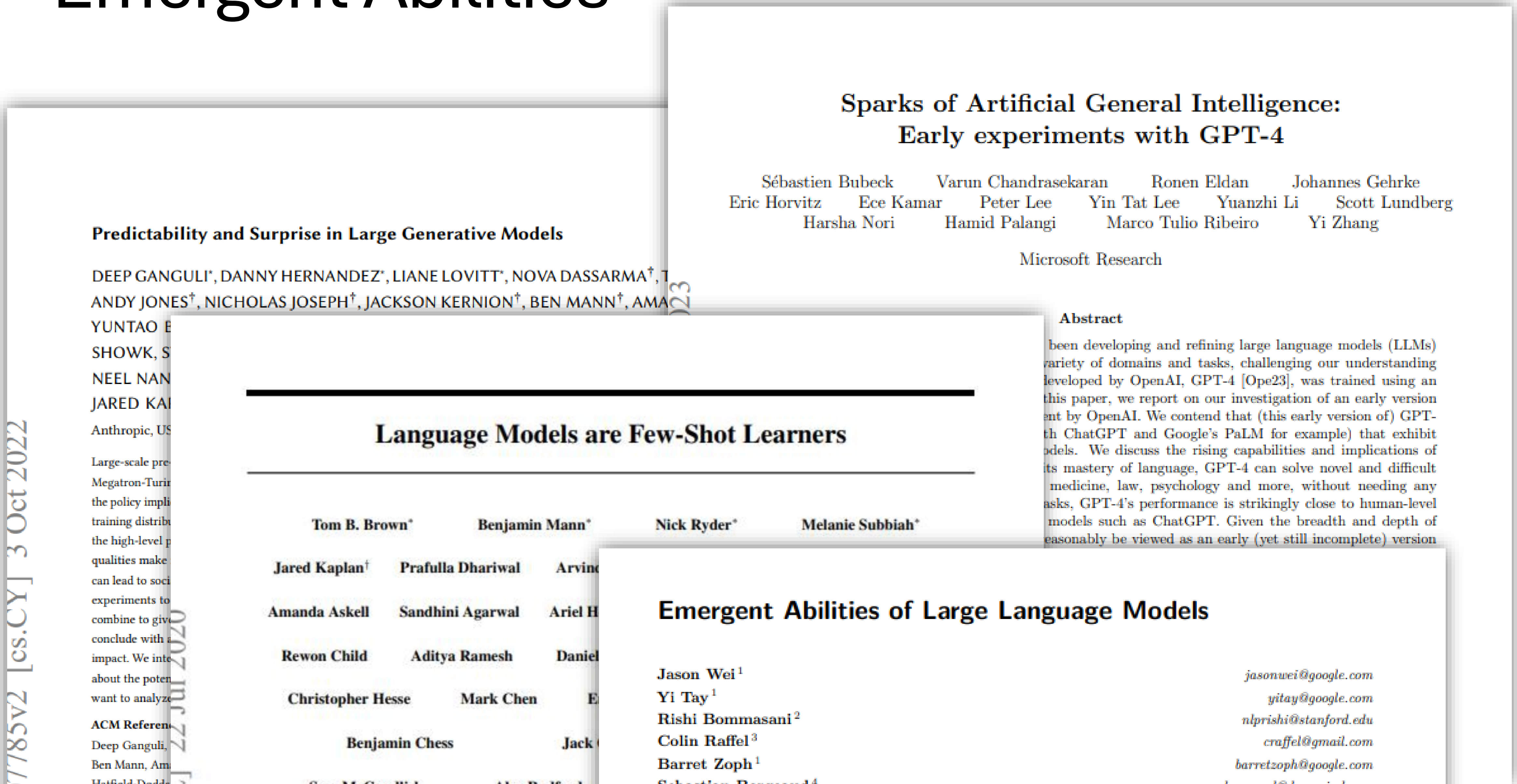
Recent work claims that large language models display *emergent abilities*, abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: that for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities, (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on BIG-Bench; and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep networks. Via all three analyses, we provide evidence that alleged emergent abilities evaporate with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

Are Emergent Abilities of LLMs a Mirage?

By Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Presented by Chao Chen (Michelle)

Emergent Abilities



Emergent Abilities are a Mirage

Sharp and unpredictable changes are induced by researcher's choice of metric. Model family's per-token error rate changes smoothly, continuously, and predictable

Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

Part III: Meta-analysis of emergent abilities

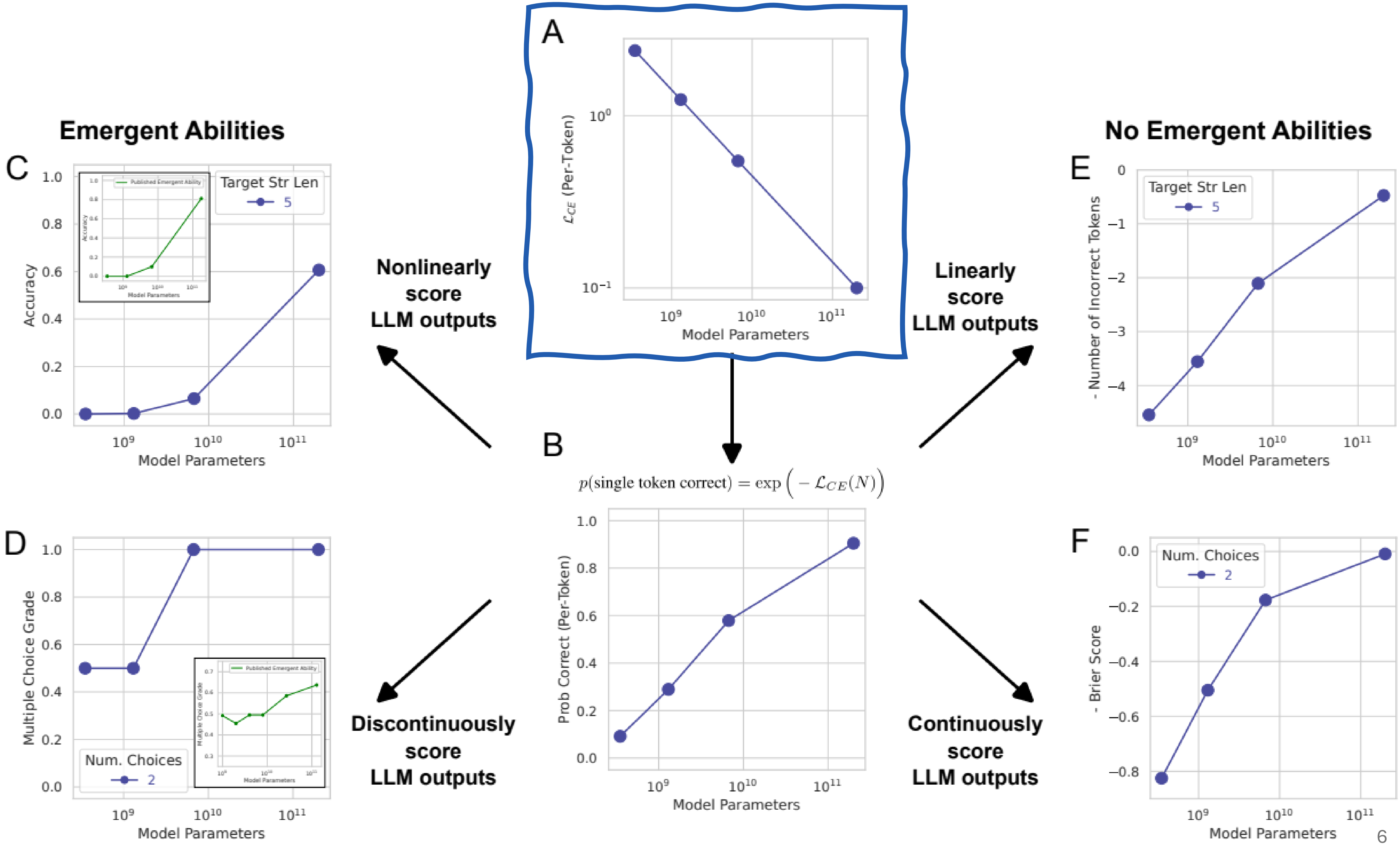
Part IV: Induce emergent abilities on vision models

Part I: Intuition on emergent abilities

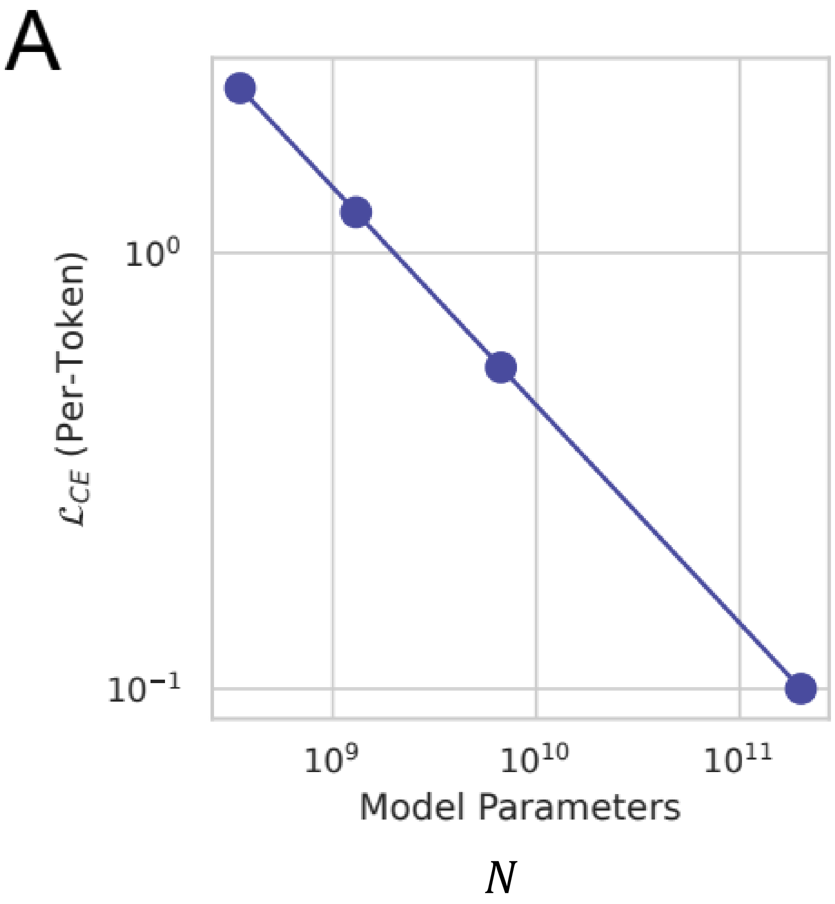
Part II: Empirically prove hypothesis InstructGPT/GPT-3

Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models



A
$\mathcal{L}_{CE}$ (Per-Token)
10^0
10^{-1}
10^9
10^{10}
10^{11}
Model Parameters
Emergent Abilities
C
Accuracy
Target Str Len
5
Model Parameters
Nonlinearly score LLM outputs
No Emergent Abilities
E
- Number of Incorrect Tokens
Target Str Len
5
Model Parameters
Linearly score LLM outputs
B
$p(\text{single token correct}) = \exp\Big(-\mathcal{L}_{CE}(N) \Big)$
Prob Correct (Per-Token)
Model Parameters
D
Multiple Choice Grade
Num. Choices
2
Model Parameters
Discontinuously score LLM outputs
F
- Brier Score
Num. Choices
2
Model Parameters
Continuously score LLM outputs



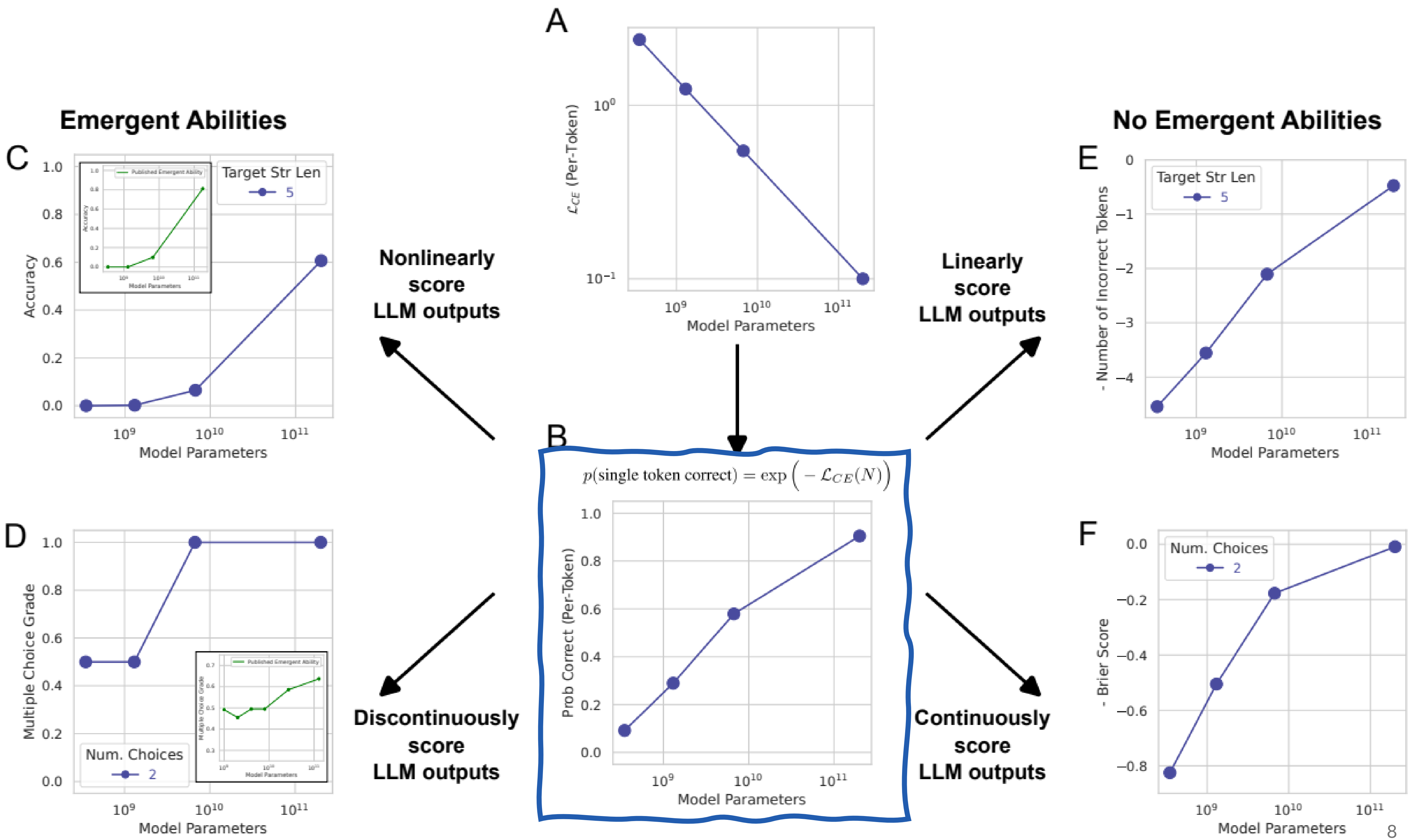
$$\mathcal{L}_{CE}(N) = \left(\frac{N}{C}\right)^{\alpha}$$

Cross Entropy Loss

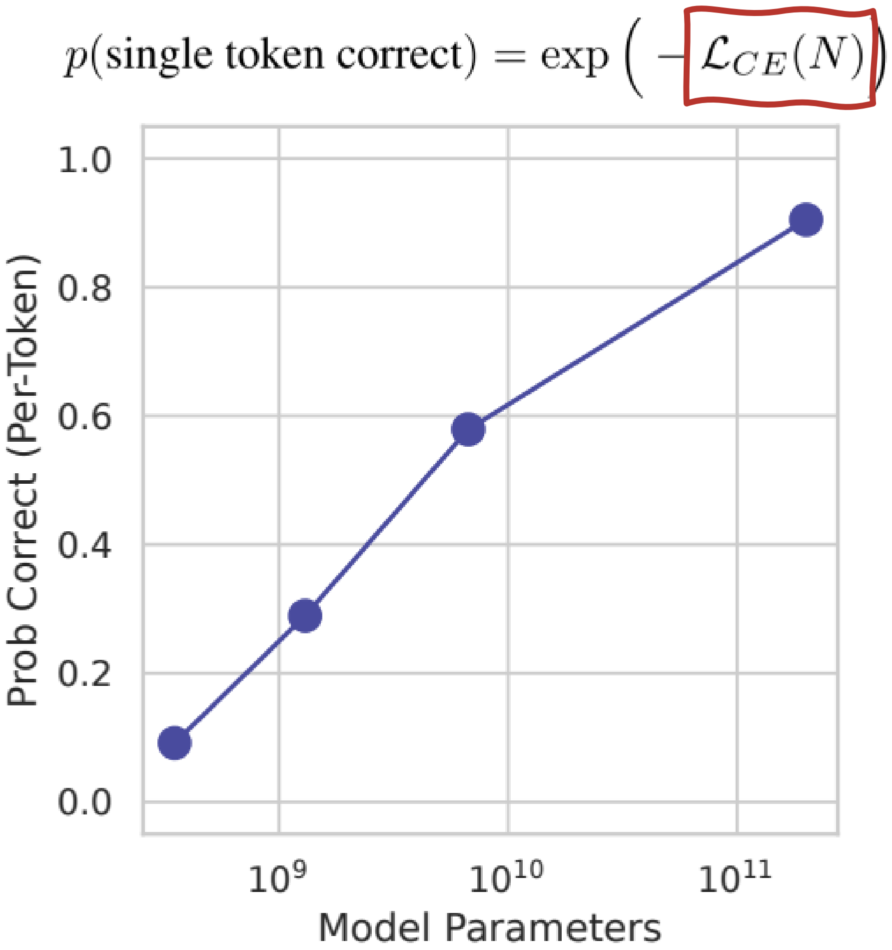
$$\mathcal{L}_{CE}(N) \coloneqq -\sum_{v \in V} p(v) \log \hat{p}_N(v)$$

One-hot distribution

$$\mathcal{L}_{CE}(N) = -\log \hat{p}_N(v^*)$$

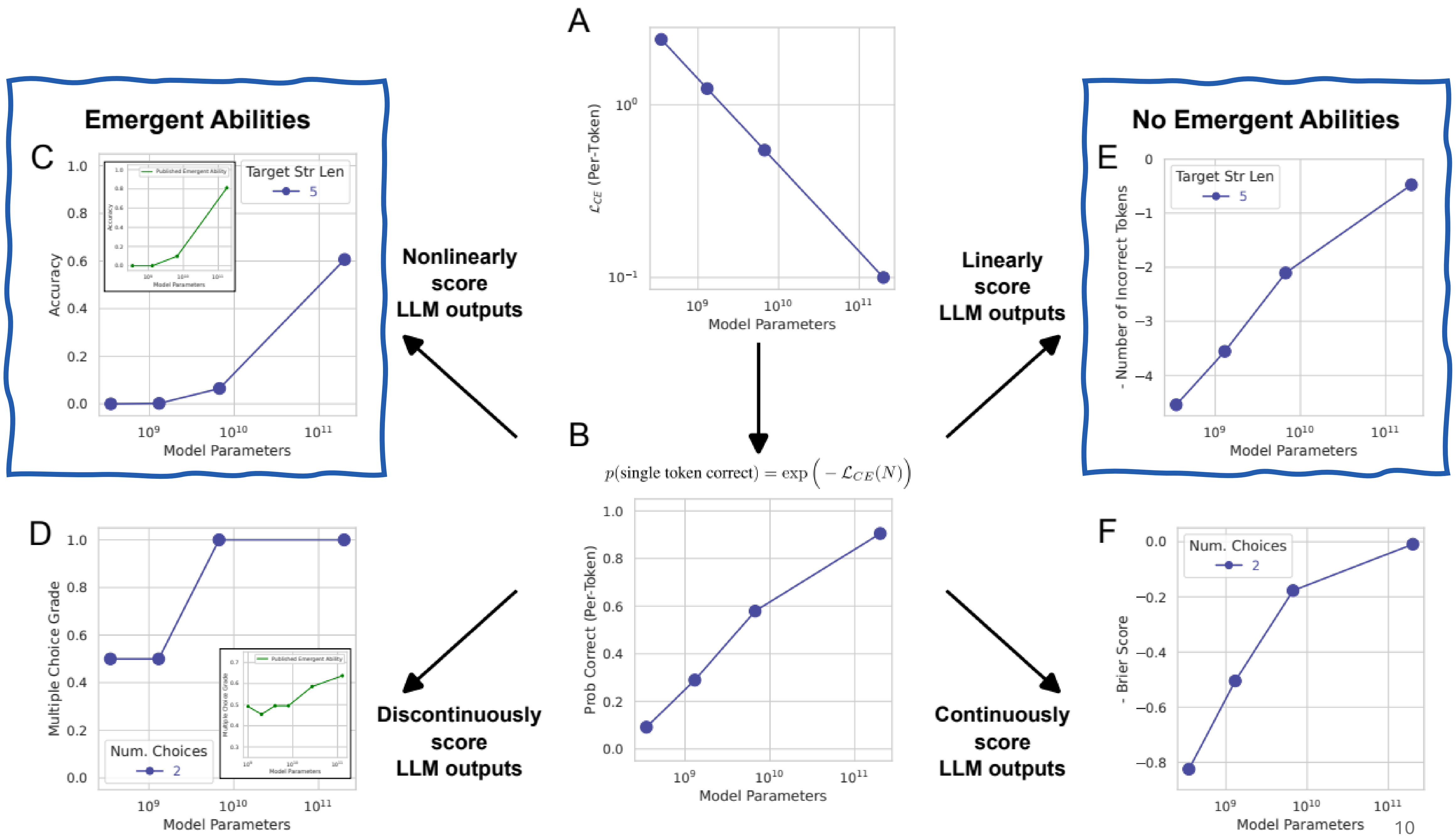


A
$\mathcal{L}_{CE}$ (Per-Token)
Model Parameters
Emergent Abilities
C
Target Str Len
5
Published Emergent Ability
Accuracy
Model Parameters
Nonlinearly score LLM outputs
Linearly score LLM outputs
No Emergent Abilities
E
Target Str Len
5
- Number of Incorrect Tokens
Model Parameters
B
$p(\text{single token correct}) = \exp\left(-\mathcal{L}_{CE}(N)\right)$
Prob Correct (Per-Token)
Model Parameters
D
Multiple Choice Grade
Num. Choices
2
Published Emergent Ability
Model Parameters
Discontinuously score LLM outputs
Continuously score LLM outputs
F
Num. Choices
2
- Brier Score
Model Parameters

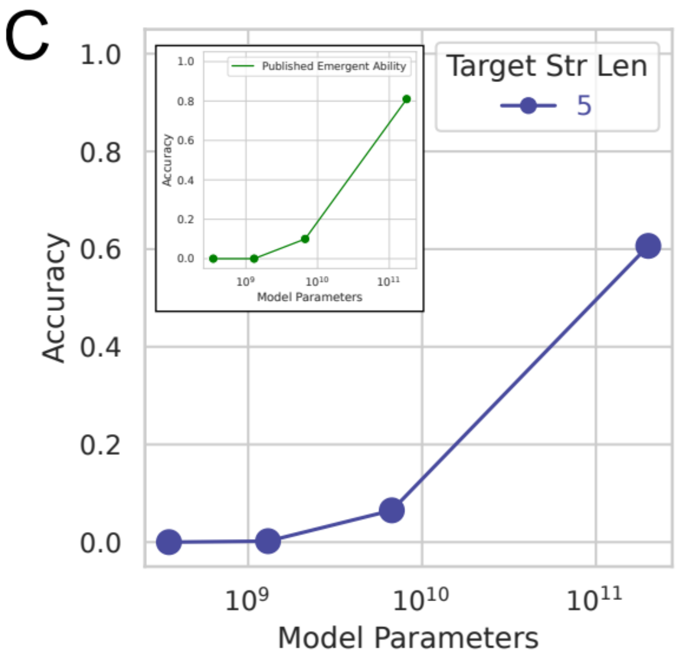


$$\mathcal{L}_{CE}(N) = -\log \hat{p}_N(v^*)$$

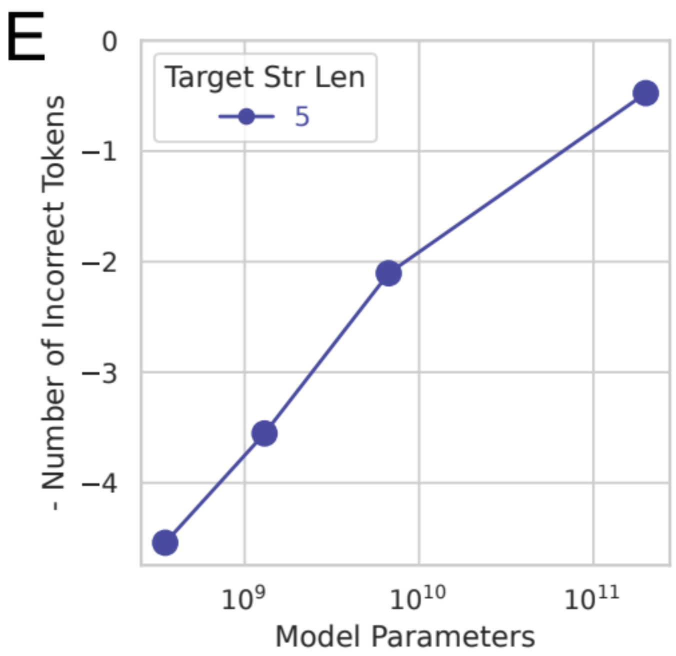
$$p_N(\text{single token correct}) = \exp\left(-\boxed{\left(\frac{N}{C}\right)^{\alpha}}\right)$$


A
B
C
D
E
F
Emergent Abilities
No Emergent Abilities
Nonlinearly score LLM outputs
Linearly score LLM outputs
Discontinuously score LLM outputs
Continuously score LLM outputs
$p(\text{single token correct}) = \exp\left(-\mathcal{L}_{CE}(N)\right)$
$\mathcal{L}_{CE}$ (Per-Token)
Prob Correct (Per-Token)
Accuracy
Multiple Choice Grade
- Number of Incorrect Tokens
- Brier Score
Model Parameters
Target Str Len
Num. Choices
Published Emergent Ability

Emergent Abilities

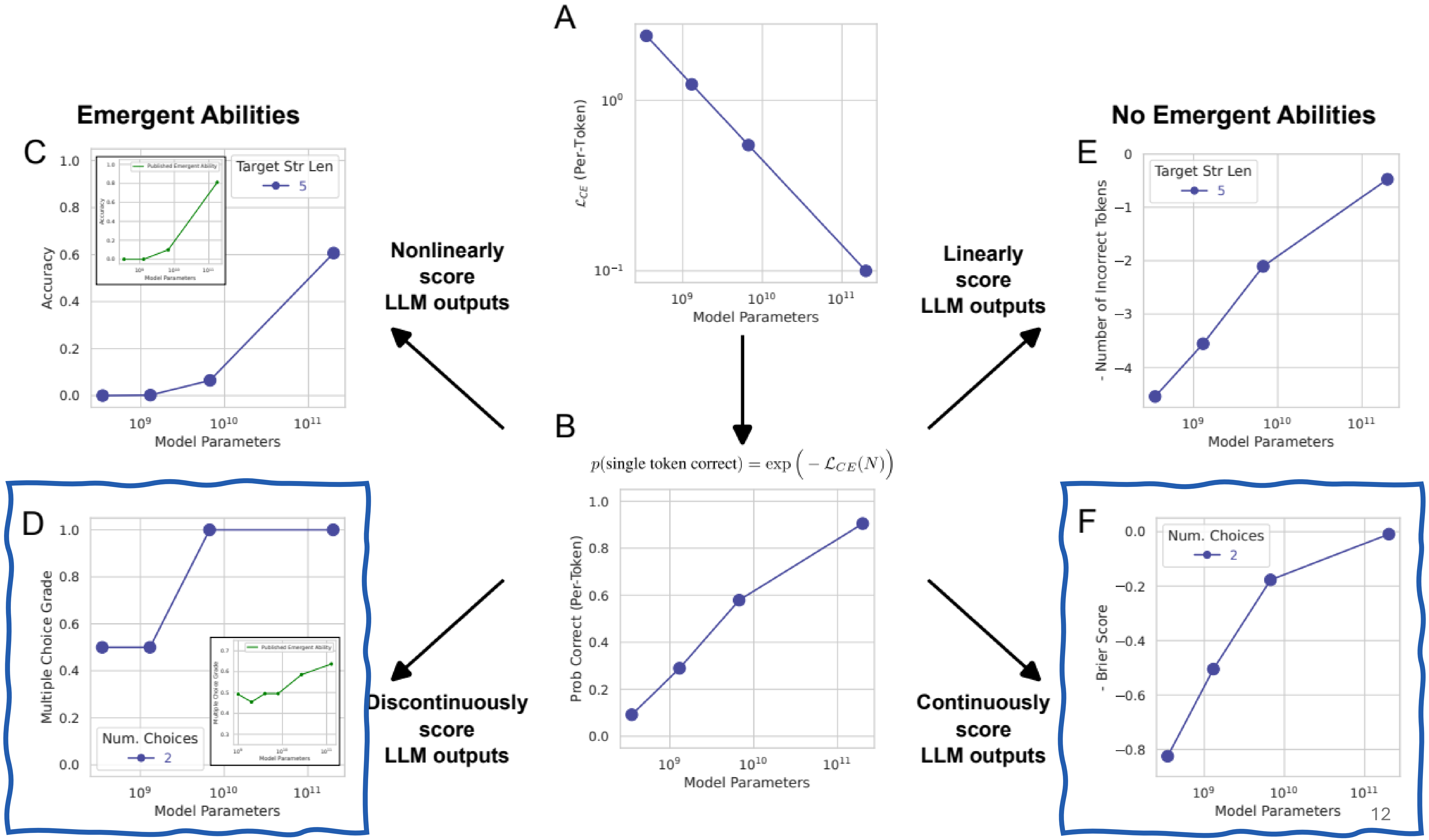


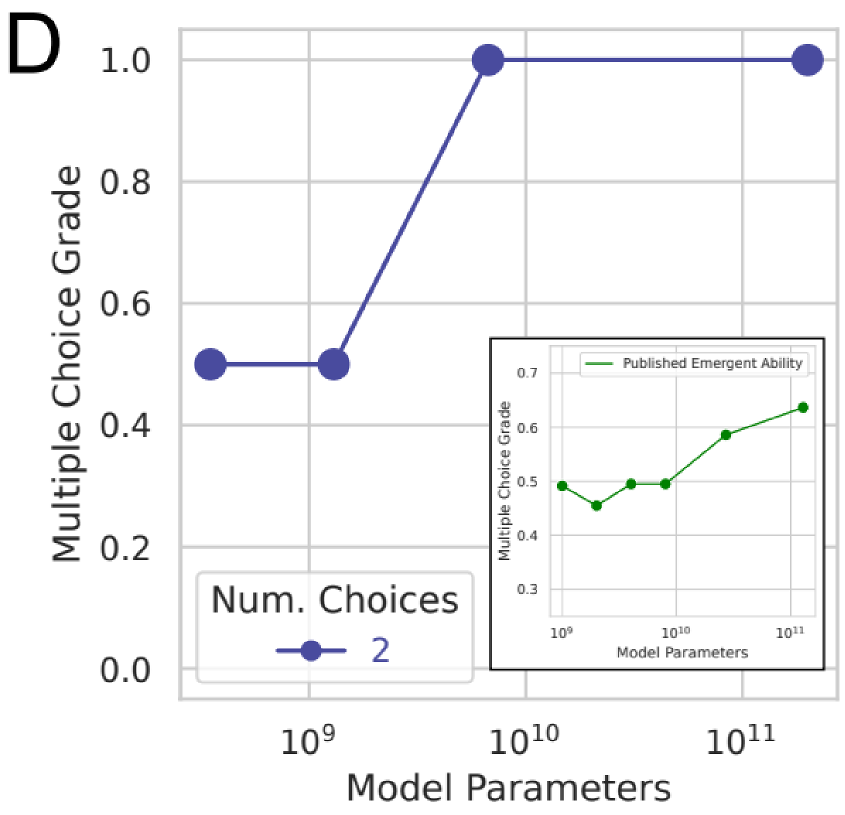
No Emergent Abilities

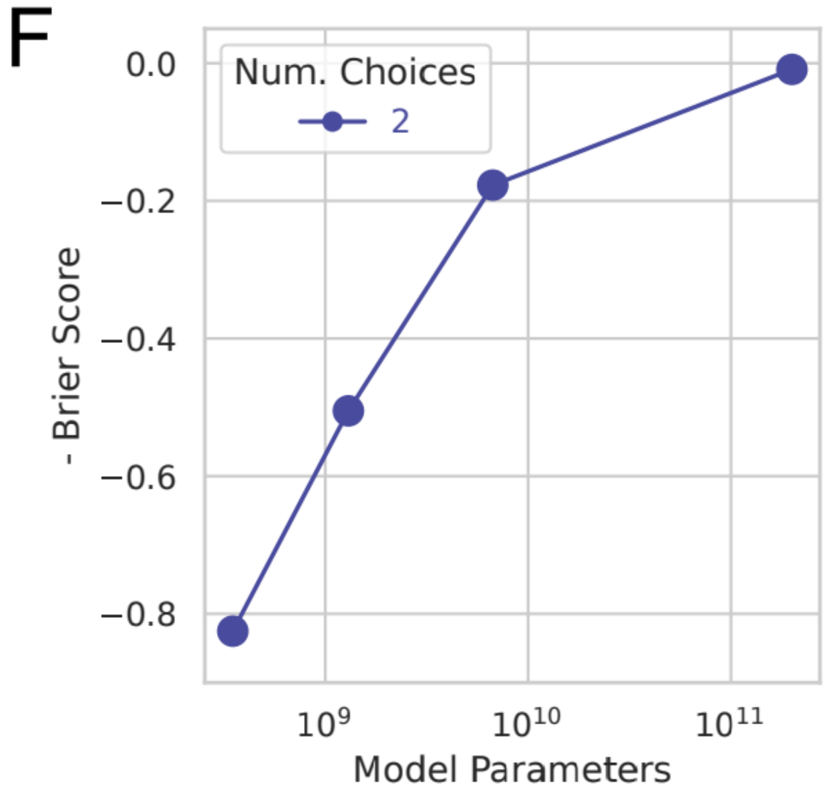


$$\mathrm{Accuracy}(N)$$
$$\approx p_N(\text{single token correct})^{\text{num of tokens}}$$
$$= \exp(-\left(\frac{N}{C}\right)^{\alpha})^{L}$$

$$\text{Token Edit Distance } (N)$$
$$\approx L \cdot \big(1 - p_N(\text{single token correct})\big)$$
$$= L \cdot (1 - \exp(-\left(\frac{N}{C}\right)^{\alpha}))$$


A
$\mathcal{L}_{CE}$ (Per-Token)
Model Parameters
B
$p(\text{single token correct}) = \exp\left(-\mathcal{L}_{CE}(N)\right)$
Prob Correct (Per-Token)
Model Parameters
Emergent Abilities
C
Accuracy
Target Str Len
5
Model Parameters
Nonlinearly
score
LLM outputs
No Emergent Abilities
E
- Number of Incorrect Tokens
Target Str Len
5
Model Parameters
Linearly
score
LLM outputs
D
Multiple Choice Grade
Num. Choices
2
Model Parameters
Discontinuously
score
LLM outputs
F
- Brier Score
Num. Choices
2
Model Parameters
Continuously
score
LLM outputs





$$\text{Multiple Choice Grade}(N) \approx \sum_{i=0}^{n} 1_{[p(v*) > p(v)]}$$

$$\text{Brier Score}(N) \approx \frac{1}{n} \sum_{i=0}^{n} (\hat{p}(v^*) - 1_{[v*]})^2$$

Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models

Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

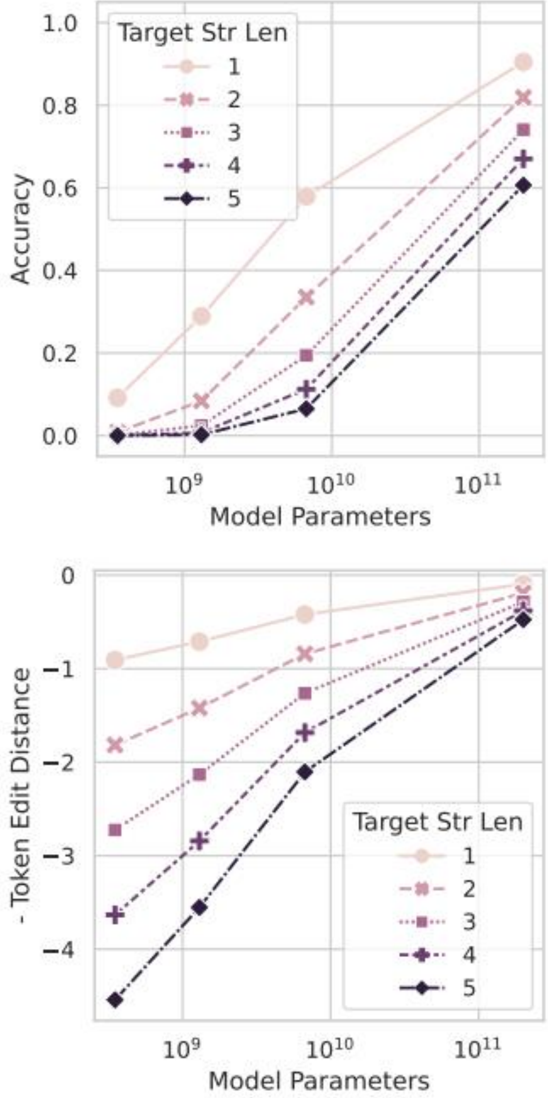
Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models

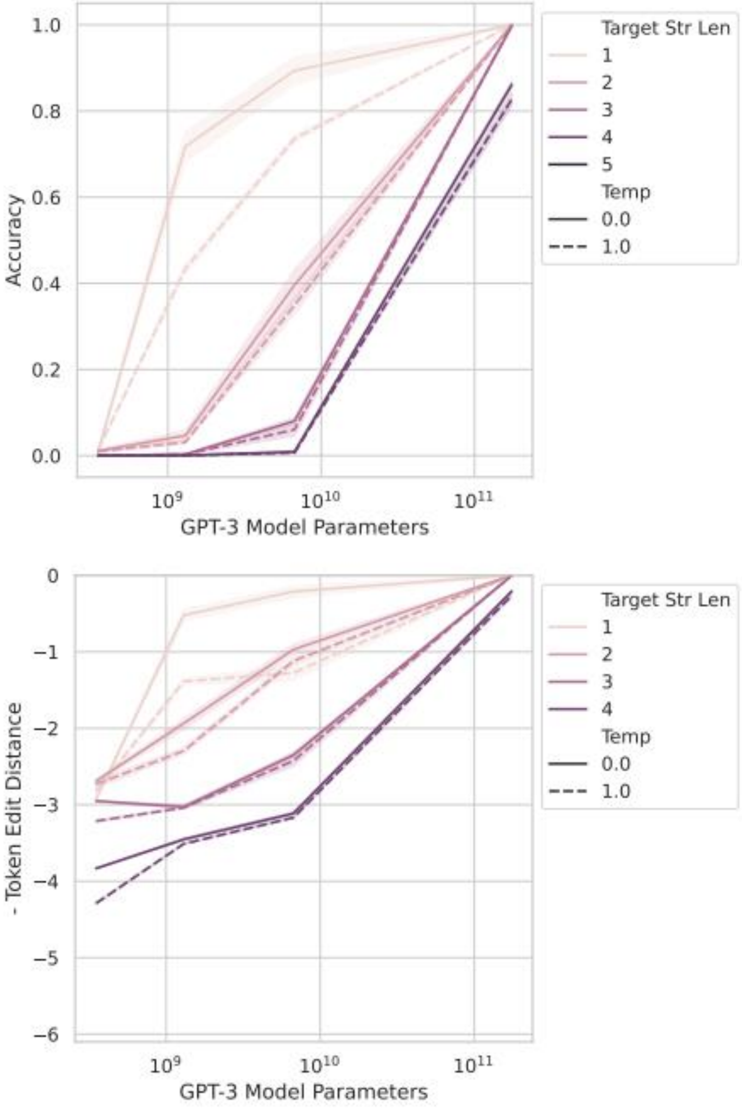
Predictions

1. Emergent abilities disappear with different metrics.
2. Emergent abilities disappear with better statistics.

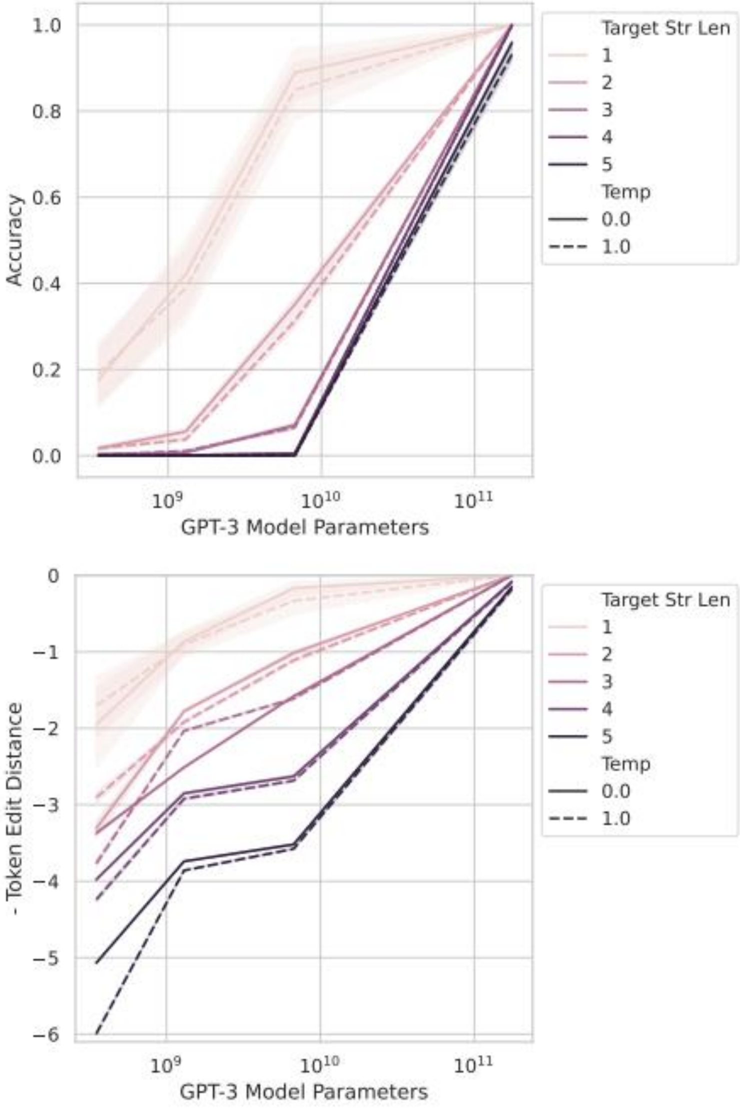
$$\mathcal{L}_{CE}(N) = \left(\frac{N}{C}\right)^{\alpha}$$



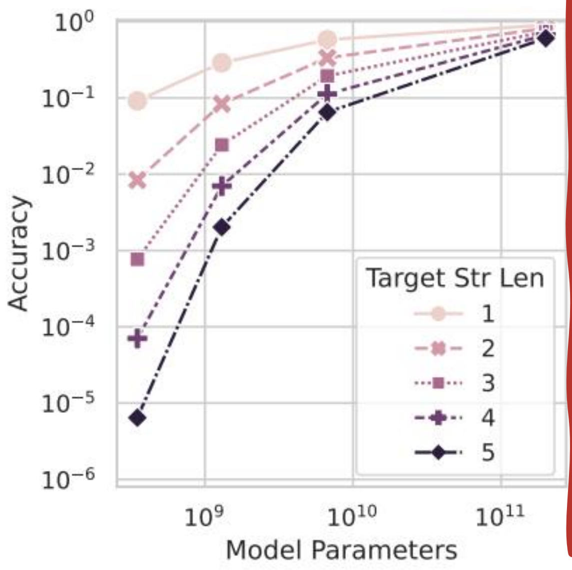
Two-digit multiplication



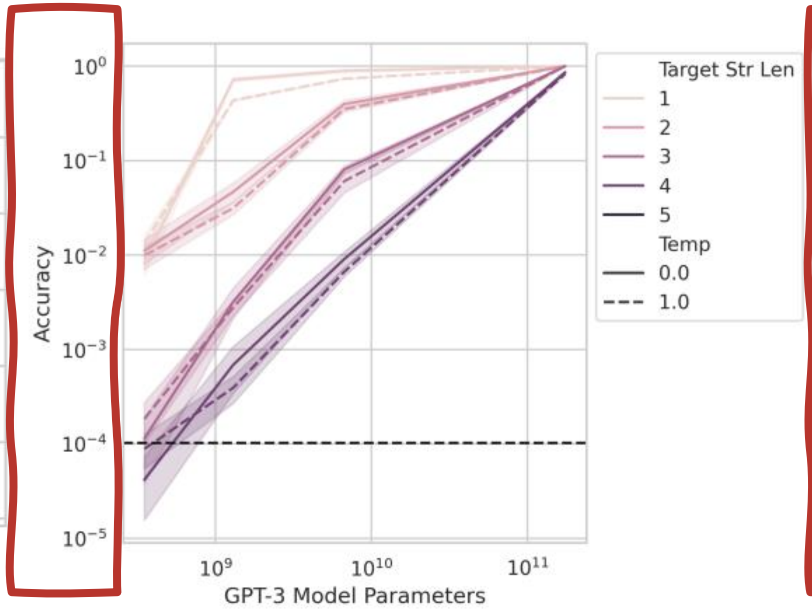
Four-digit addition



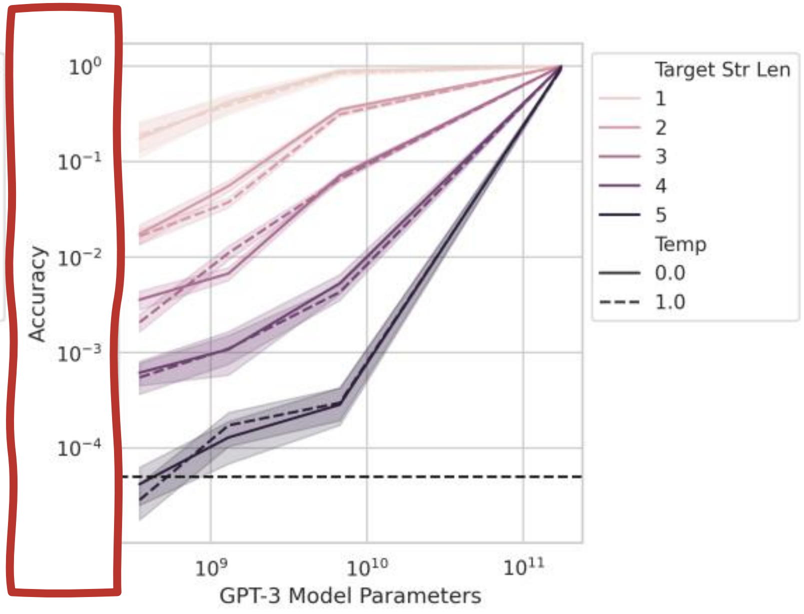
$$\mathcal{L}_{CE}(N) = \left(\frac{N}{C}\right)^{\alpha}$$



Two-digit multiplication



Four-digit addition



Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models

Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

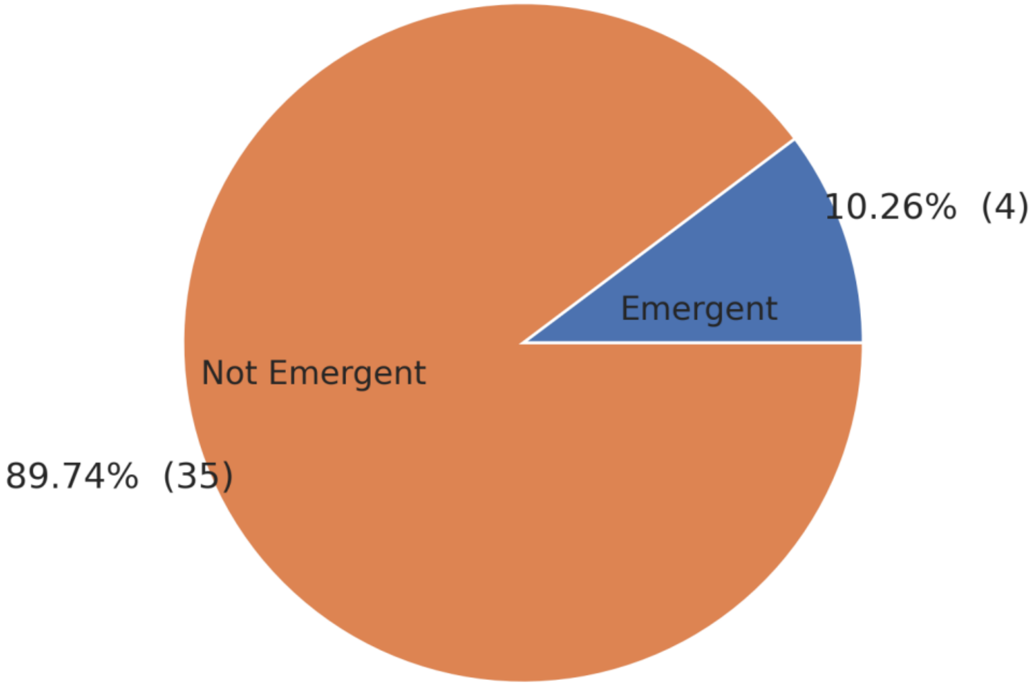
Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models

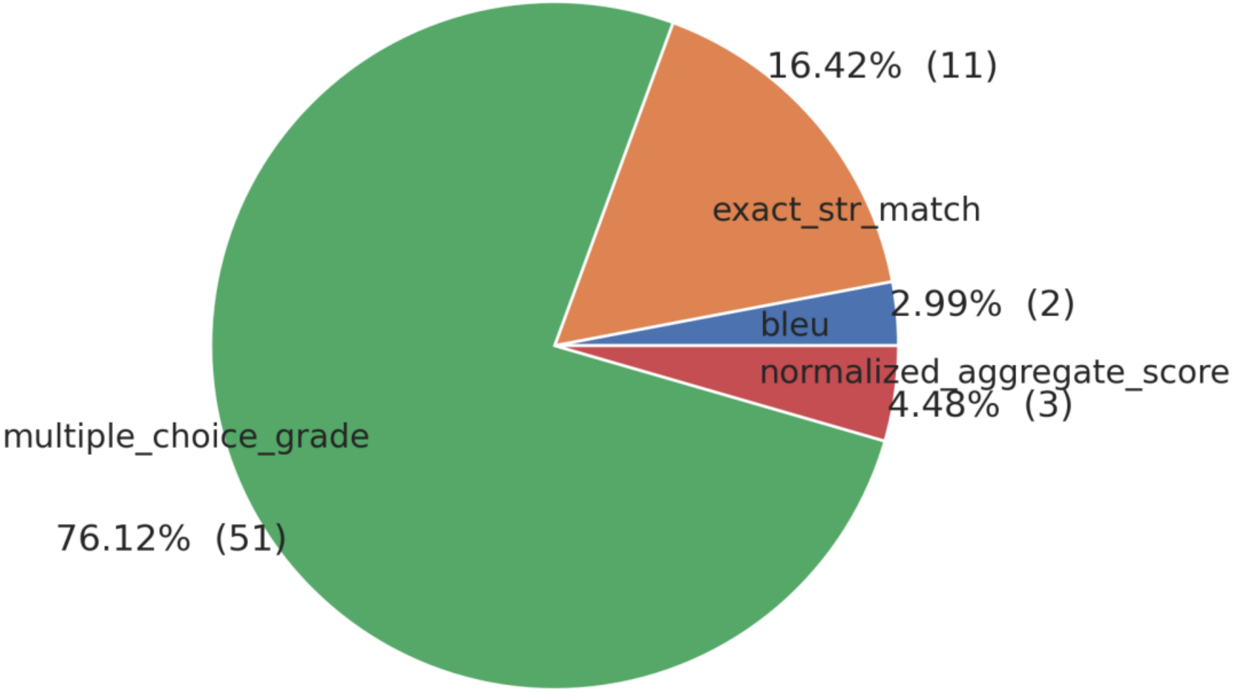
Predictions

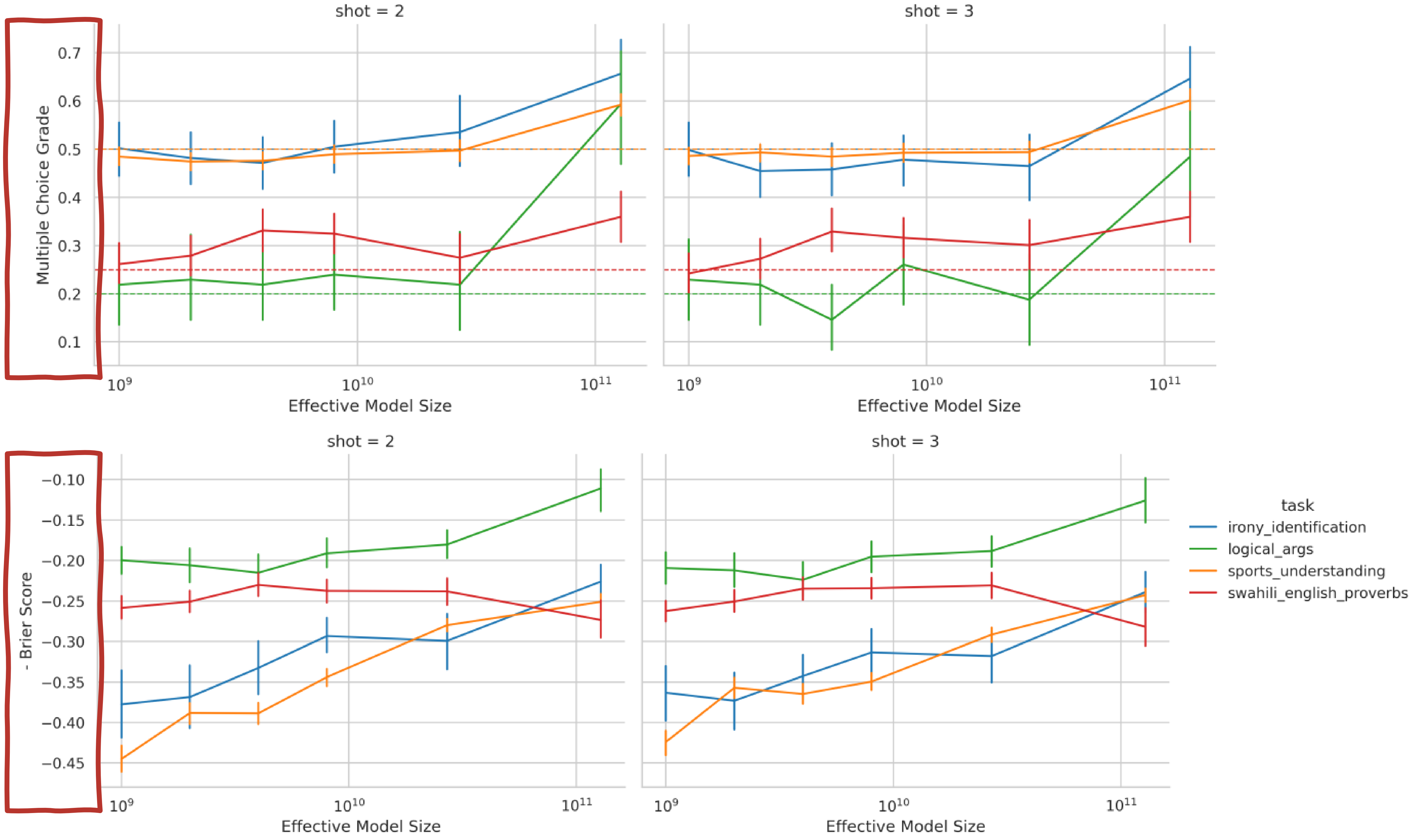
- Emergent abilities appear with discontinuous/non-linear metrics.
- Emergent abilities disappear after changing metric.

% of Metrics with > 1 Model-Task Pair Exhibiting Emergent Abilities



Metrics of Model-Task Pairs Exhibiting Emergent Abilities




shot = 2
shot = 3
Multiple Choice Grade
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10^9
10^{10}
10^{11}
Effective Model Size
- Brier Score
−0.10
−0.15
−0.20
−0.25
−0.30
−0.35
−0.40
−0.45
task
irony_identification
logical_args
sports_understanding
swahili_english_proverbs

Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models

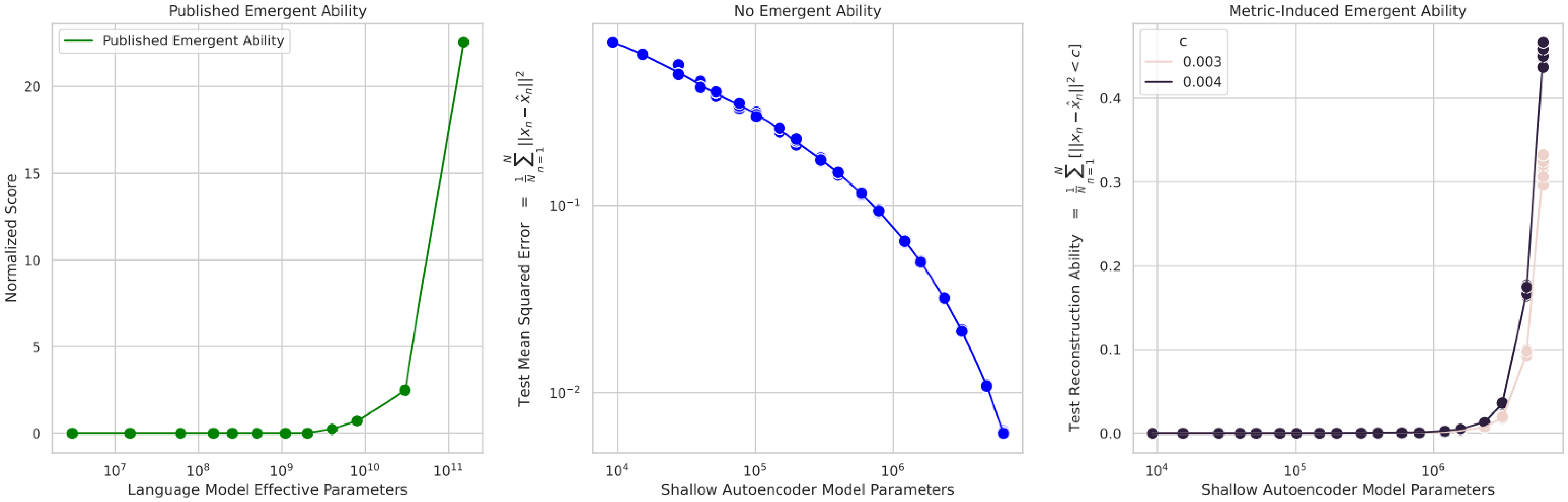
Part I: Intuition on emergent abilities

Part II: Empirically prove hypothesis InstructGPT/GPT-3

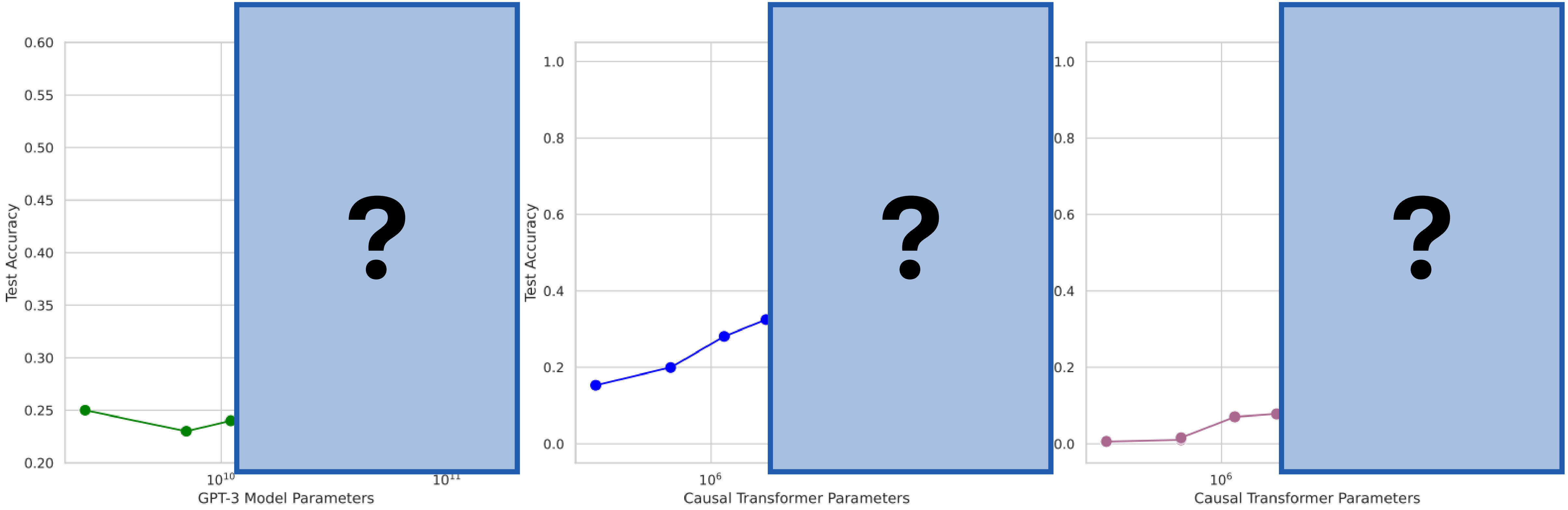
Part III: Meta-analysis of emergent abilities

Part IV: Induce emergent abilities on vision models

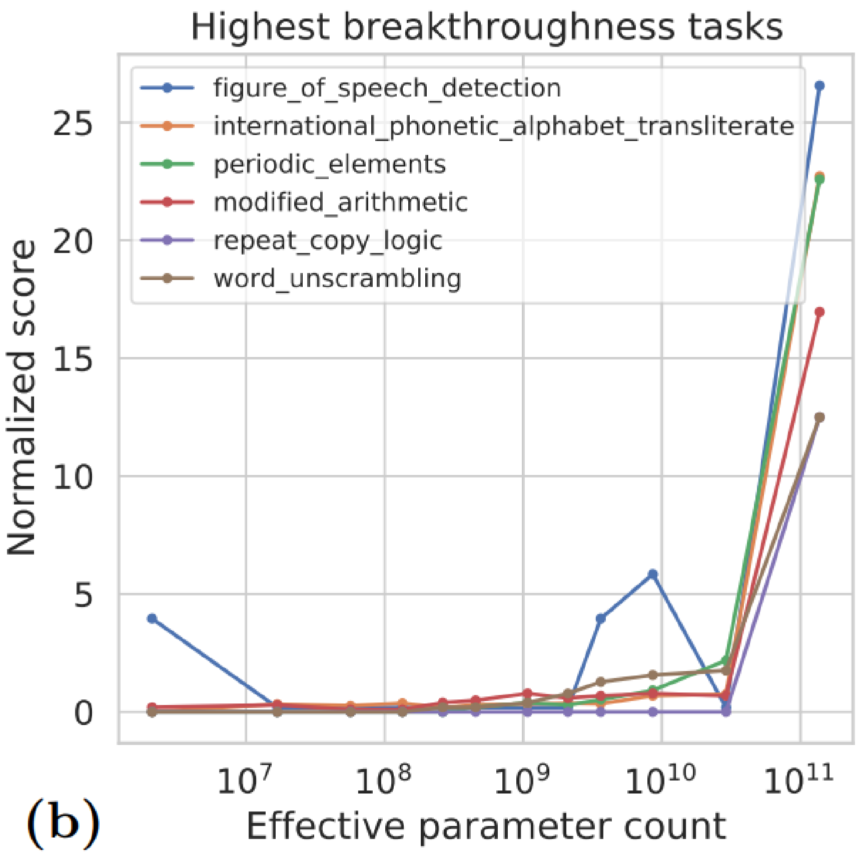
Shallow non-linear autoencoder for CIFAR100



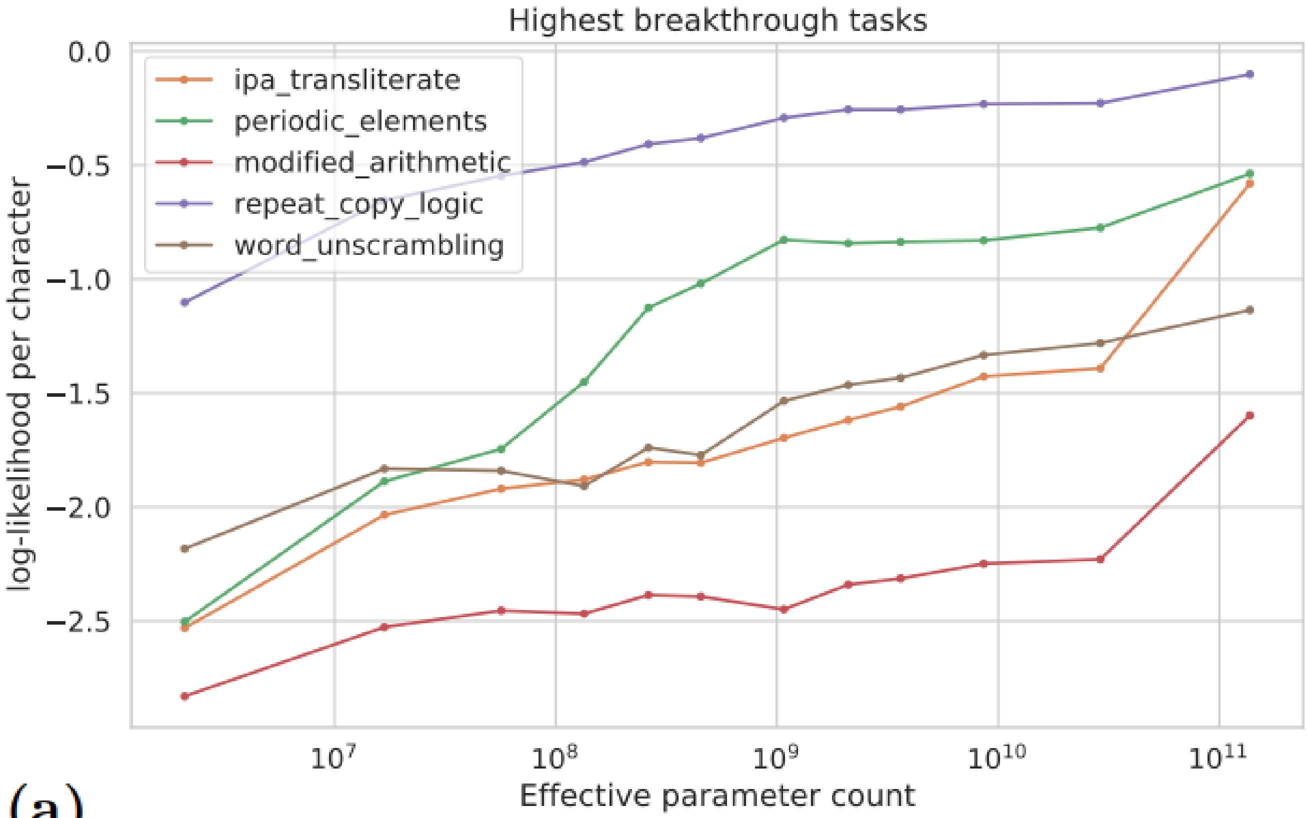
Transformer for classifying Omniglot characters



BIG Benchmark

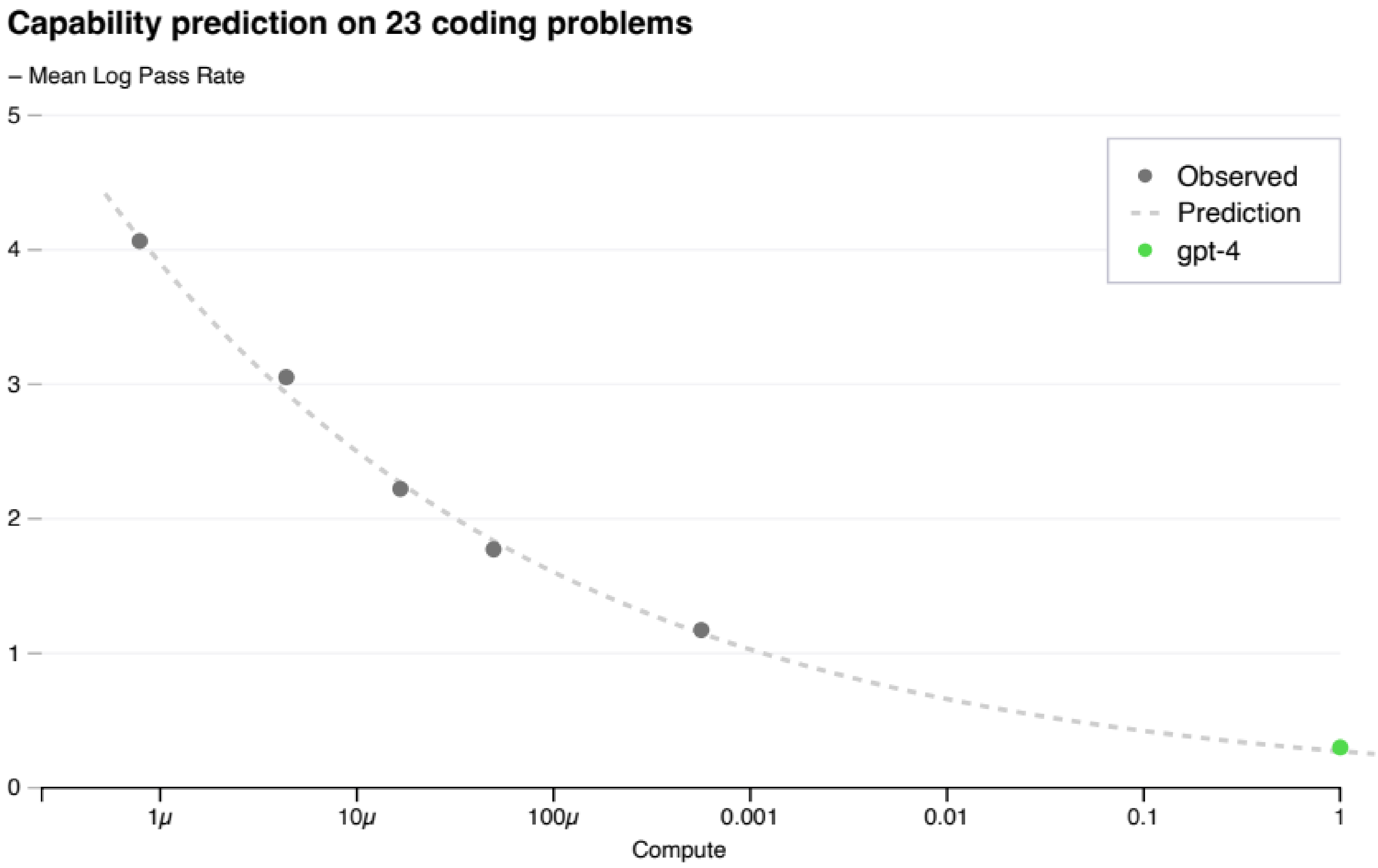


(b)



(a)

GPT-4 Technical Report



Are Emergent Abilities a Mirage?


Published in Transactions on Machine Learning Research (08/2022)

Emergent Abilities of Large Language Models

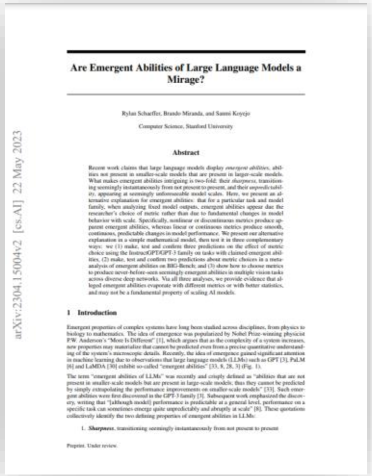
Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus

arXiv:2206.07682v2 [cs.CL] 26 Oct 2022

Abstract

1 Introduction

Emergent abilities only occur with certain metrics.


Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Computer Science, Stanford University

arXiv:2304.15004v2 [cs.AI] 22 May 2023

Abstract

1 Introduction

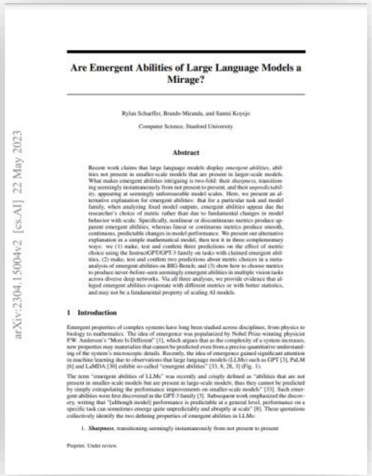
Those metrics are the ones that matter.

Are Emergent Abilities a Mirage?


Published in Transactions on Machine Learning Research (08/2022)

Emergent Abilities of Large Language Models

Plots have large jumps due to log-scaled x-axis.


Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

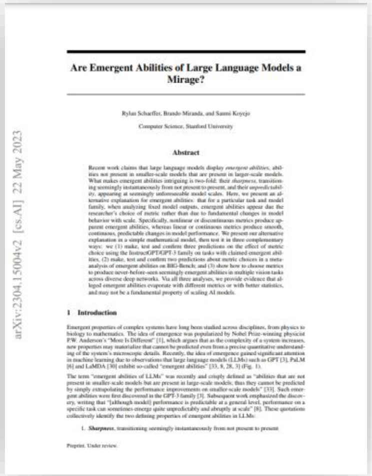
Computer Science, Stanford University

Linear scaling of x-axis also shows jumps.

Are Emergent Abilities a Mirage?


Emergent Abilities of Large Language Models

X-axis is not sampled densely enough.


Are Emergent Abilities of Large Language Models a Mirage?

The trend cannot be extrapolated.

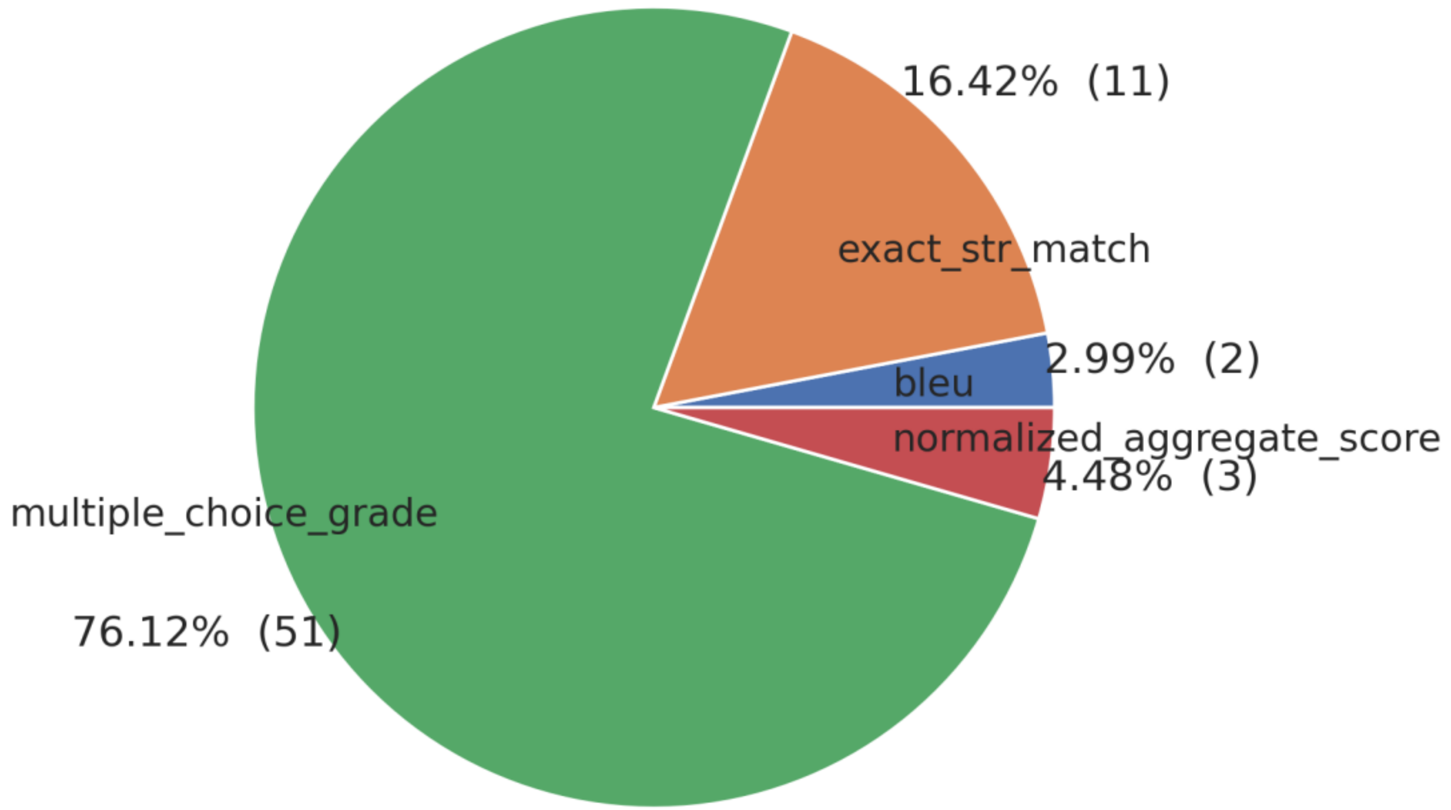
Discussion

Strengths:

- Multiple arguments to support their hypothesis.
- Clear explanation for unpredictable trends.

Weaknesses:

- Unpredictability of improvement.


16.42% (11)
exact_str_match
2.99% (2)
bleu
normalized_aggregate_score
4.48% (3)
multiple_choice_grade
76.12% (51)

Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Computer Science, Stanford University

Abstract

Recent work claims that large language models display *emergent abilities*, abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: that for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities, (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on BIG-Bench; and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep networks. Via all three analyses, we provide evidence that alleged emergent abilities evaporate with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

Are Emergent Abilities of LLMs a Mirage

By Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Presented by Chao Chen (Michelle)

Predictability and Surprise in LGM

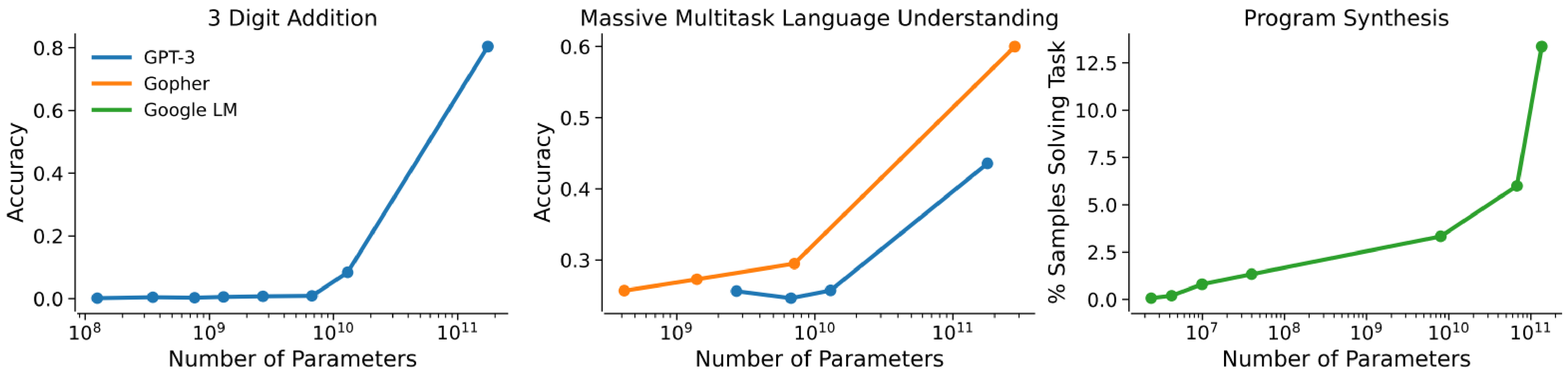


Fig. 2 Three examples of abrupt specific capability scaling described in Section 2.2, based on three different models: GPT-3 (blue), Gopher (orange), and a Google language model (green). **(Left)** 3-Digit addition with GPT-3 [11]. **(Middle)** Language understanding with GPT-3 and Gopher [62]. **(Right)** Program synthesis with Google language models [4].