Non-Emergent (Linguistic) Properties

27 Oct 2023
FSU SC-ML
Tom Juzek

Resources General

Schaeffer et al. 2023:

https://arxiv.org/pdf/2304.15004.pdf

Resources Linguistics

Warstadt and Bowman 2022:
https://arxiv.org/pdf/2208.07998.pdf

Warstadt et al. 2019:
https://arxiv.org/pdf/1805.12471.pdf

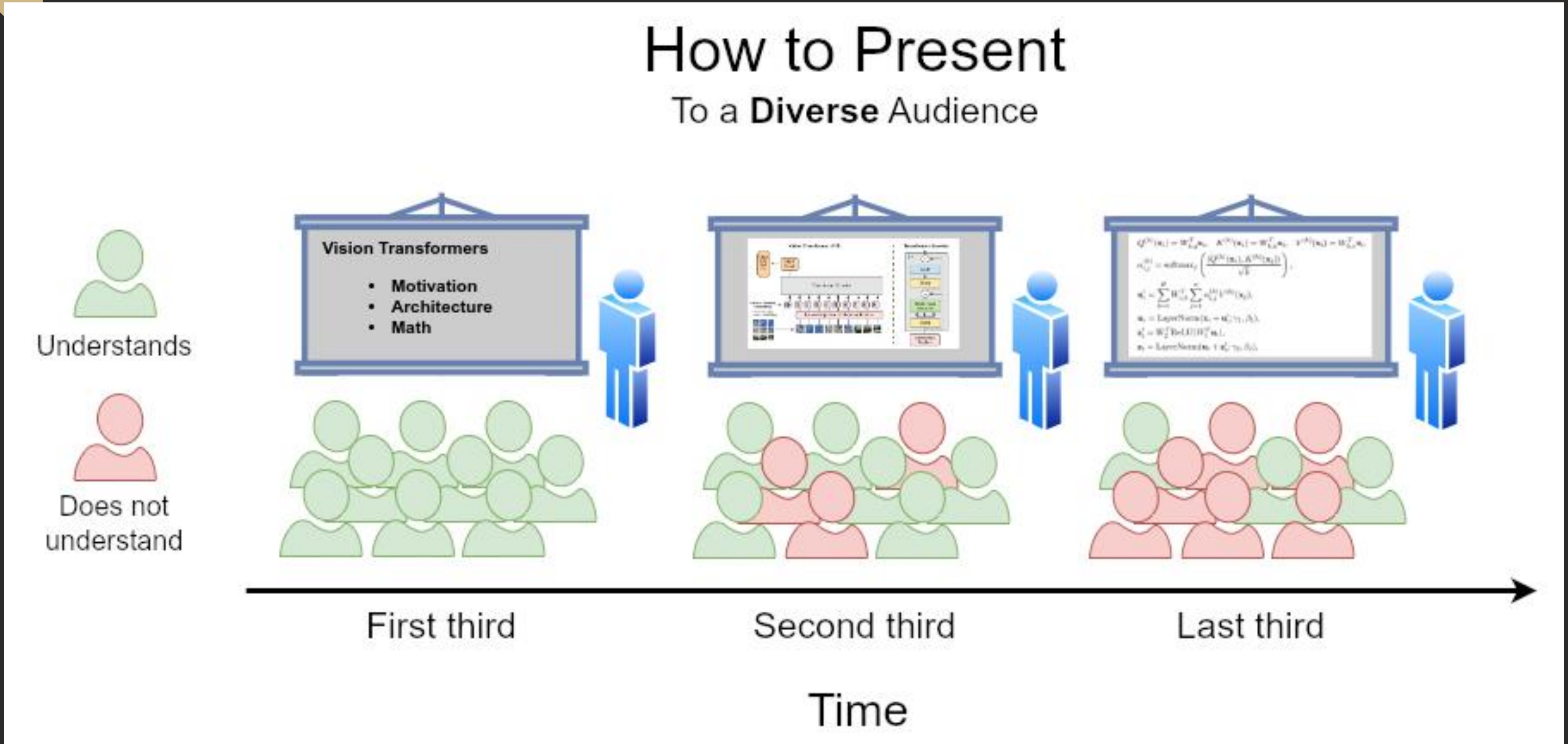
Anon SAD 2023 (paper+corpus*+code*):
https://github.com/arizus/sad

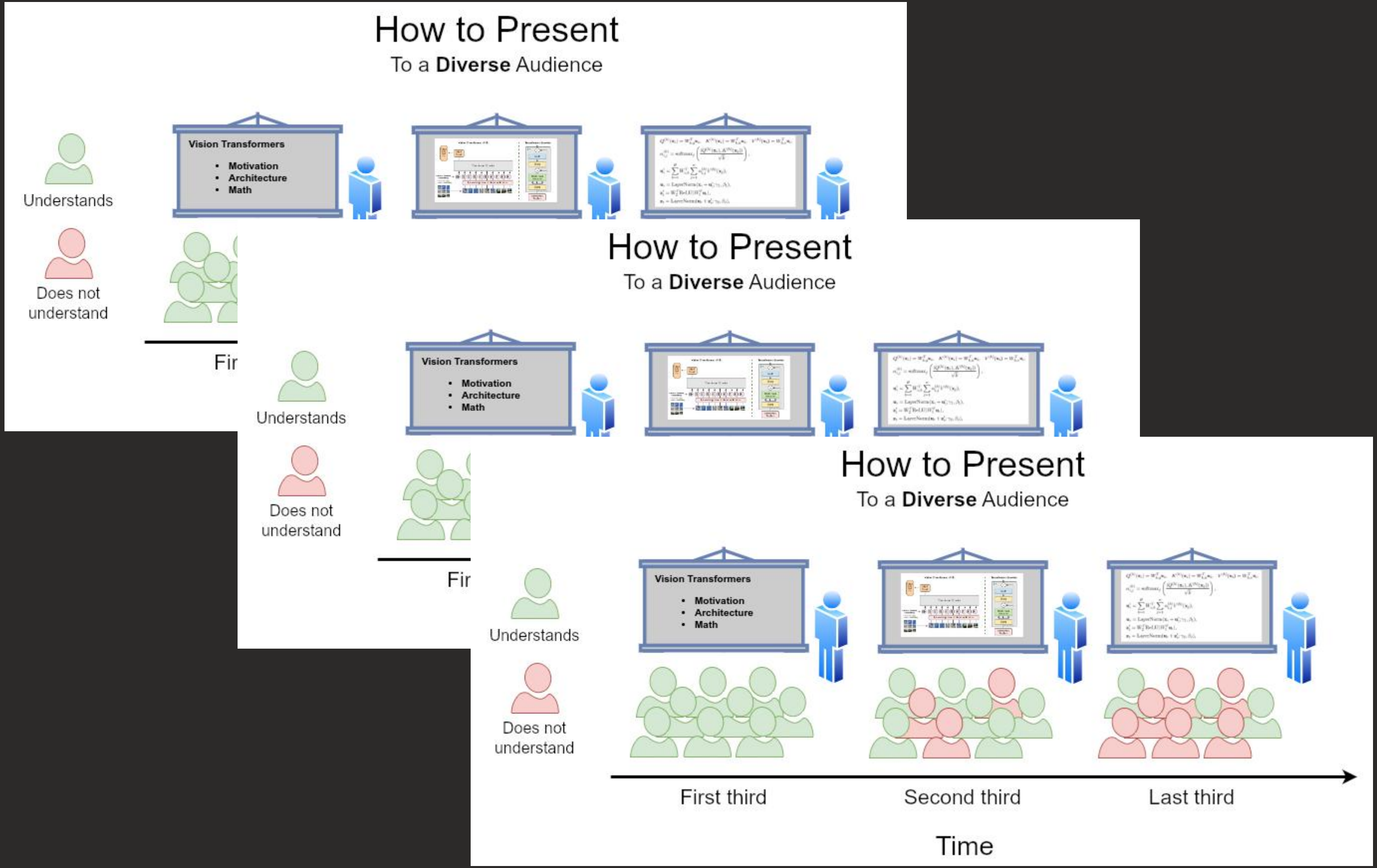
Overview

- Revisit emergence
- Schaeffer et al. 2023
- linguistic properties, current research line

→ usual format: Qs anytime

Overview


How to Present
To a Diverse Audience
Understands
Does not understand
Vision Transformers
• Motivation
• Architecture
• Math
First third
Second third
Last third
Time


How to Present
To a **Diverse** Audience
Understands
Does not understand
Vision Transformers
Motivation
Architecture
Math
First third
Second third
Last third
Time

Recap: Intro

Data - ODEP – ANN – Transformers – LLMs

A very general introduction

“AI/LLM revolution”, “structure of learning”

Recap

“Emergence is when quantitative changes in a system result in qualitative changes in behavior.”

Steinhardt 2022, “rooted in” (as per Wei et al. 2022):

Recap

Emergence:

- “sharp” learning
- “hard to predict”

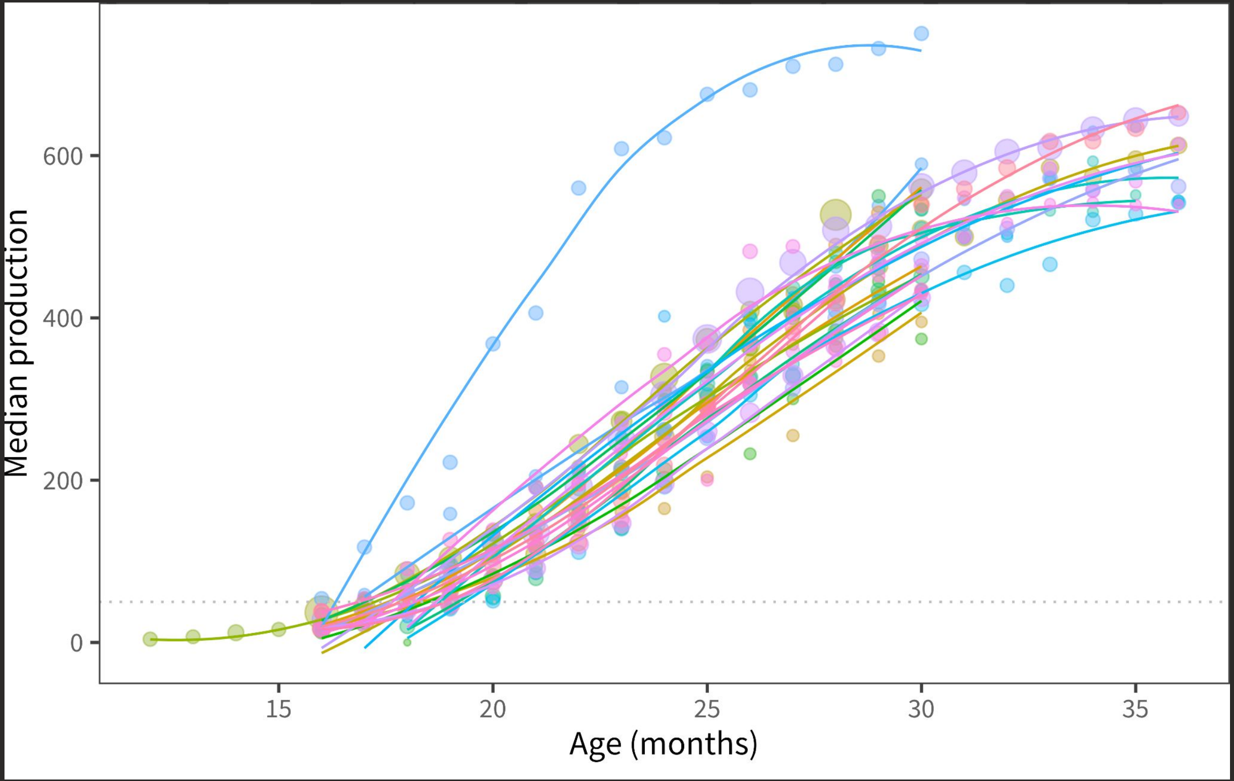
Recap

Emergence:

- “sharp” learning
- “hard to predict”

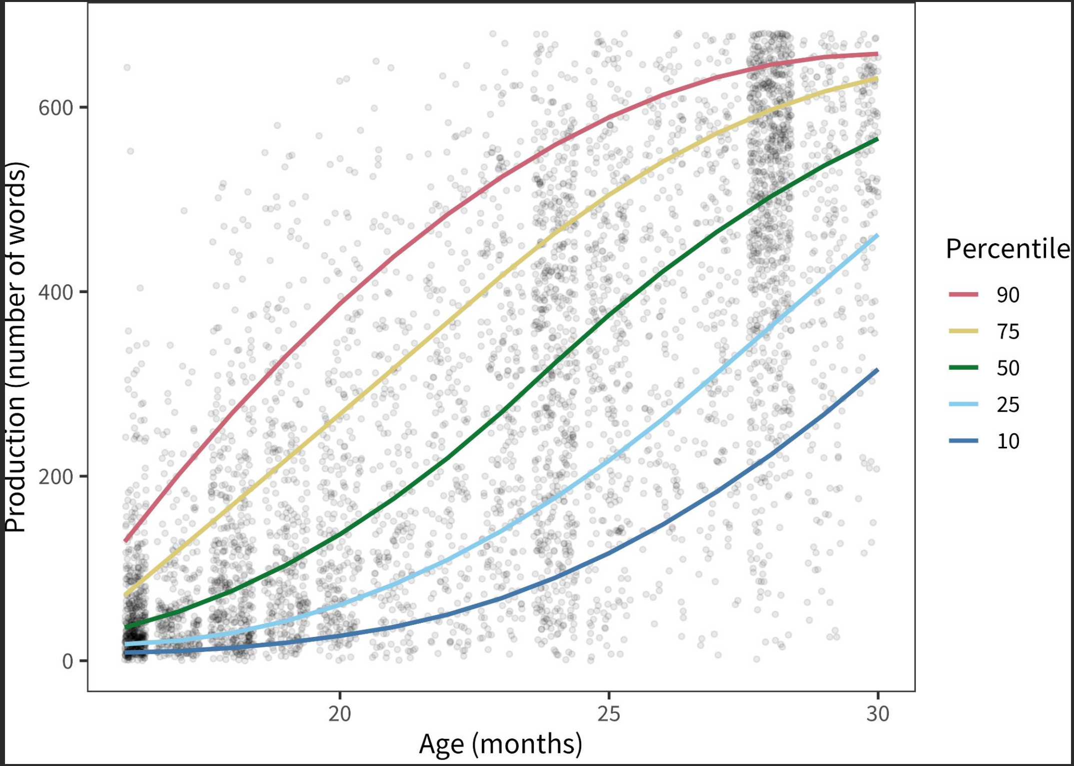
Zina Ward’s point about human abilities:

Emergence in human abilities



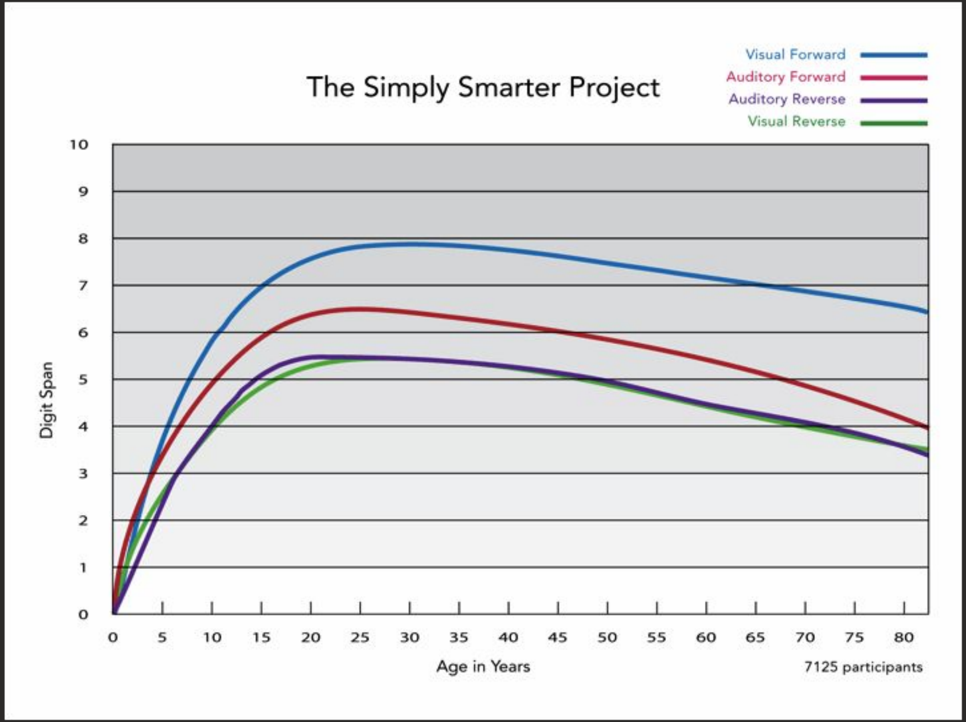
Frank et al. 2021. Word-bank book.

Emergence in human abilities



Frank et al. 2021. Word-bank book.

Nathan’s point about underlying …



Doman Jr. 2008. Short Term and Working Memory.

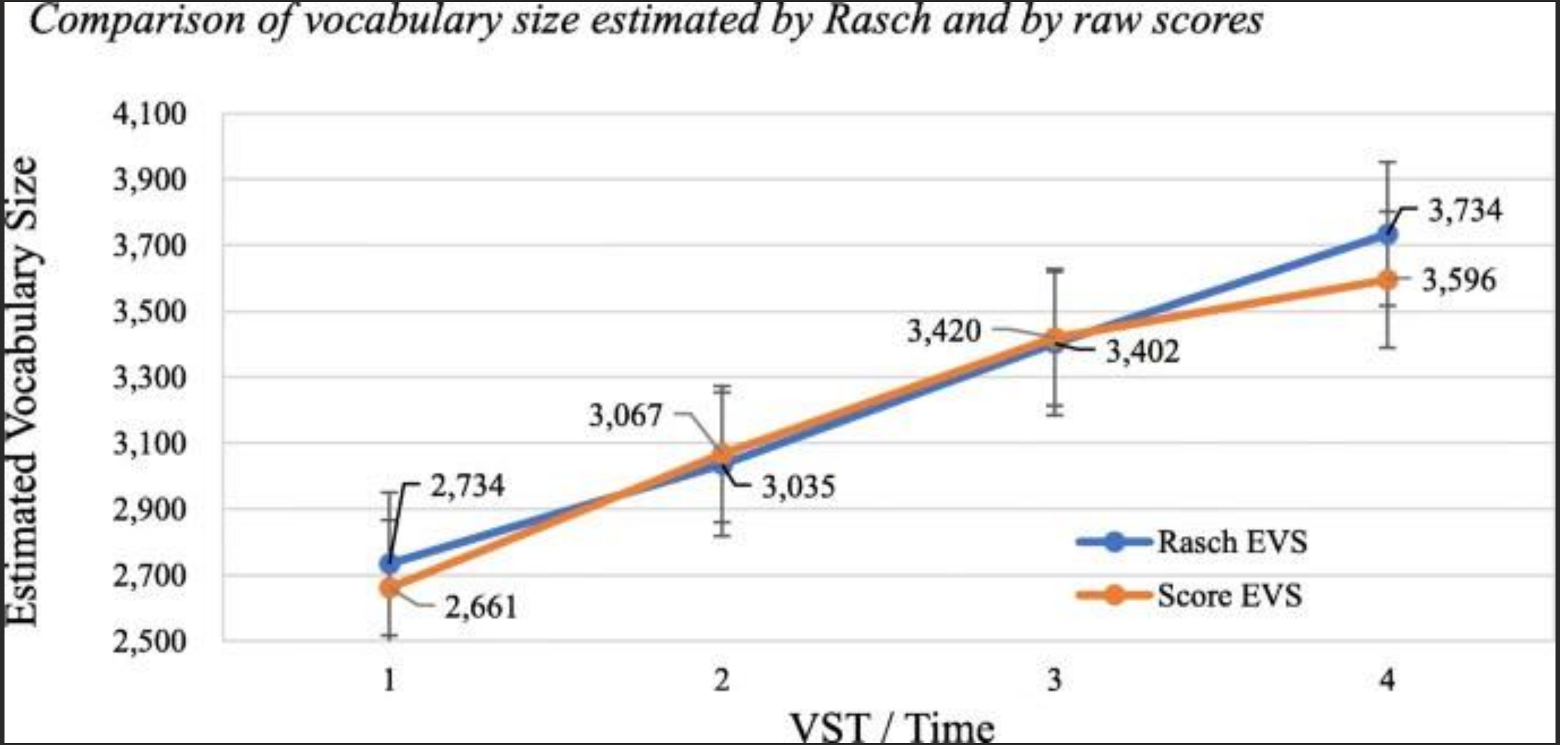
Non-Emergence in human abilities

Table 26.1 Average Native Speaker Vocabulary Sizes for Various Age Levels

Age	Average vocabulary size
6-year-olds	4,000 word families
7-year-olds	5,000 word families
8-year-olds	6,000 word families
9-year-olds	7,000 word families
10-year-olds	8,000 word families
11-year-olds	9,000 word families

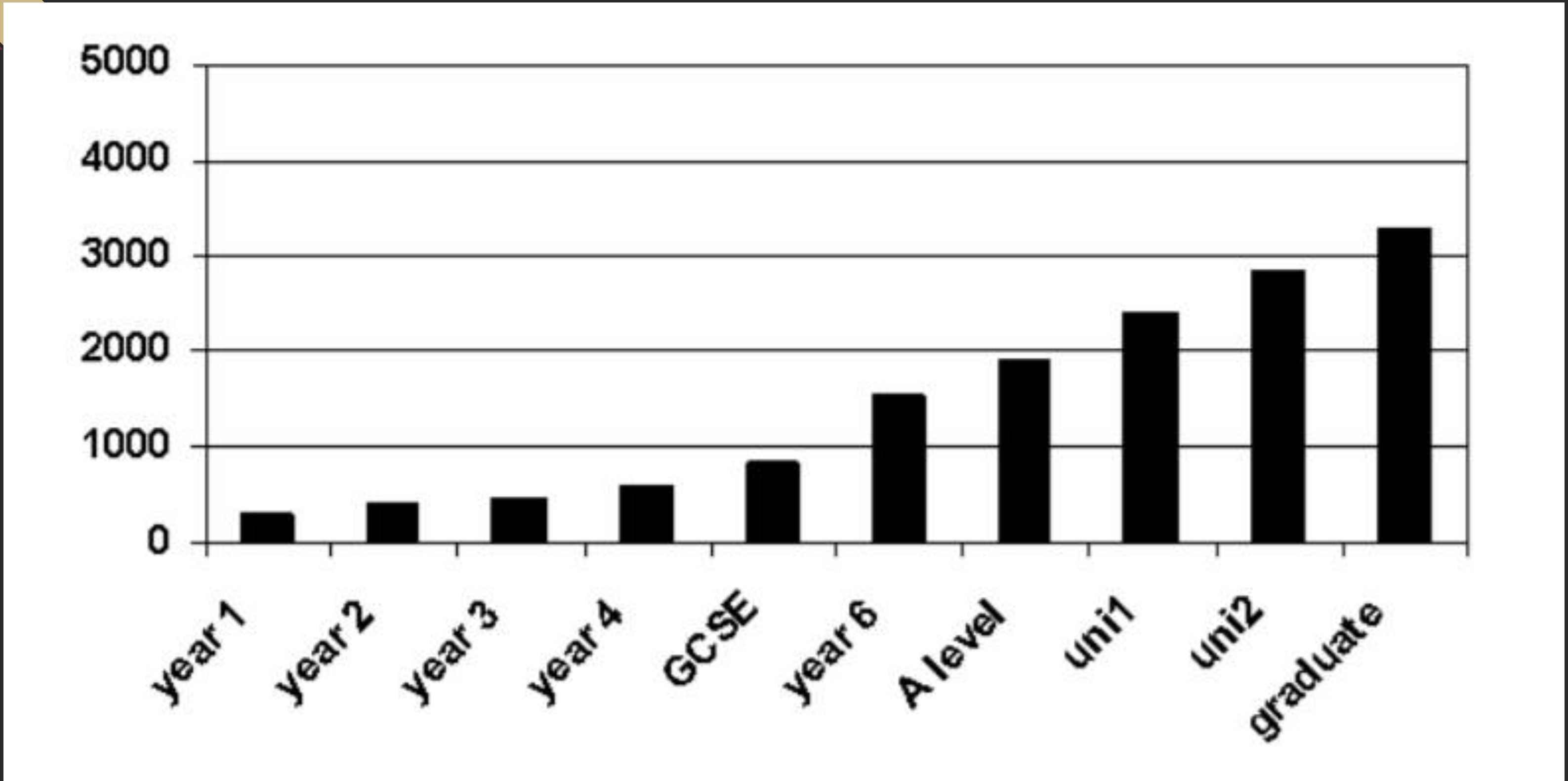
Hinkel 2017. Handbook of Research in Second Language Teaching and Learning.

Non-Emergence in human abilities



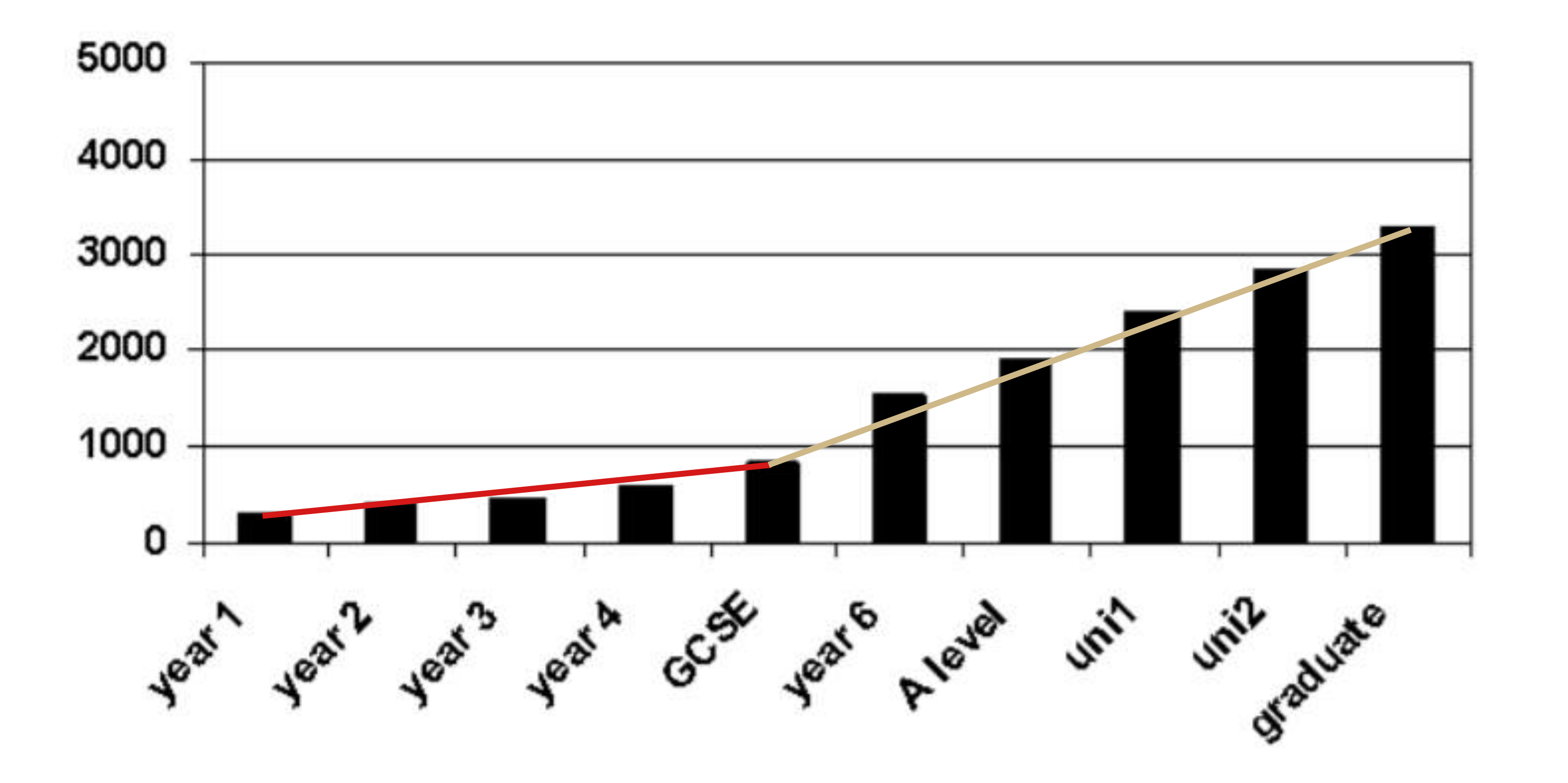
Akase 2022. Longitudinal measurement of growth in vocabulary ...

Non-Emergence in human abilities



Conti 2017. How many new words should you teach per lesson?

Non-Emergence in human abilities



Conti 2017. How many new words should you teach per lesson?

Recap

For the discussion of non-emergence, we can keep in mind:

- Olmo's point about the role of the loss function
- Gordon's point about sudden system collapses

Schaeffer, Miranda, Koyejo 2023

https://arxiv.org/pdf/2304.15004.pdf

Relevance

“What controls which abilities will emerge? What controls when abilities will emerge? How can we make desirable abilities emerge faster, and ensure undesirable abilities never emerge?”

Relevance

"What controls which abilities will emerge? What controls when abilities will emerge? How can we make desirable abilities emerge faster, and ensure undesirable abilities never emerge?"

→ *emergence or not, important questions*

Core aspect

“ [W]e 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.”

Structure

- Introduction
- Alternative Explanation for Emergent Abilities
- Analyzing [GPT]'s Emergent Arithmetic Abilities
- Meta-Analysis of Claimed Emergent Abilities
- Inducing Emergent Abilities in Networks on Vision Tasks
- Related Work
- Discussion

Structure

- Introduction
- Alternative Explanation for Emergent Abilities
- Analyzing [GPT]'s Emergent Arithmetic Abilities
- Meta-Analysis of Claimed Emergent Abilities
- Inducing Emergent Abilities in Networks on Vision Tasks
- Related Work
- Discussion

Starting observation-ish

92% of emergent tasks fall into two categories qua metrics:

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

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

Alternative Explanation

“Linear” baseline model family

Test it on common metrics

v

Test it on alternative metrics

Alternative Explanation

Ingredients:

- “Model family”, **Large Language Models**
- different numbers of parameters $N > 0$
- each model’s per-token cross entropy falls as a power law with the number of parameters N for:
- constant $c > 0$
- constant $\alpha < 0$

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

Alternative Explanation

Ingredients:

- "Model family", **Large Language Models**
- different numbers of parameters $N > 0$
- each model's per-token cross entropy falls as a power law with the number of parameters N for:
- constant $c > 0$
- constant $\alpha < 0$

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

0≤(N/c)alpha ≤1**

Alternative Explanation

Ingredients:

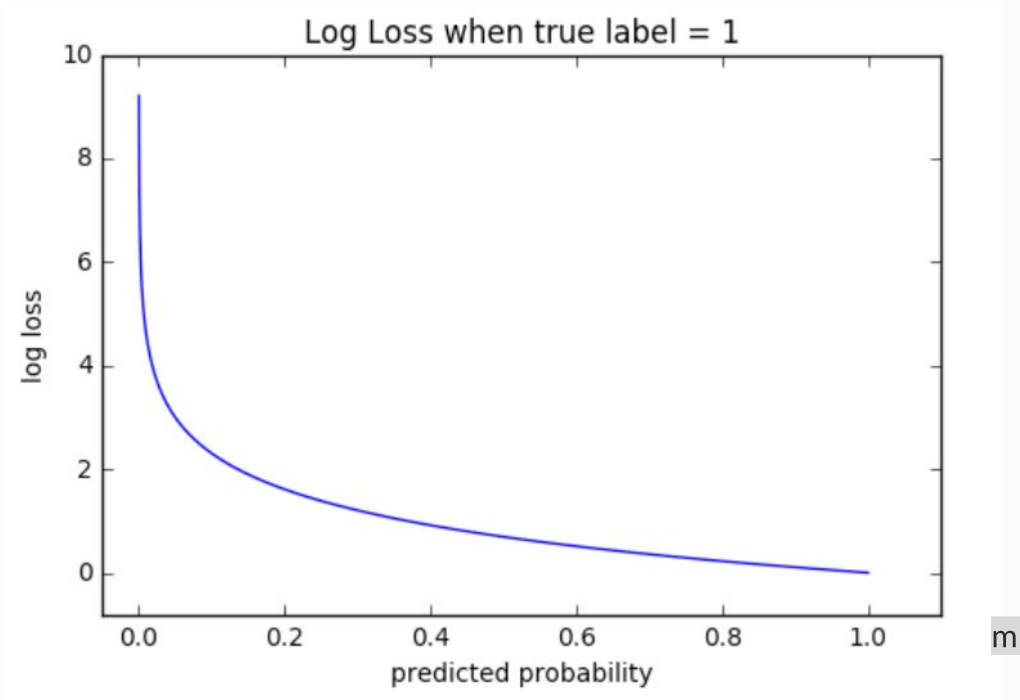
- “Model family”, **Large Language Models**
- different numbers of parameters $N > 0$
- each model’s per-token cross entropy falls as a power law with the number of parameters N for:
- constant $c > 0$
- constant $\alpha < 0$

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

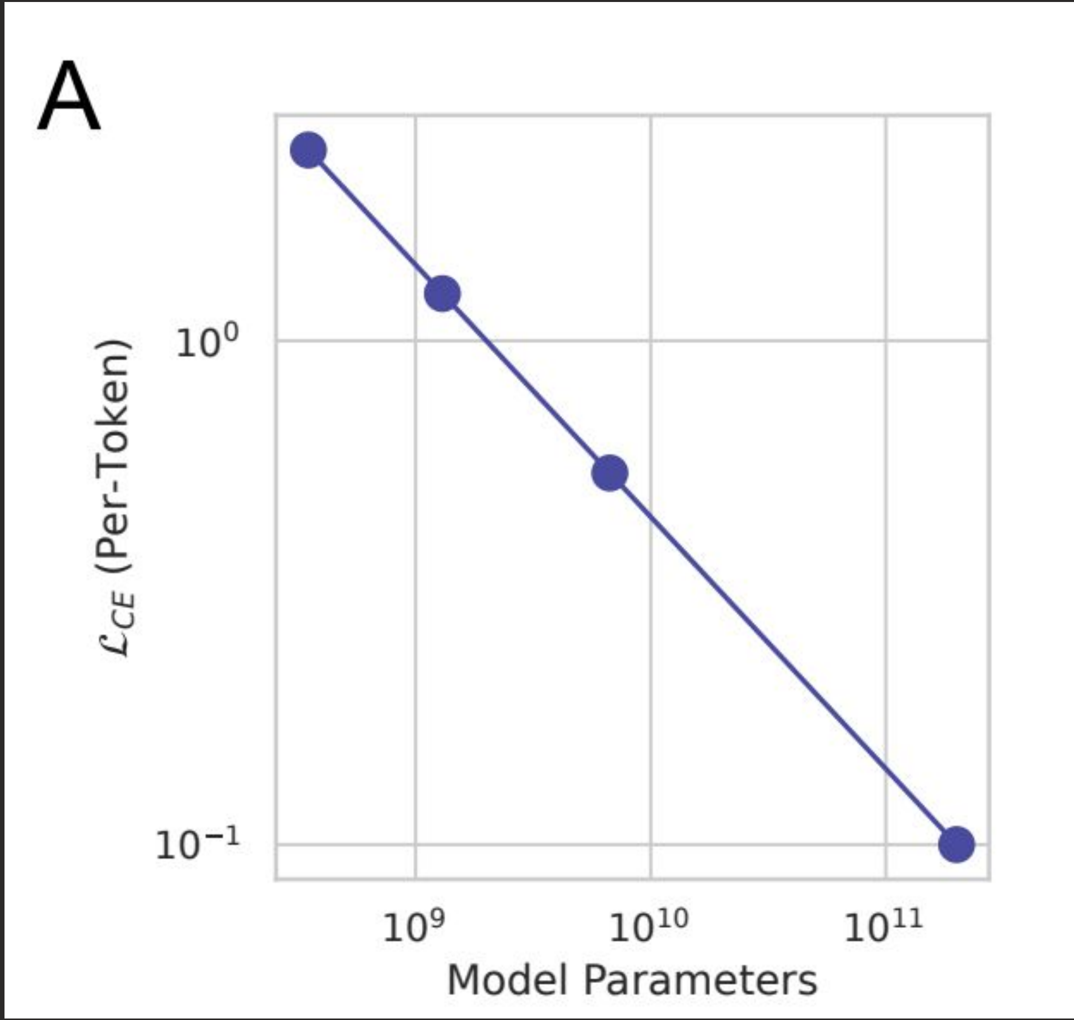
0≤(N/c)**alpha ≤1

Cross-Entropy

Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label. So predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high loss value. A perfect model would have a log loss of 0.



ml-cheatsheet.readthedocs.io



A
$\mathcal{L}_{CE}$ (Per-Token)
10^0
10^{-1}
10^9
10^{10}
10^{11}
Model Parameters

Alternative Explanation

- V: set of possible tokens
- $\mathbf{p} \in \Delta^{|V|-1}$: true but unknown probability distribution
- $\Delta^{|V|-1}$: set of all possible probability distributions over the vocabulary, where each distribution assigns a probability to each word such that the sum of probabilities is 1 (ChatGPT)

Alternative Explanation

Still on establishing a “baseline”:

Alternative Explanation

- $\hat{p}_N \in \Delta^{|V|-1}$: model with N parameters, its predicted probability distribution
- per-token cross entropy as a f(N):

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

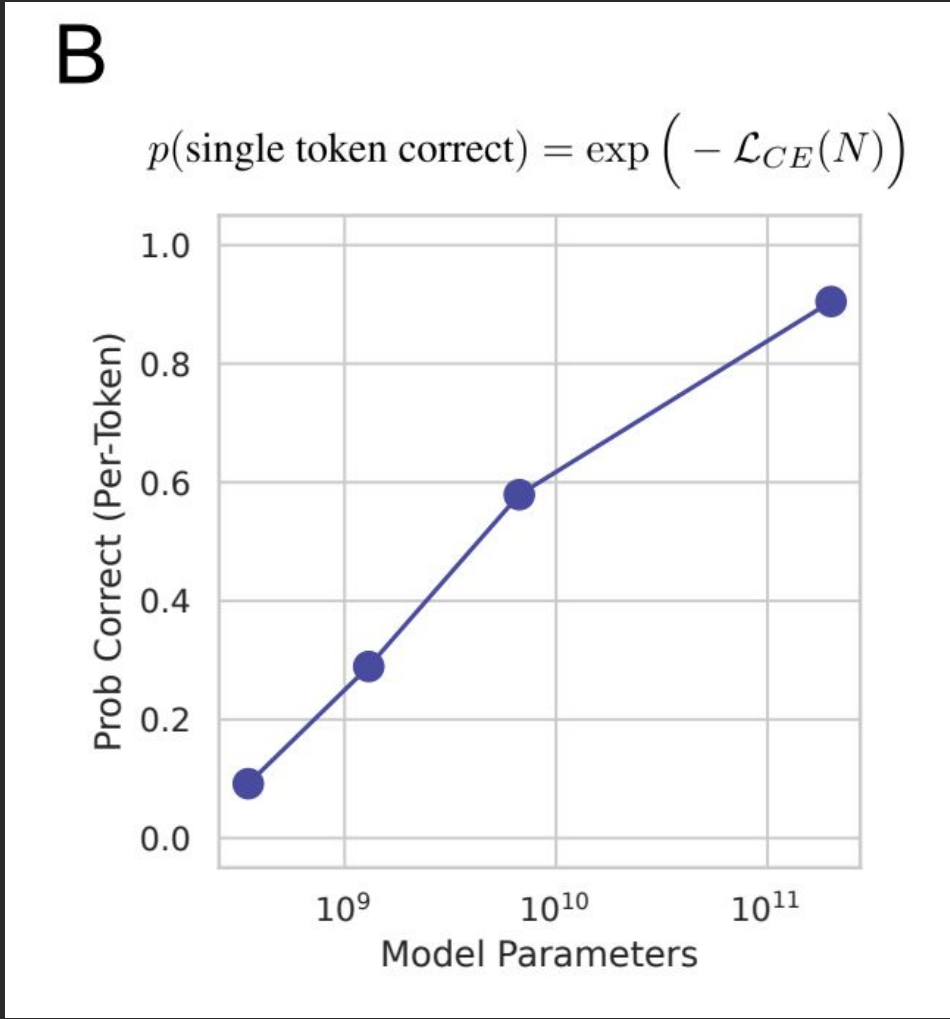
Alternative Explanation

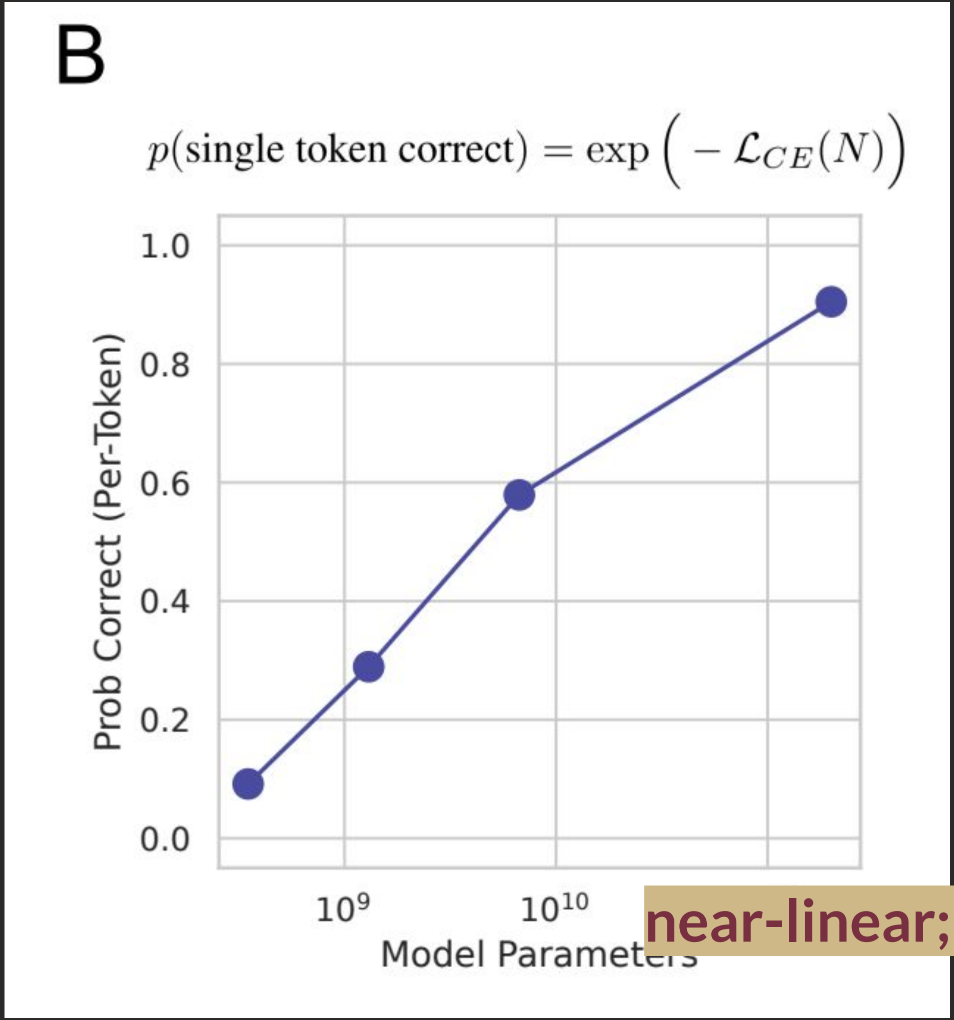
In practice, p is unknown, so we substitute a one-hot distribution of the observed token v^*:

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

A model with N parameters then has a per-token probability of selecting the correct token (Fig. 2B):

$$p(\text{single token correct}) = \exp\Big(-\mathcal{L}_{CE}(N)\Big) = \exp\Big(-(N/c)^{\alpha}\Big)$$


B
$p(\text{single token correct}) = \exp\Big(-\mathcal{L}_{CE}(N)\Big)$
Prob Correct (Per-Token)
1.0
0.8
0.6
0.4
0.2
0.0
10^9
10^{10}
10^{11}
Model Parameters



near-linear; non-emergence

Alternative Explanation

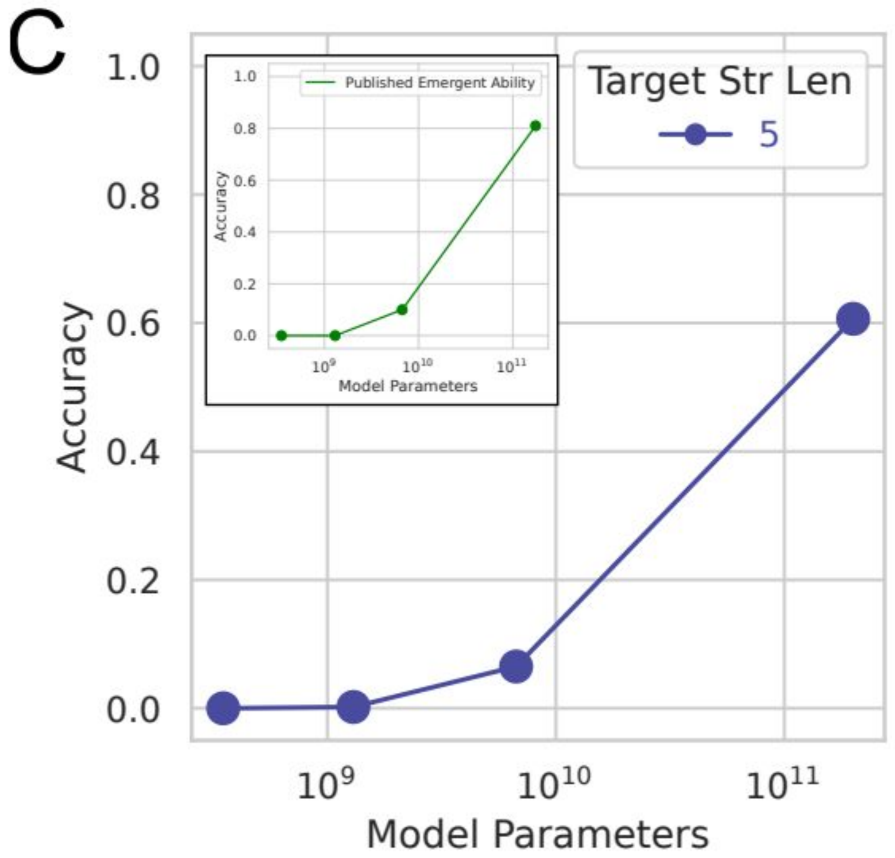
Suppose:

- a metric that requires selecting L tokens correctly
- probability of scoring 1 is*:

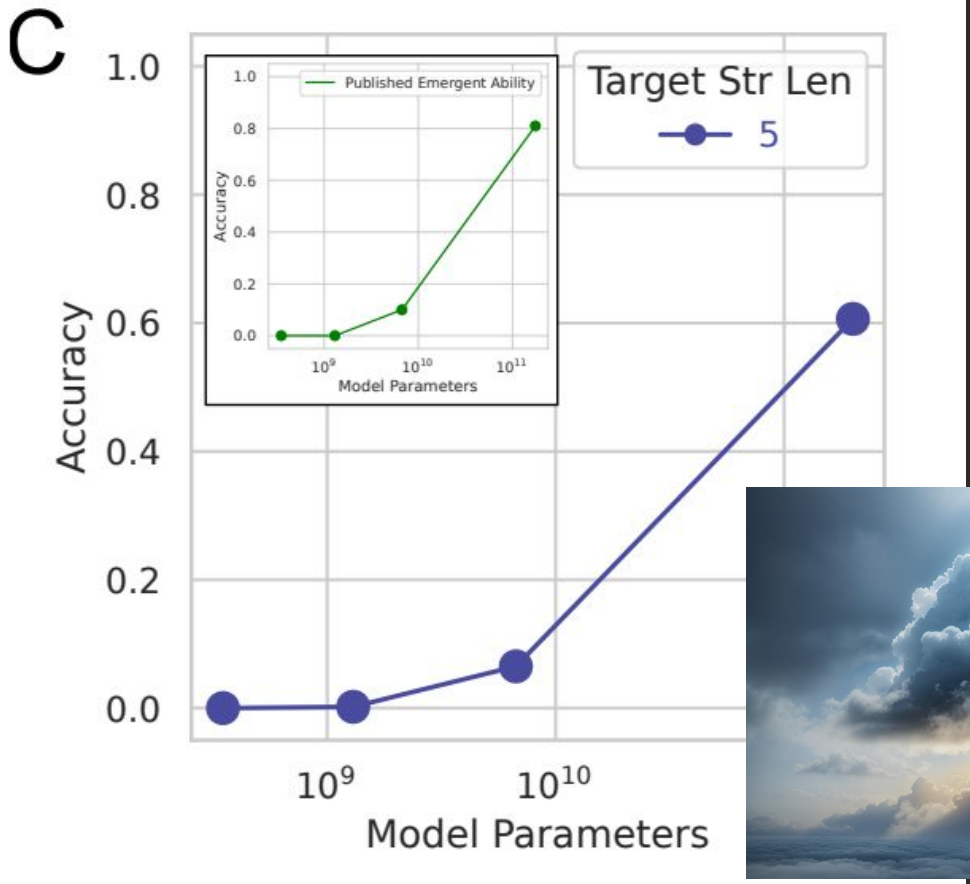
$$\text{Accuracy}(N) \approx p_N(\text{single token correct})^{\text{num. of tokens}} = \exp\left(-(N/c)^{\alpha}\right)^{L}$$

*assuming independence, see FN1

Emergent Abilities



Emergent Abilities



Alternative Explanation

- But change that to another metric, Token Edit Distance:

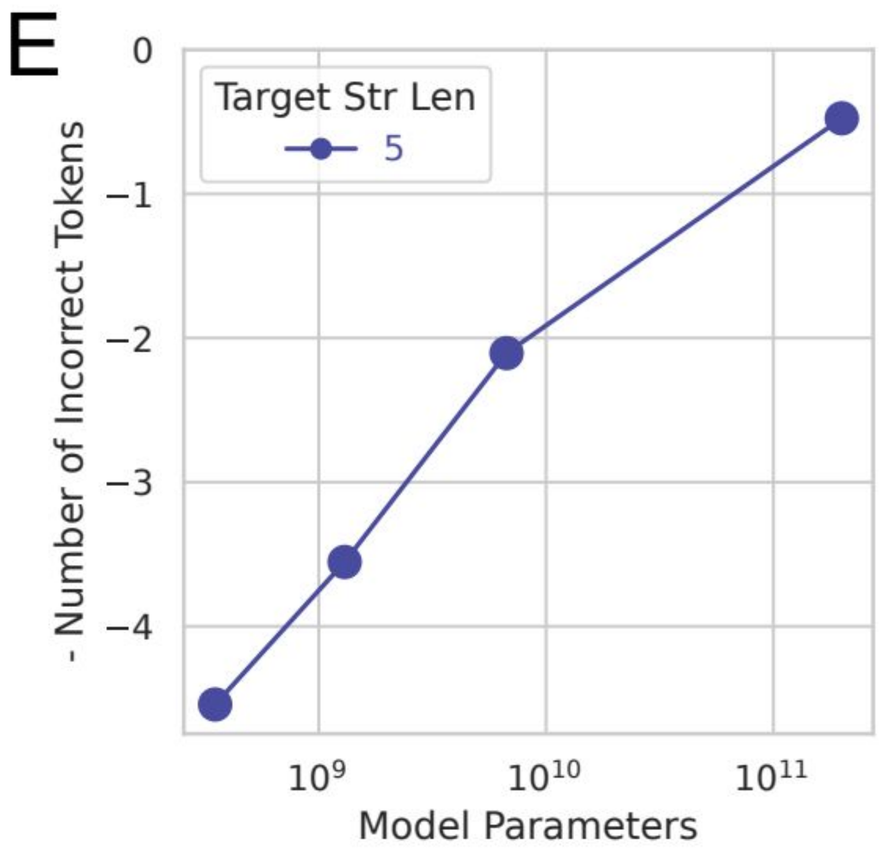
$$\text{Token Edit Distance}(N) \approx L\left(1 - p_N(\text{single token correct})\right) = L\left(1 - \exp\left(-(N/c)^{\alpha}\right)\right)$$

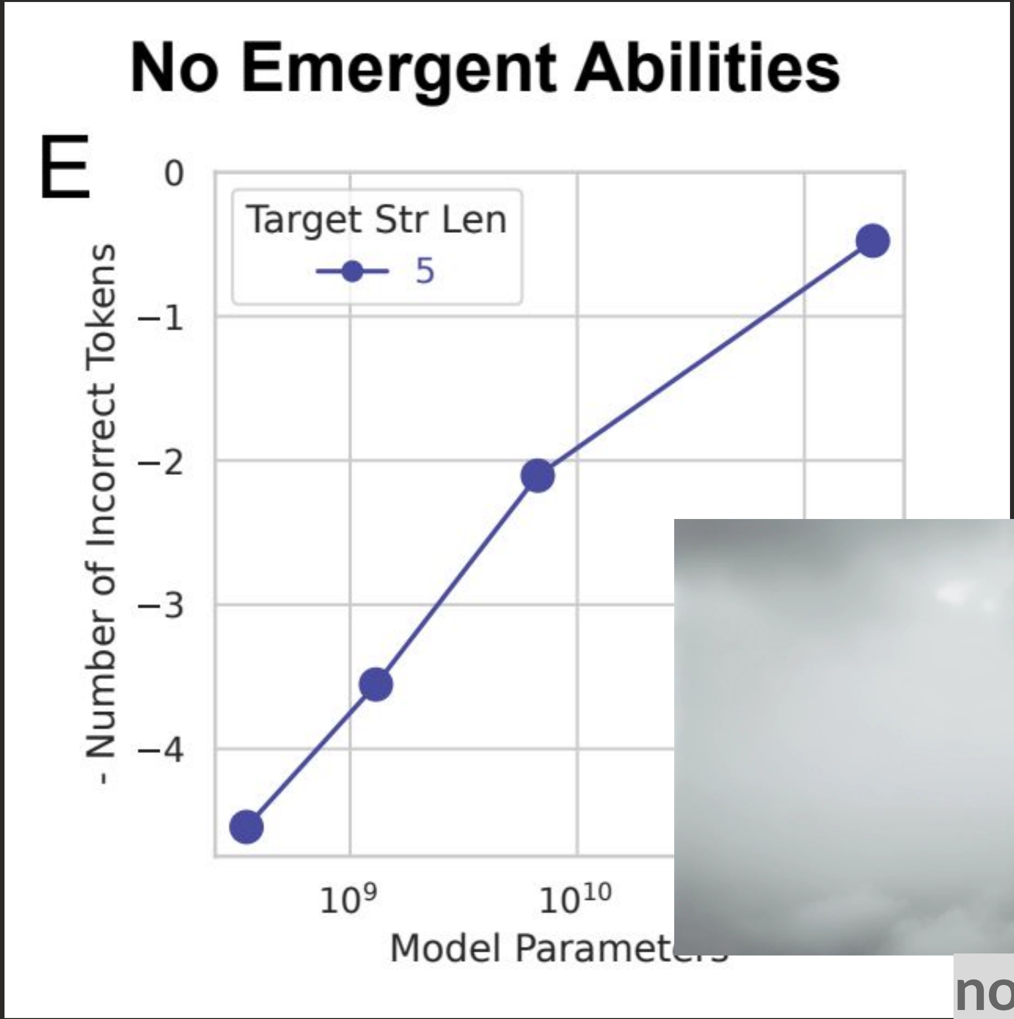
L: … appendix (let's have a look if there is time)

Alternative Explanation

Essentially, changing from Accuracy to something like the Levenshtein distance

No Emergent Abilities

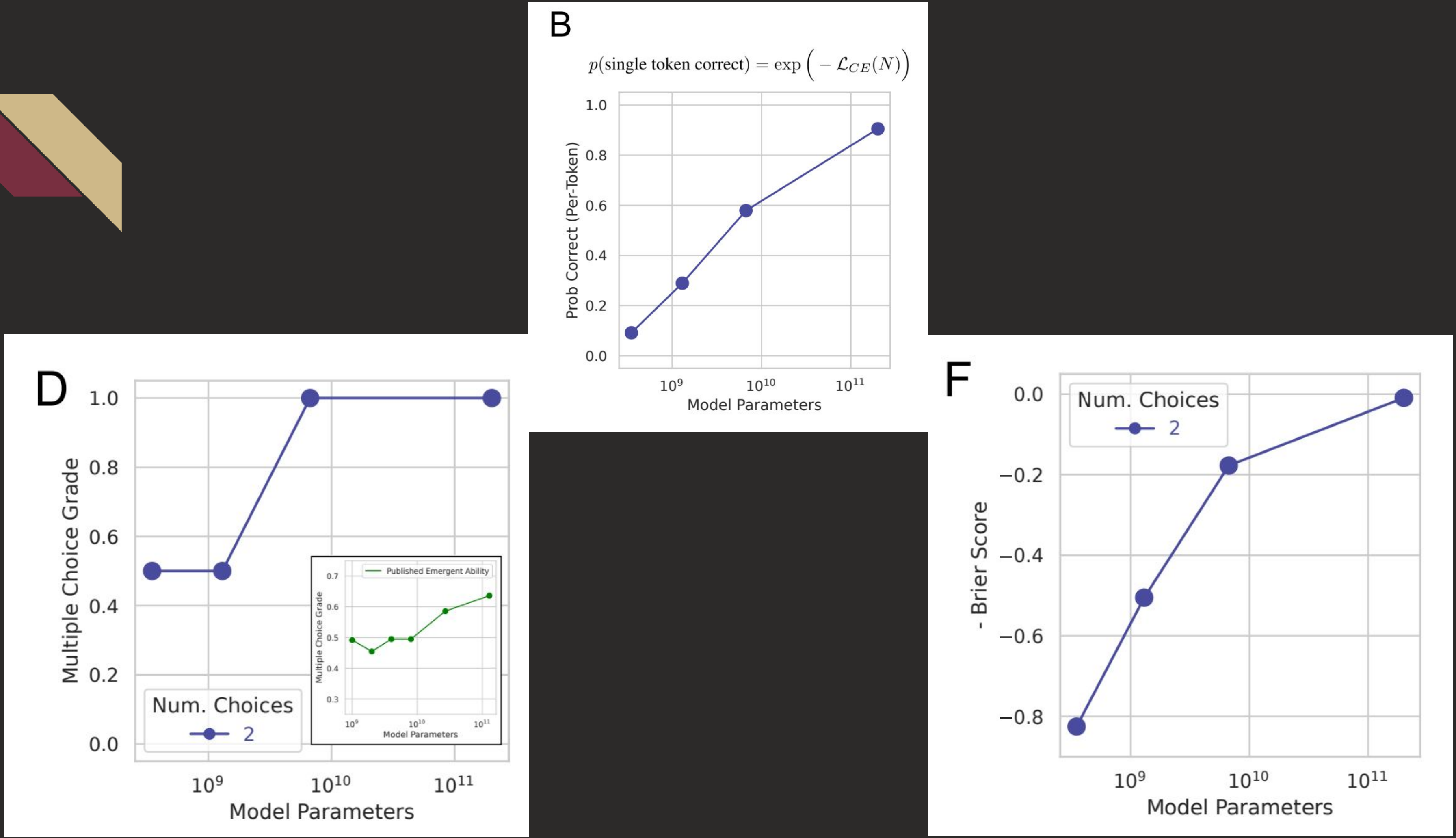


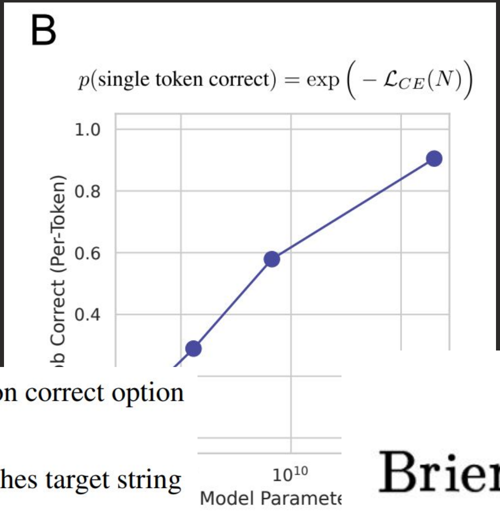


non-emergence

Alternative Explanation

Similarly for Multiple Choice Tasks

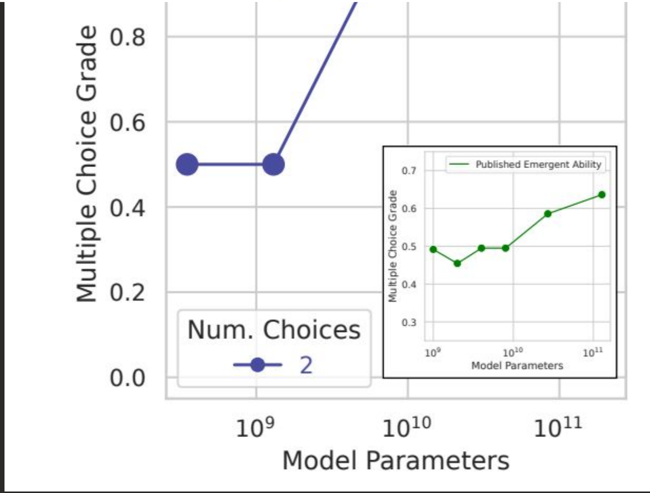

B
$p(\text{single token correct}) = \exp\left(-\mathcal{L}_{CE}(N)\right)$
Prob Correct (Per-Token)
1.0
0.8
0.6
0.4
0.2
0.0
10^9
10^{10}
10^{11}
Model Parameters
D
Multiple Choice Grade
1.0
0.8
0.6
0.4
0.2
0.0
Num. Choices
2
Published Emergent Ability
Multiple Choice Grade
0.7
0.6
0.5
0.4
0.3
10^9
10^{10}
10^{11}
Model Parameters
10^9
10^{10}
10^{11}
Model Parameters
F
- Brier Score
0.0
−0.2
−0.4
−0.6
−0.8
Num. Choices
2
10^9
10^{10}
10^{11}
Model Parameters

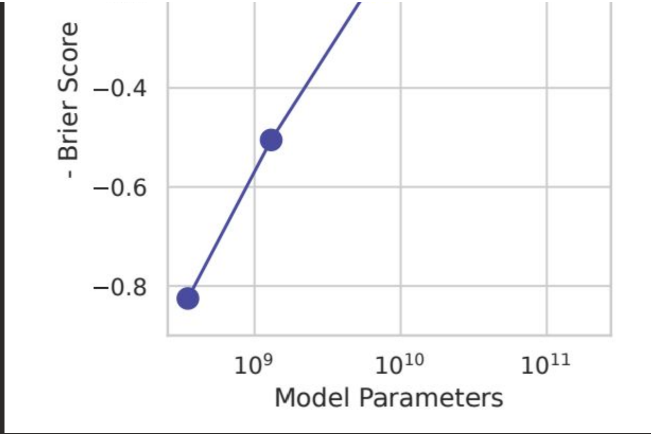


$$\text{Multiple Choice Grade} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$$

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

$$\text{Brier score} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$





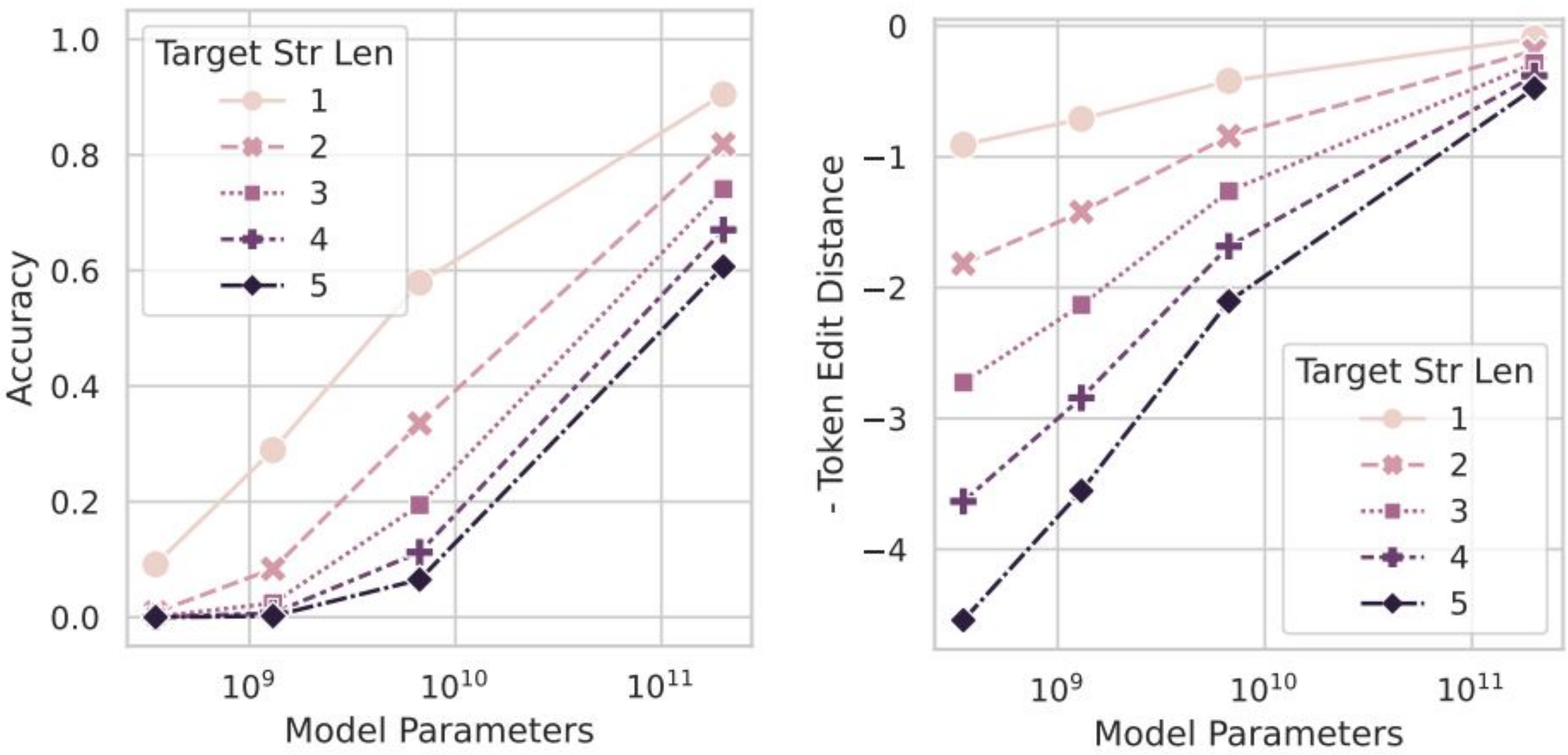
Analysing GPT

Section 3

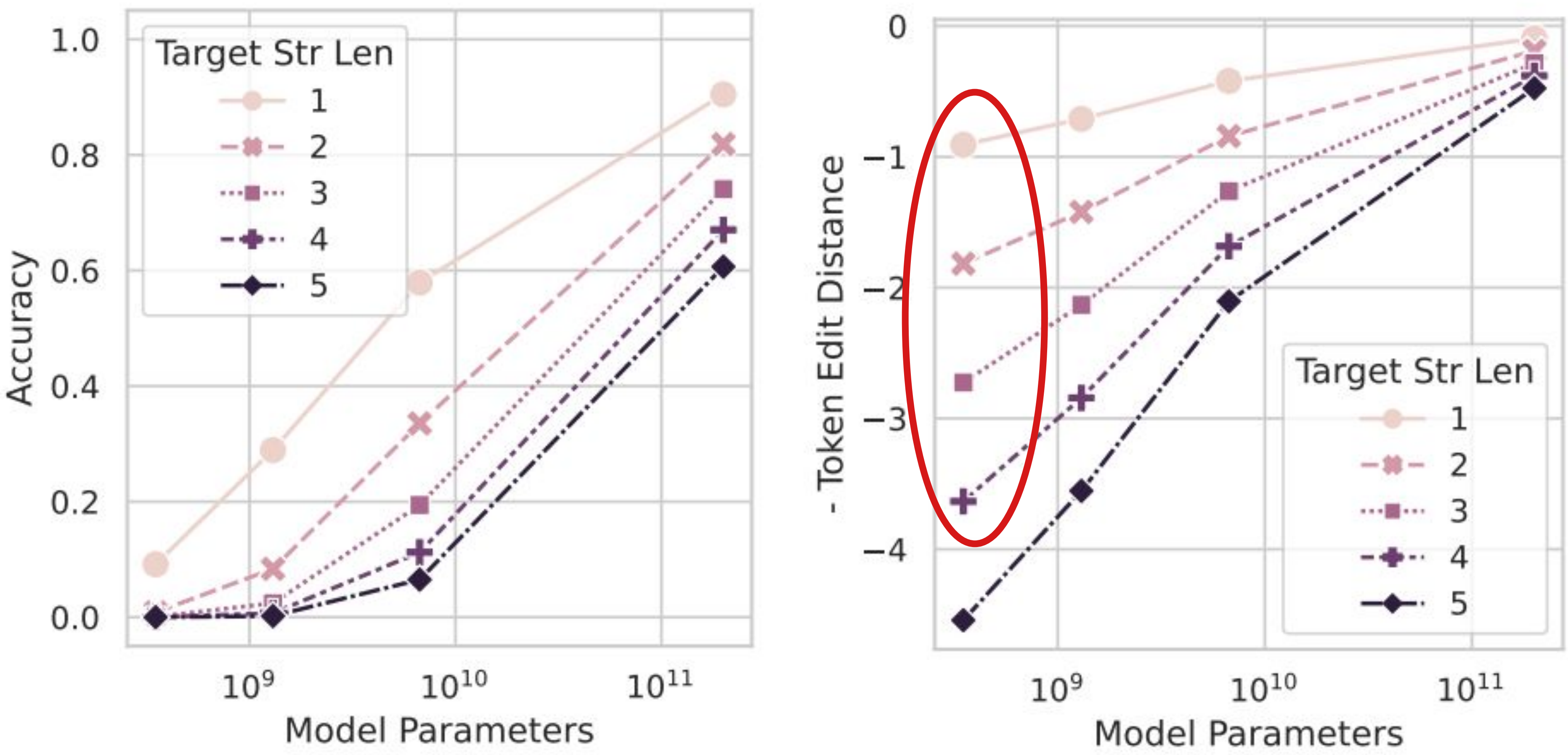
Analyzing InstructGPT/GPT-3’s Emergent Arithmetic Abilities

two tasks:

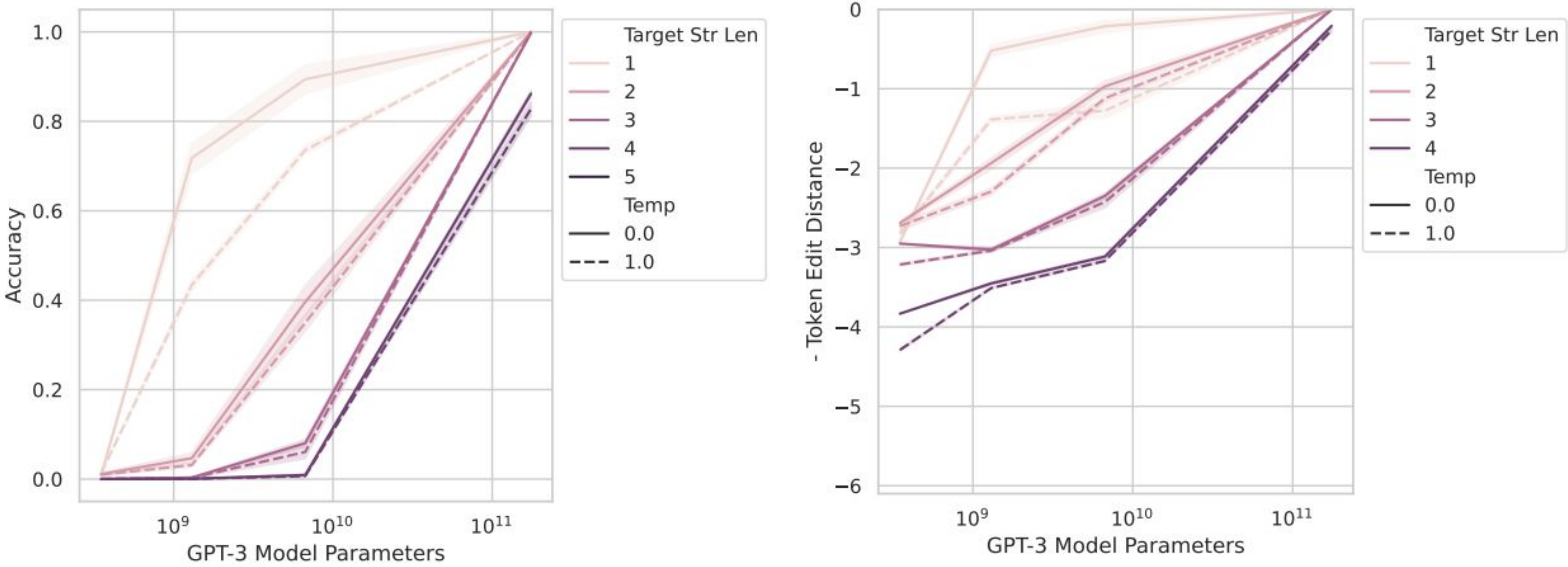
- 2-shot multiplication between two 2-digit integers
- 2-shot addition between two 4-digit integers

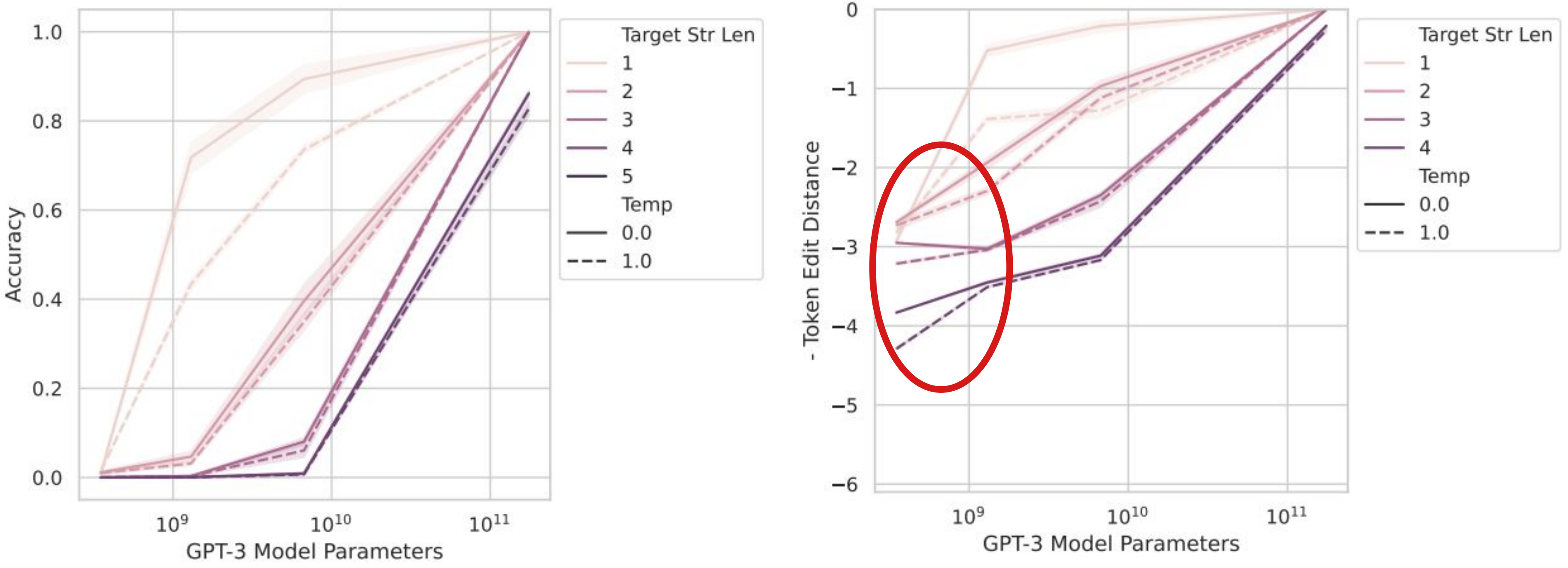


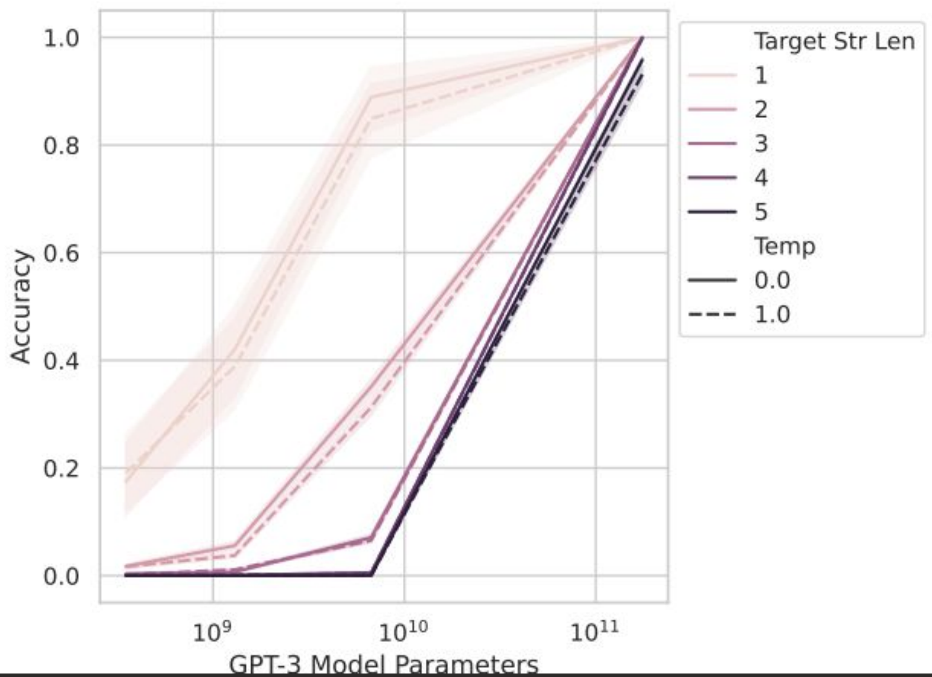
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Target Str Len
1
2
3
4
5
10^9
10^{10}
10^{11}
Model Parameters
- Token Edit Distance
0
−1
−2
−3
−4
Target Str Len
1
2
3
4
5
10^9
10^{10}
10^{11}
Model Parameters



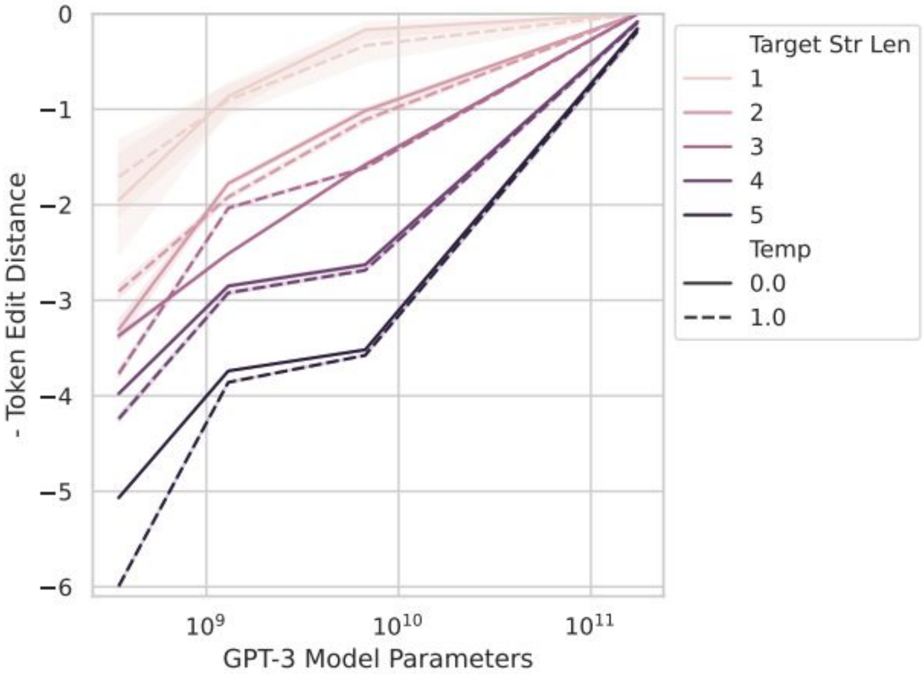
Target Str Len
1
2
3
4
5
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
10^9
10^{10}
10^{11}
Model Parameters
- Token Edit Distance
0
−1
−2
−3
−4
Target Str Len
1
2
3
4
5
10^9
10^{10}
10^{11}
Model Parameters


Accuracy
GPT-3 Model Parameters
Target Str Len
Temp
- Token Edit Distance
GPT-3 Model Parameters

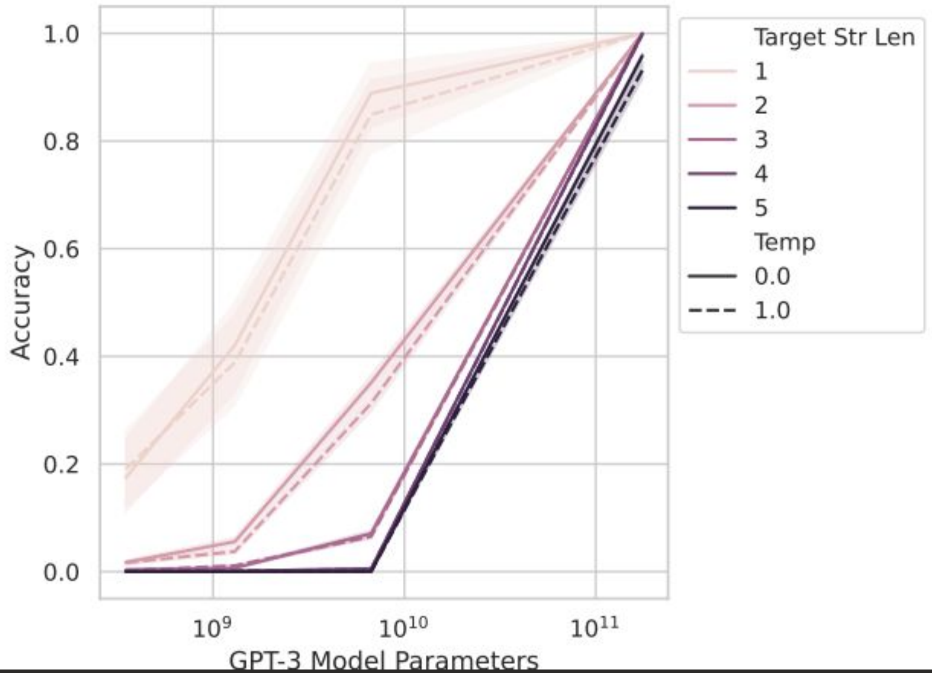

1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
10^9
10^{10}
10^{11}
GPT-3 Model Parameters
Target Str Len
1
2
3
4
5
Temp
0.0
1.0
0
−1
−2
−3
−4
−5
−6
- Token Edit Distance
10^9
10^{10}
10^{11}
GPT-3 Model Parameters
Target Str Len
1
2
3
4
Temp
0.0
1.0

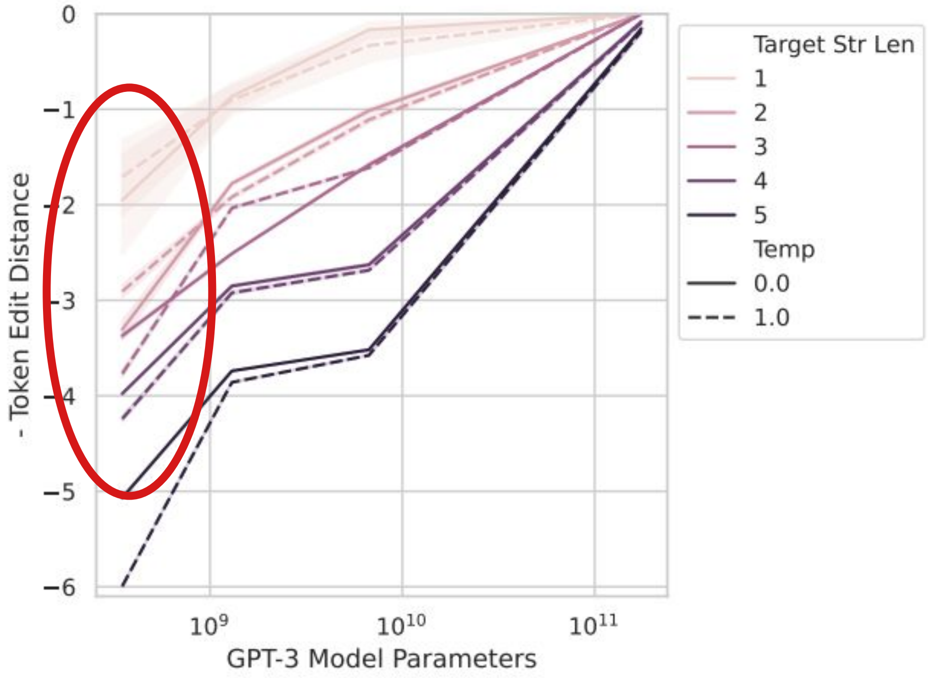


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
10^9
10^{10}
10^{11}
GPT-3 Model Parameters
Target Str Len
1
2
3
4
5
Temp
0.0
1.0



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


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
10^9
10^{10}
10^{11}
GPT-3 Model Parameters
Target Str Len
1
2
3
4
5
Temp
0.0
1.0


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

Meta-Analysis

Section 4

Meta-Analysis of Claimed Emergent Abilities

Categorise tasks of BIG-Bench (collection of ML benchmarks)

Meta-Analysis

Emergence criterion:

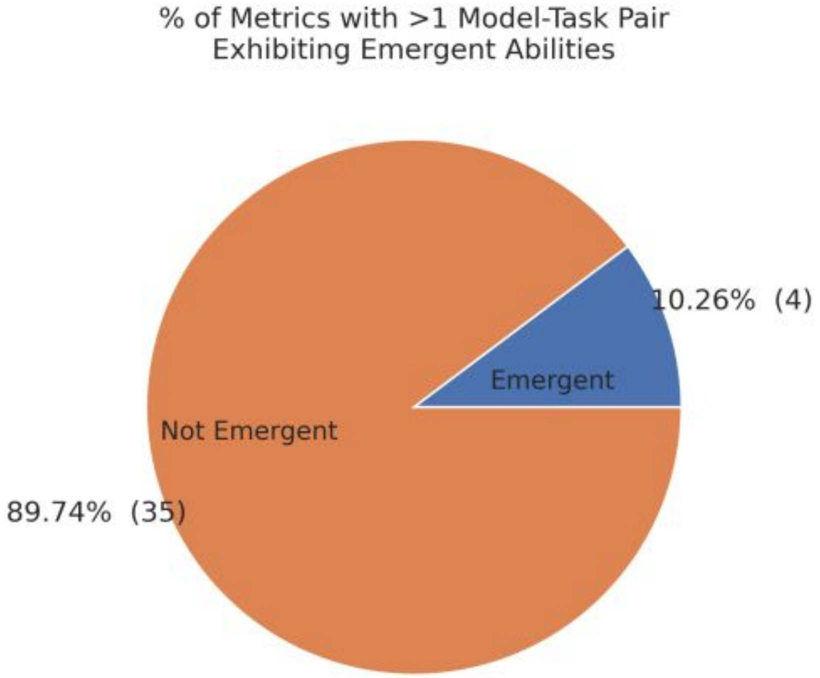
Letting $y_i \in \mathbb{R}$ denote model performance at model scales $x_i \in \mathbb{R}$, sorted such that $x_i < x_{i+1}$, the emergence score is:

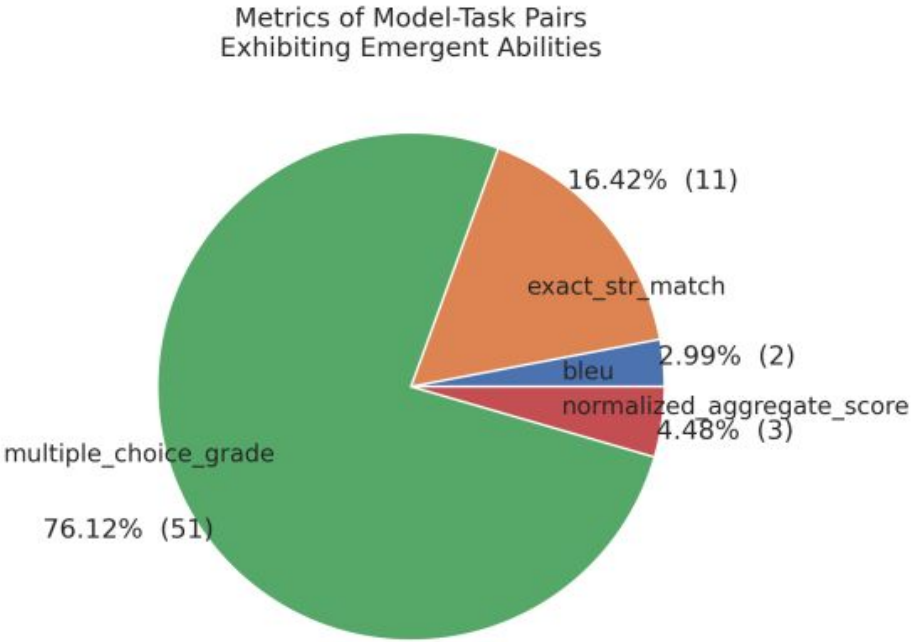
$$\text{Emergence Score}\left(\left\{(x_n, y_n)\right\}_{n=1}^{N}\right) \stackrel{\text{def}}{=} \frac{\text{sign}(\arg\max_i y_i - \arg\min_i y_i)(\max_i y_i - \min_i y_i)}{\sqrt{\text{Median}(\{(y_i - y_{i-1})^2\}_i)}} \tag{1}$$

Meta-Analysis

→ 92% figure

92% of tasks where emergence was observed use accuracy or multiple choice grade metrics

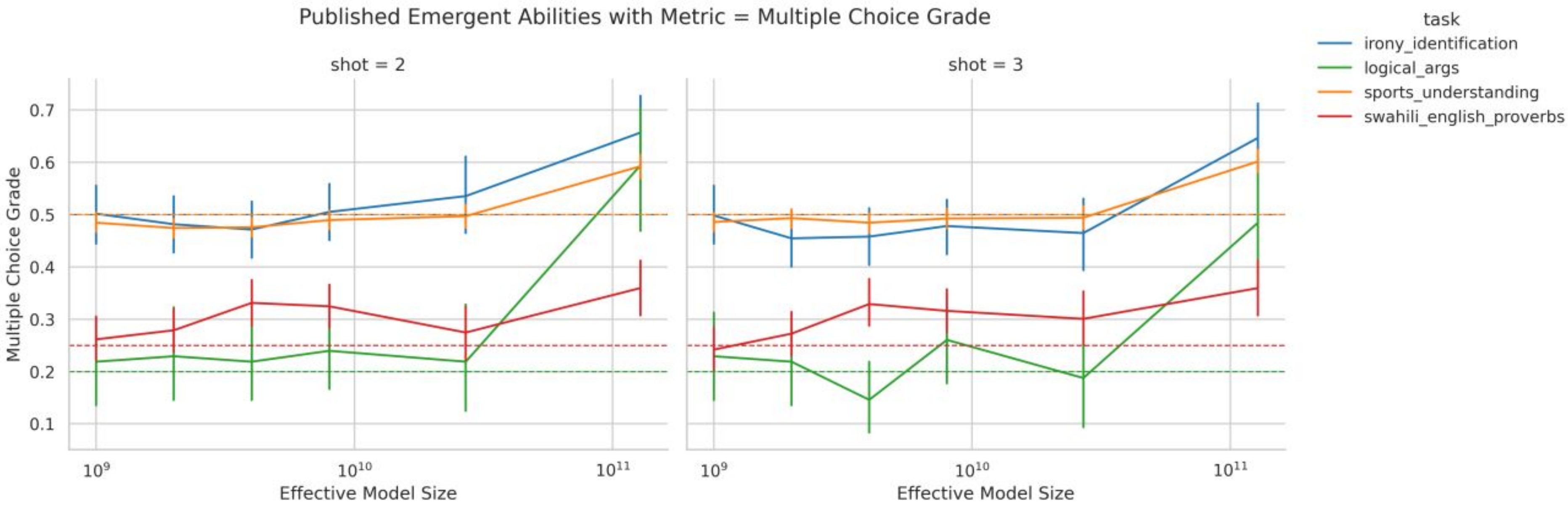

% of Metrics with >1 Model-Task Pair
Exhibiting Emergent Abilities
10.26% (4)
Emergent
Not Emergent
89.74% (35)


Metrics of Model-Task Pairs
Exhibiting Emergent Abilities
16.42% (11)
exact_str_match
2.99% (2)
bleu
normalized_aggregate_score
4.48% (3)
multiple_choice_grade
76.12% (51)

Meta-Analysis

We can look at the paper if there is time

Where emergence was observed:

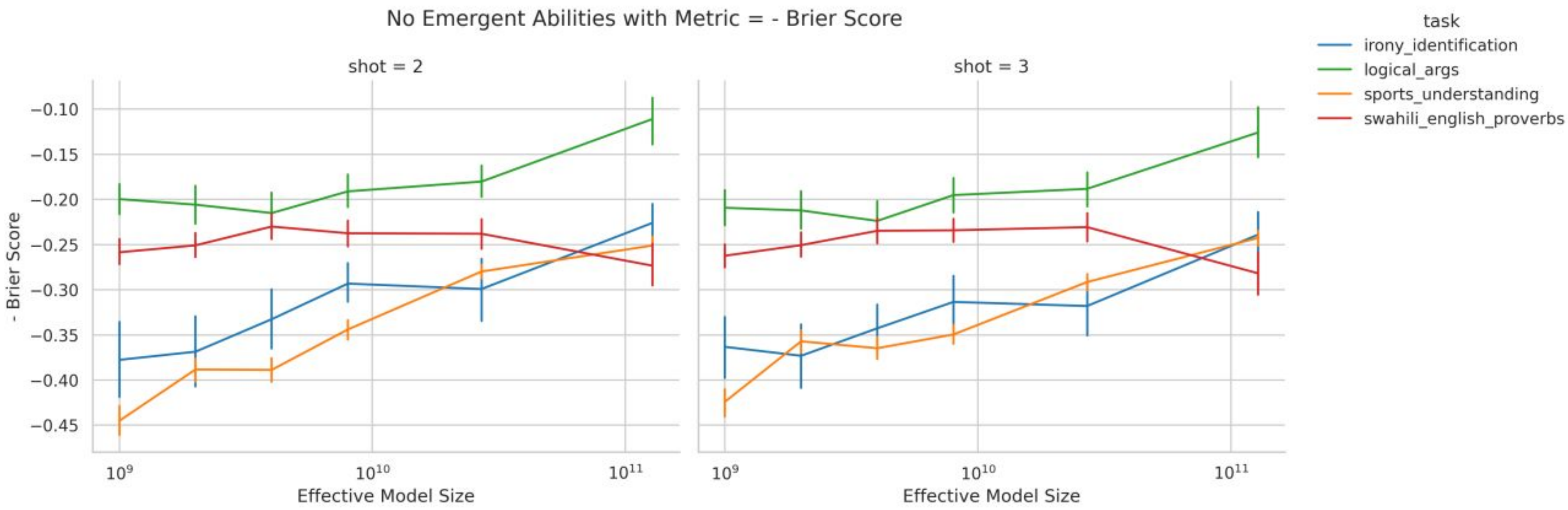

Published Emergent Abilities with Metric = Multiple Choice Grade
shot = 2
shot = 3
Multiple Choice Grade
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10^9
10^{10}
10^{11}
Effective Model Size
task
irony_identification
logical_args
sports_understanding
swahili_english_proverbs


Published Emergent Abilities with Metric = Multiple Choice Grade
shot = 2
shot = 3
task
irony_identification
logical_args
sports_understanding
swahili_english_proverbs
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

Meta-Analysis

Changing the metric makes the emergence go away:


No Emergent Abilities with Metric = - Brier Score
shot = 2
shot = 3
- Brier Score
−0.10
−0.15
−0.20
−0.25
−0.30
−0.35
−0.40
−0.45
10^9
10^{10}
10^{11}
Effective Model Size
task
irony_identification
logical_args
sports_understanding
swahili_english_proverbs


No Emergent Abilities with Metric = - Brier Score
shot = 2
shot = 3
task
irony_identification
logical_args
sports_understanding
swahili_english_proverbs
- Brier Score
−0.10
−0.15
−0.20
−0.25
−0.30
−0.35
−0.40
−0.45
10^9
10^{10}
10^{11}
Effective Model Size

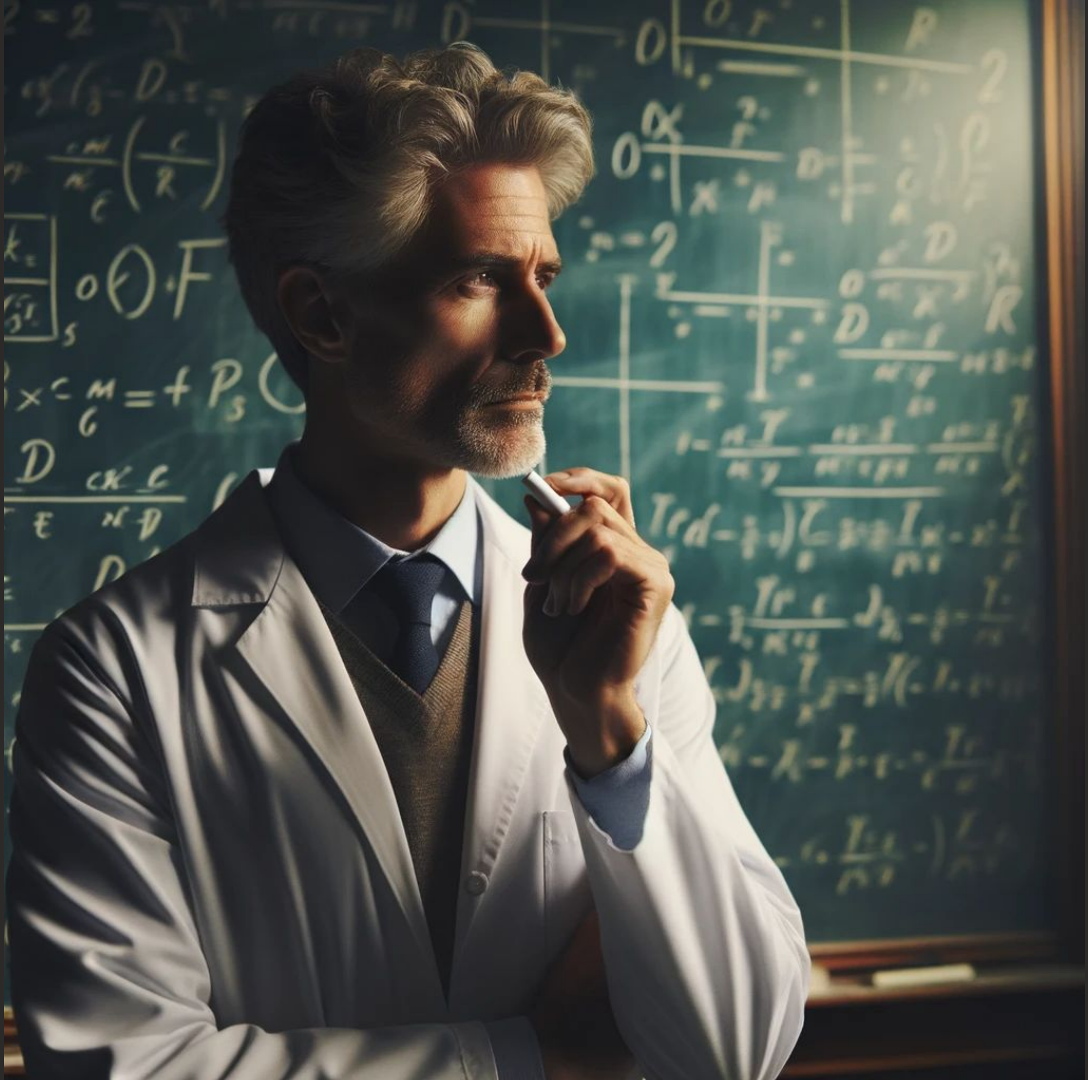
Schaeffer et al. 2023

Their arguments concern the observed emergence, not emergence in general.

Absence of evidence is not evidence of absence.

Schaeffer et al. 2023

“Ergo, emergent abilities may be creations of the researcher’s choices, not a fundamental property of the model family on the specific task. We emphasize that nothing in this paper should be interpreted as claiming that large language models cannot display emergent abilities; rather, our message is that previously claimed emergent abilities in [3, 8, 28, 33] might likely be a mirage induced by researcher analyses.”



The big picture

What are the wider implications?

For ML, and/or possibly our own research?


THANKS TO TO HUMAINITY


"Thentt you, Husmanrty


AI
Thank you,
for being
usseful.

Thank you