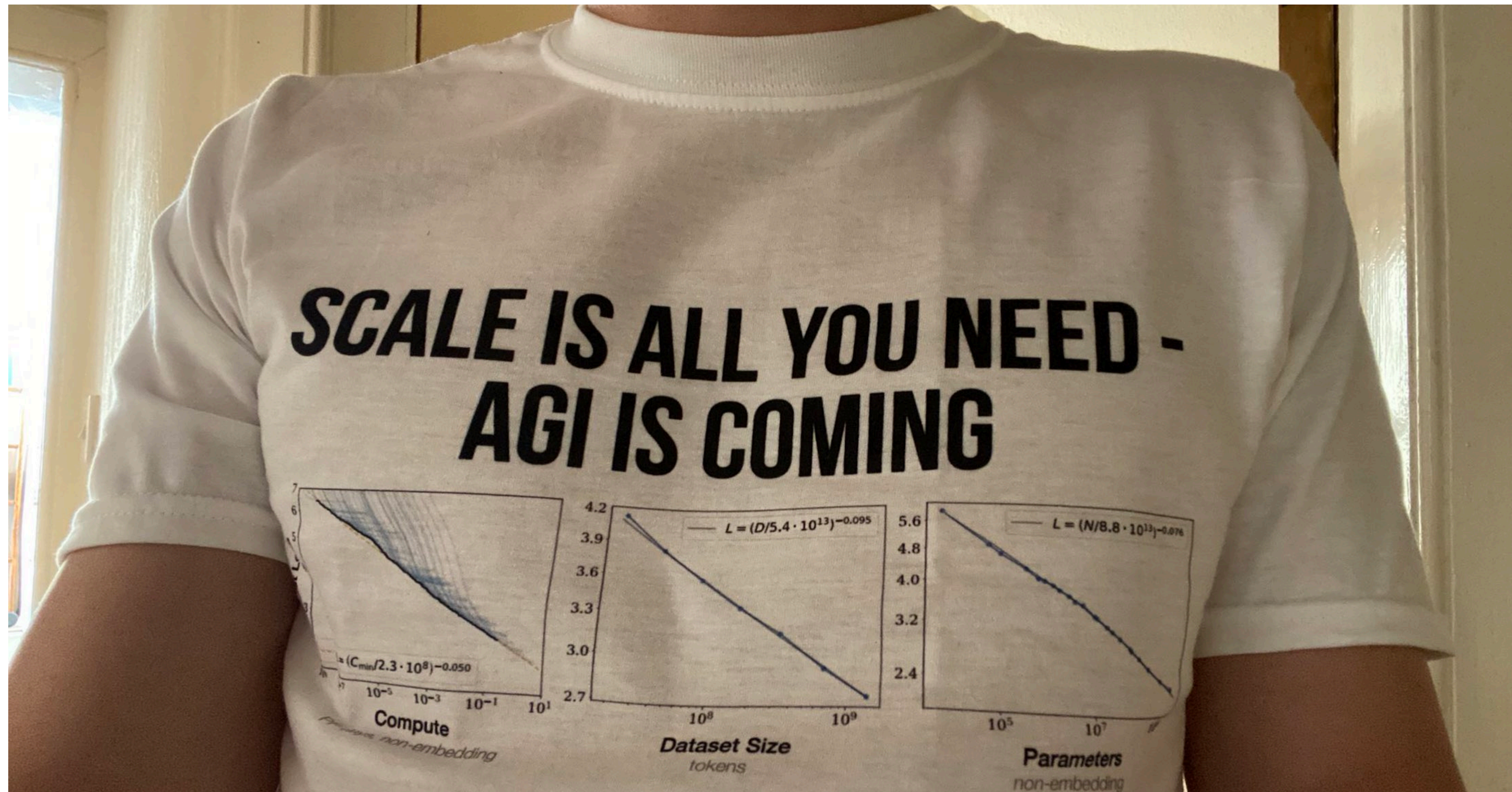


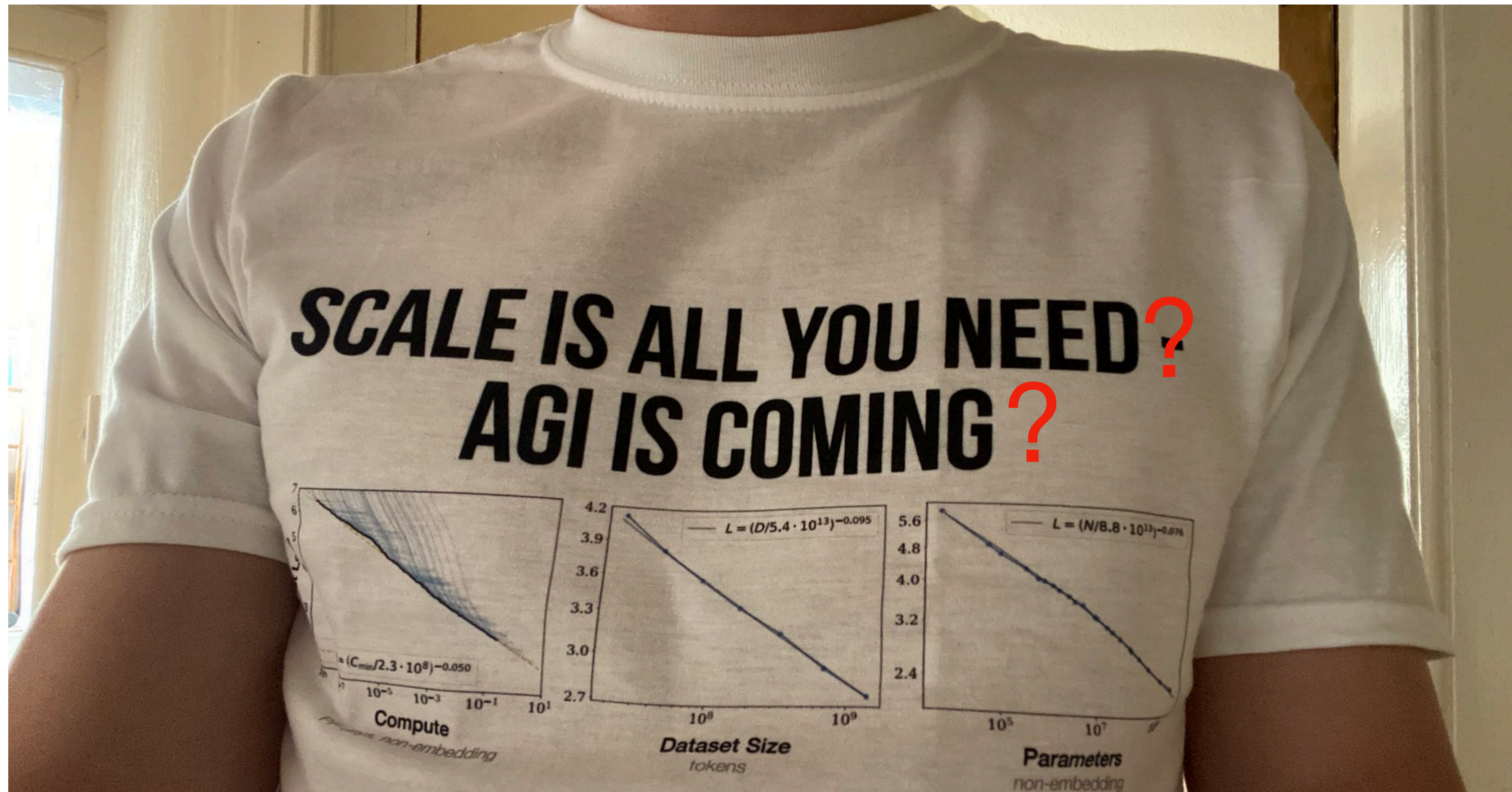
The Tale of Neural Scaling Laws

Speaker: Ziming Liu, April 2023



The Tale of Neural Scaling Laws

Speaker: Ziming Liu, April 2023



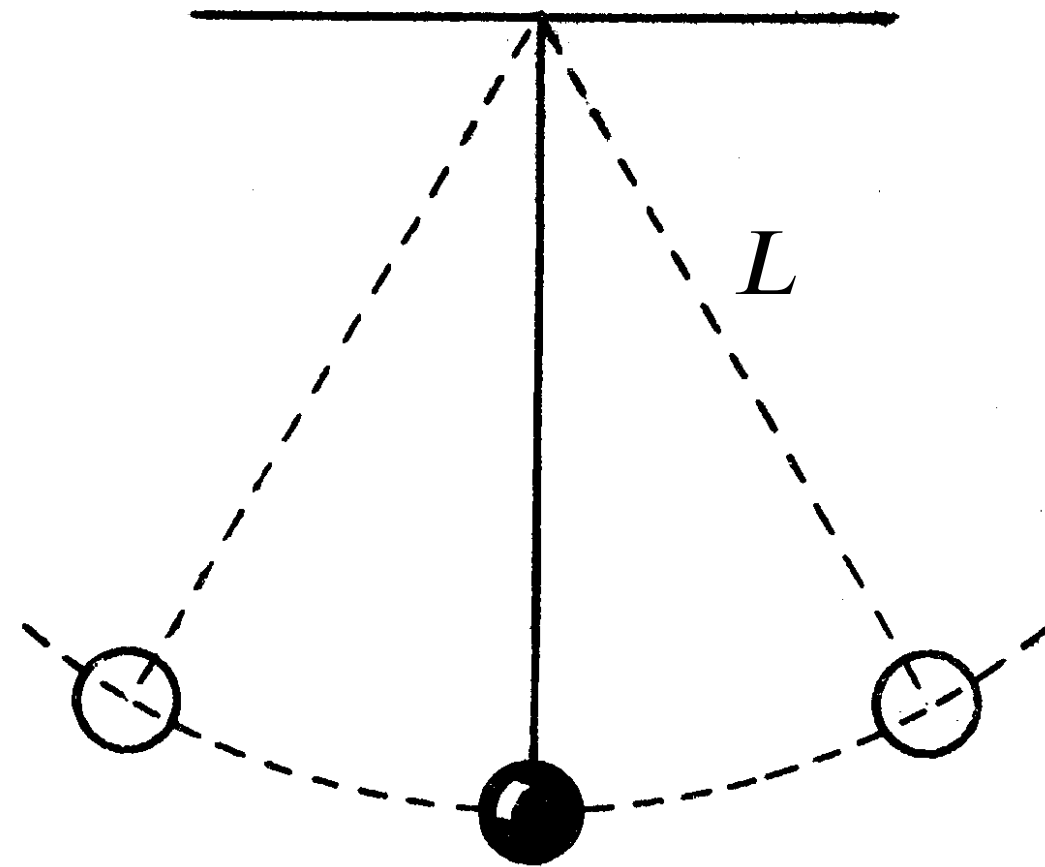
Overview

- Empirical evidence for neural scaling laws and other exotic behavior
- Two theories of neural scaling laws: “classical” and “quantum”
- Panel discussion: AI (LLM) and Science

Empirical Evidence of Neural Scaling Laws and Other Exotic Behaviour

Scaling Laws in science

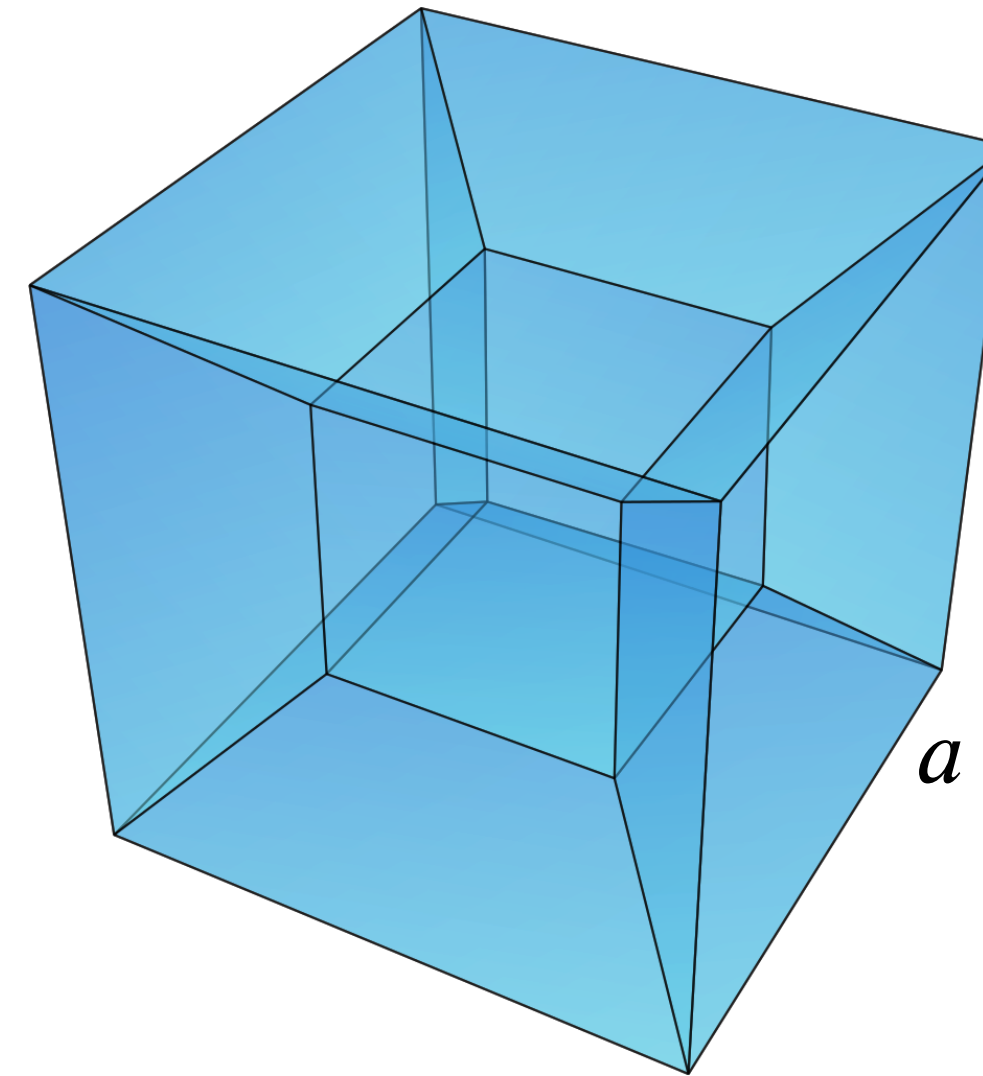
Mechanical



Period of pendulum:

$$T = 2\pi \sqrt{\frac{L}{g}} \propto L^{\frac{1}{2}}$$

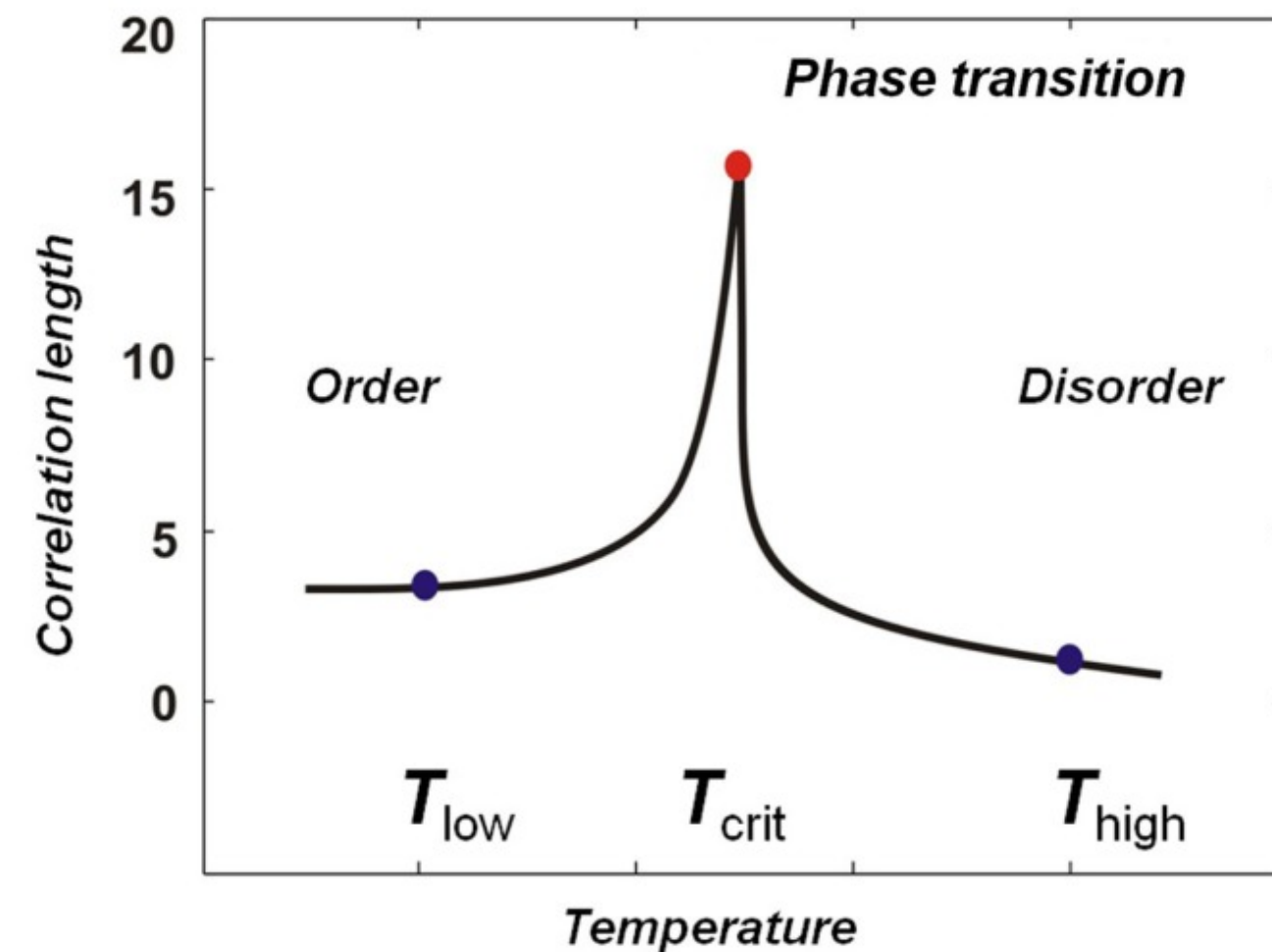
Geometrical



Volume of hypercube:

$$V = a^d$$

Statistical

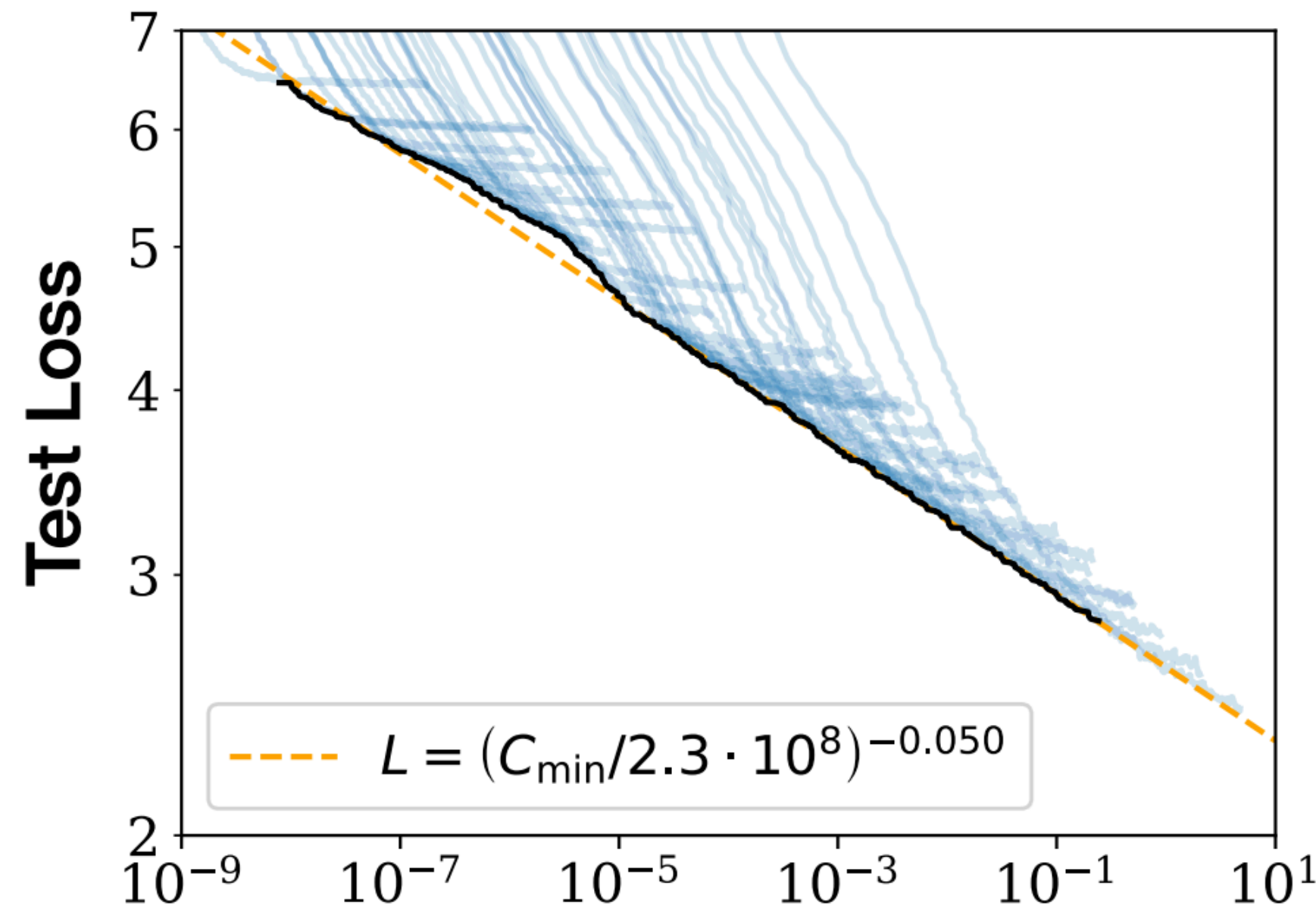


Correlation function: $\langle s_x s_y \rangle = G(|x - y|)$

Far from criticality: $G(|x - y|) \propto \exp\left(-\frac{|x - y|}{\xi}\right)$, ξ correlation length

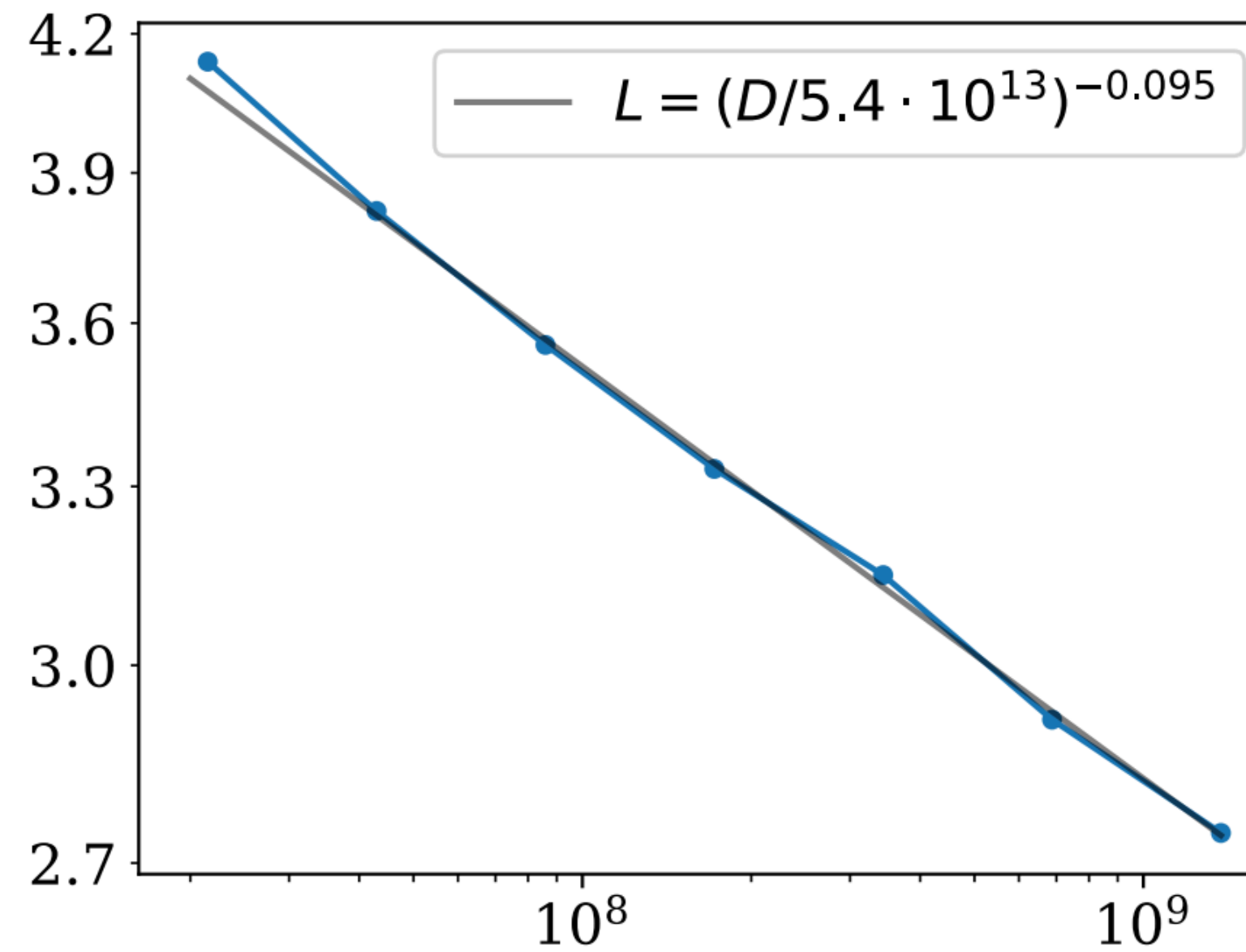
Near criticality: $G(|x - y|) \propto |x - y|^{-\alpha}$

Neural Scaling Laws (NSL)



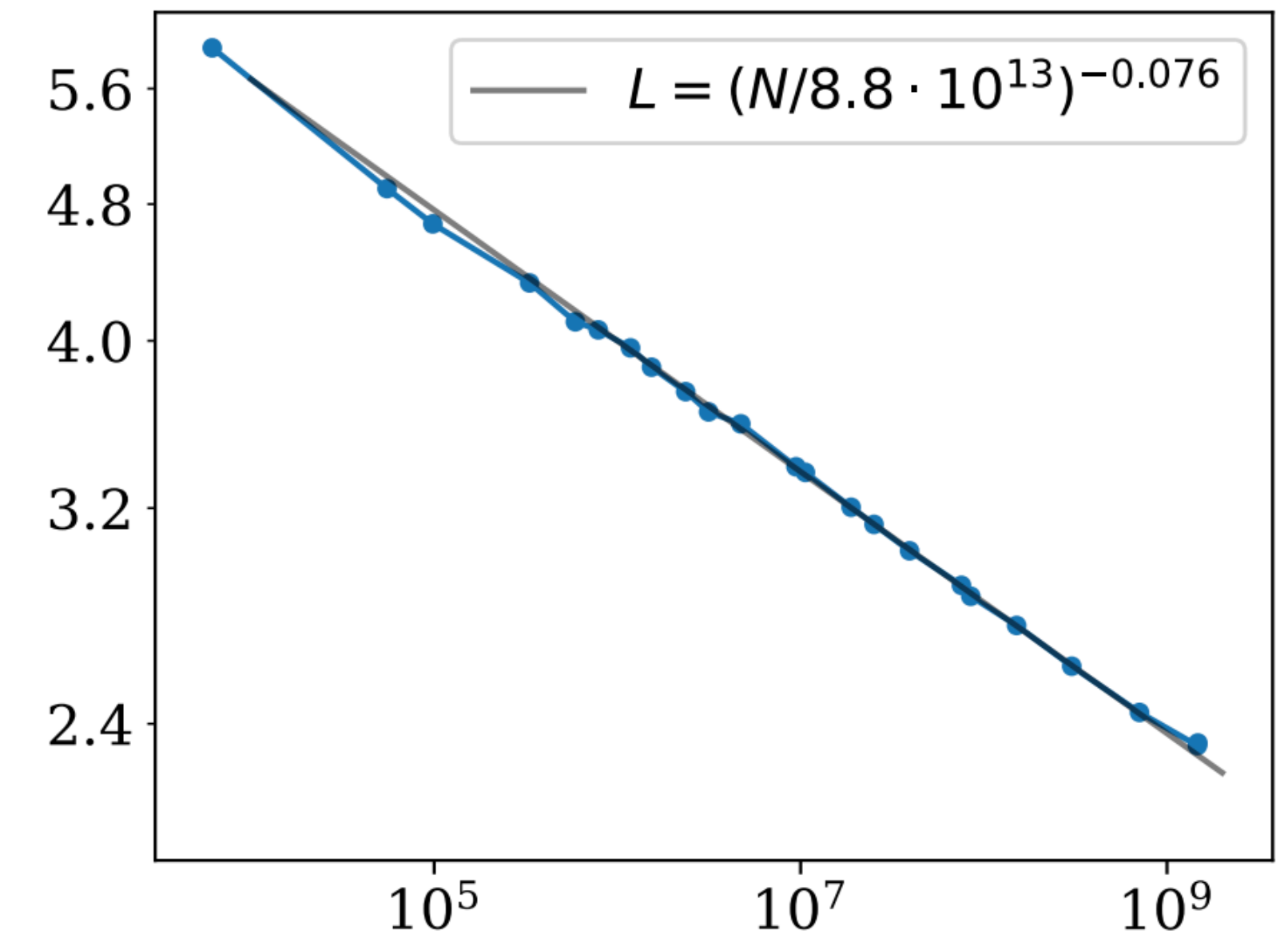
Compute

PF-days, non-embedding



Dataset Size

tokens



Parameters

non-embedding

arXiv > cs > arXiv:2001.08361

Search...

Help | Advance

Computer Science > Machine Learning

[Submitted on 23 Jan 2020]

Scaling Laws for Neural Language Models

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei

NSL is not complete

arXiv > cs > arXiv:2201.02177

1. Grokking

Computer Science > Machine Learning

[Submitted on 6 Jan 2022]

Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra

arXiv > cs > arXiv:1912.02292

2. Double descent

Computer Science > Machine Learning

[Submitted on 4 Dec 2019]

Deep Double Descent: Where Bigger Models and More Data Hurt

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, Ilya Sutskever

arXiv > cs > arXiv:2210.14891

3. Broken scaling laws

Computer Science > Machine Learning

[Submitted on 26 Oct 2022 (v1), last revised 27 Mar 2023 (this version, v10)]

Broken Neural Scaling Laws

Ethan Caballero, Kshitij Gupta, Irina Rish, David Krueger

arXiv > cs > arXiv:2206.07682

4. Emergent abilities

Computer Science > Computation and Language

[Submitted on 15 Jun 2022 (v1), last revised 26 Oct 2022 (this version, v2)]

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

5. Inverse scaling

irmckenzie.co.uk/round2

Inverse Scaling Prize: Second Round Winners

At the end of the second and final round of the [Inverse Scaling Prize](#), we're awarding 7 more Third Prizes. The Prize aimed to identify important tasks on which language models (LMs) perform worse the larger they are ("inverse scaling"). Inverse scaling may reveal cases where LM training actively encourages behaviors that are misaligned with human preferences. The contest started on June 27th and concluded on October 27th, 2022 – thanks to everyone who participated! Across the two rounds, we had over 80 unique submissions and gave out a total of 11 Third Prizes.

arXiv > cs > arXiv:2206.14486

6. Exponential scaling

Computer Science > Machine Learning

[Submitted on 29 Jun 2022 (v1), last revised 15 Nov 2022 (this version, v5)]

Beyond neural scaling laws: beating power law scaling via data pruning

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, Ari S. Morcos

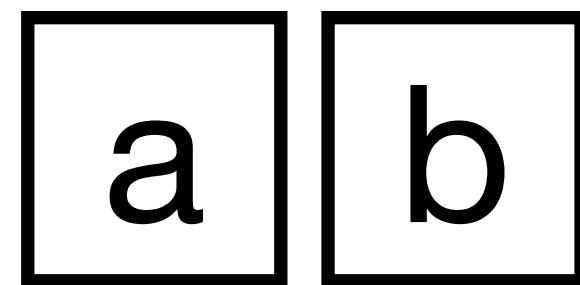
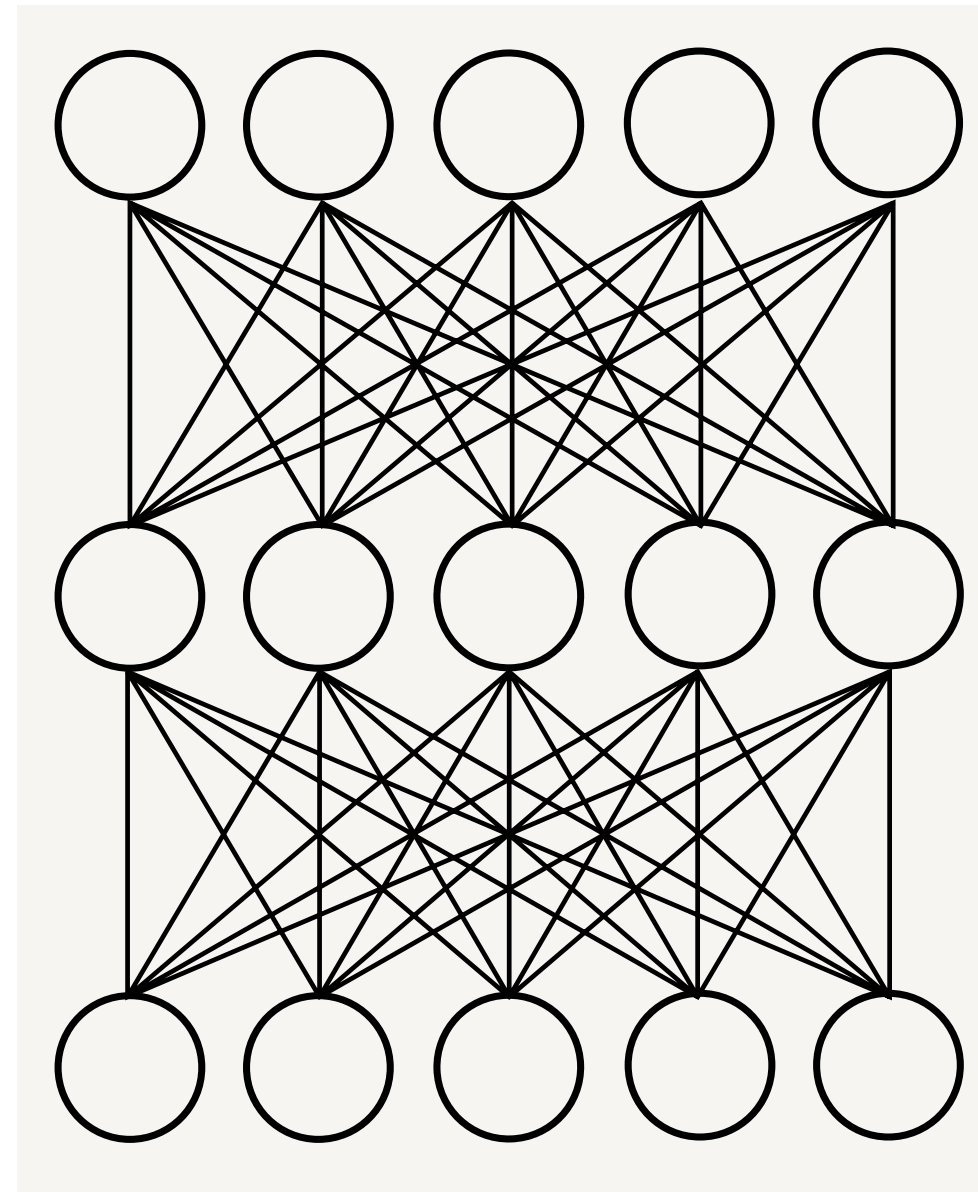
1. Grokking

Train a neural network to learn binary operations

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

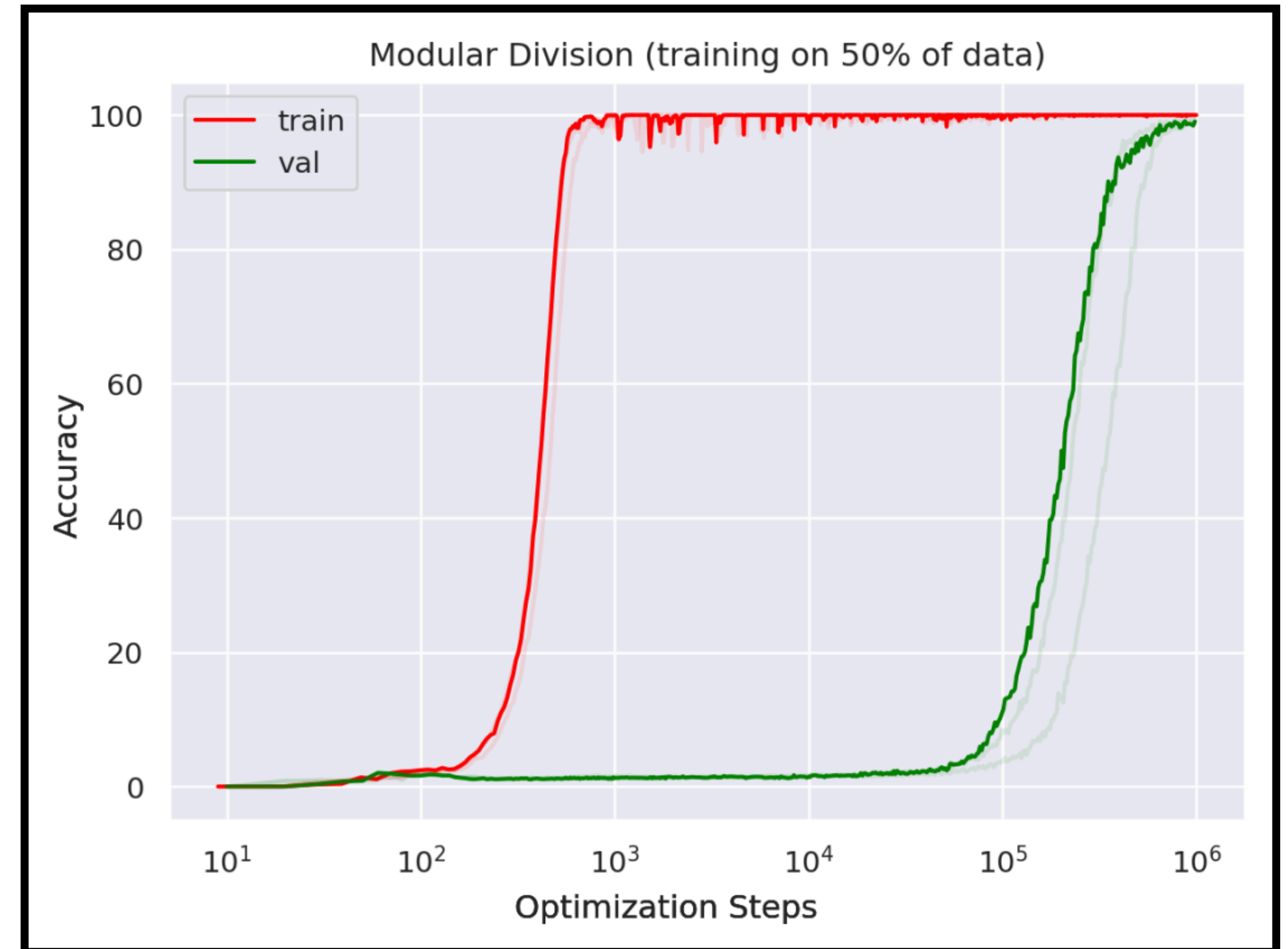
$$a \circ b = c$$

Logits for a, b, c, ...



Trainable
Embeddings

Phase transition behavior



arXiv > cs > arXiv:2201.02177

Computer Science > Machine Learning

[Submitted on 6 Jan 2022]

Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra

1. Grokking

arXiv > cs > arXiv:2205.10343

Computer Science > Machine Learning

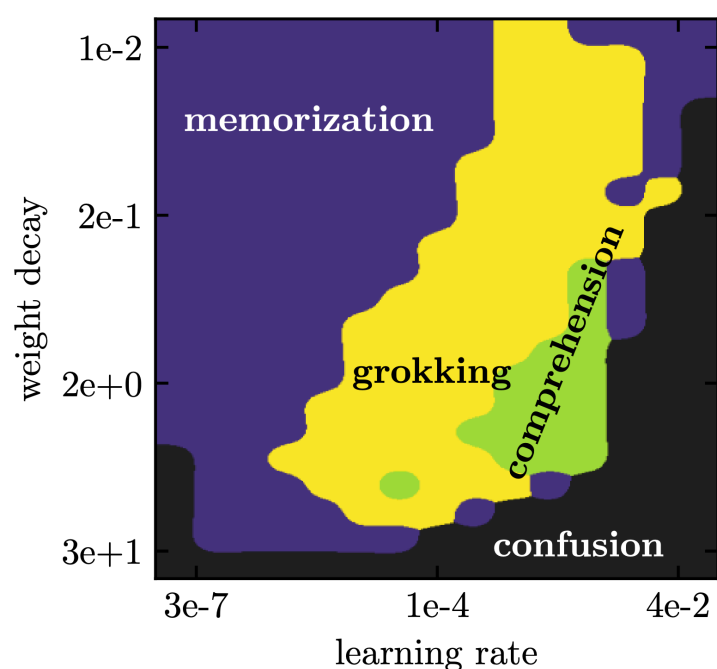
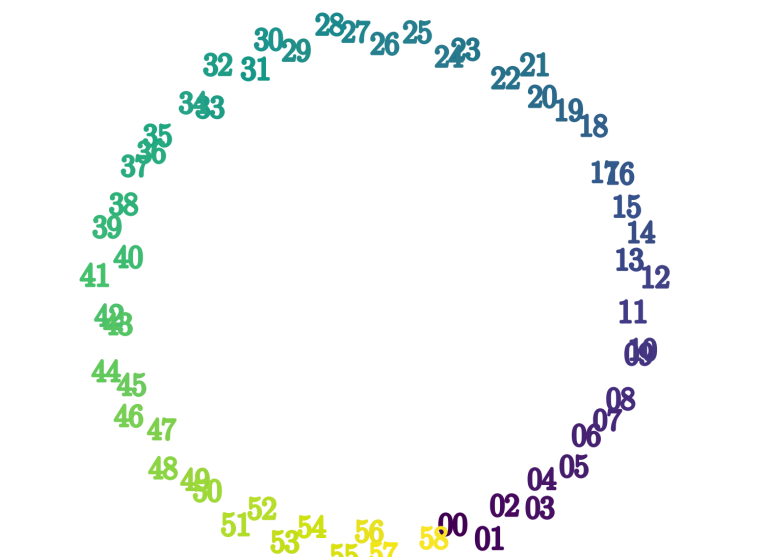
[Submitted on 20 May 2022 (v1), last revised 14 Oct 2022 (this version, v2)]

Towards Understanding Grokking: An Effective Theory of Representation Learning

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams

Representation learning and phase diagrams are key to understand grokking.

Representation Learning (20000 iterations)
train acc: 1.0 — val acc: 1.0



arXiv > cs > arXiv:2301.05217

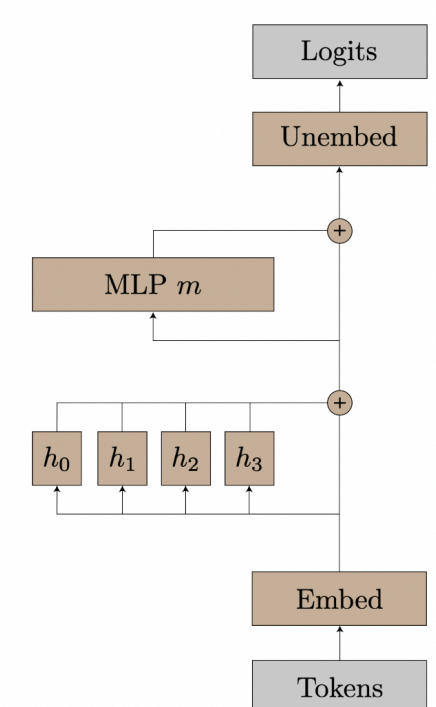
Computer Science > Machine Learning

[Submitted on 12 Jan 2023 (v1), last revised 13 Jan 2023 (this version, v2)]

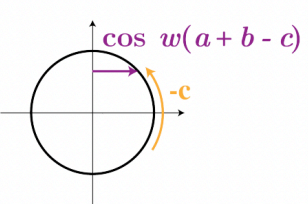
Progress measures for grokking via mechanistic interpretability

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, Jacob Steinhardt

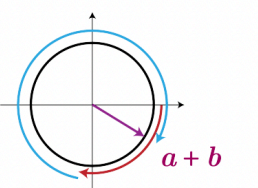
Grokking = Formation of the “clock math” circuit



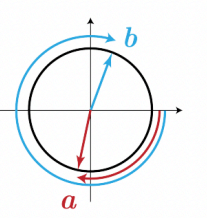
Computes logits using further trig identities:
 $\text{Logit}(c) \propto \cos(w(a + b - c))$
 $= \cos(w(a + b)) \cos(wc) + \sin(w(a + b)) \sin(wc)$



Calculates sine and cosine of $a + b$ using trig identities:
 $\sin(w(a + b)) = \sin(wa) \cos(wb) + \cos(wa) \sin(wb)$
 $\cos(w(a + b)) = \cos(wa) \cos(wb) - \sin(wa) \sin(wb)$



Translates one-hot a, b to Fourier basis:
 $a \rightarrow \sin(wa), \cos(wa)$
 $b \rightarrow \sin(wb), \cos(wb)$



arXiv > cs > arXiv:2210.01117

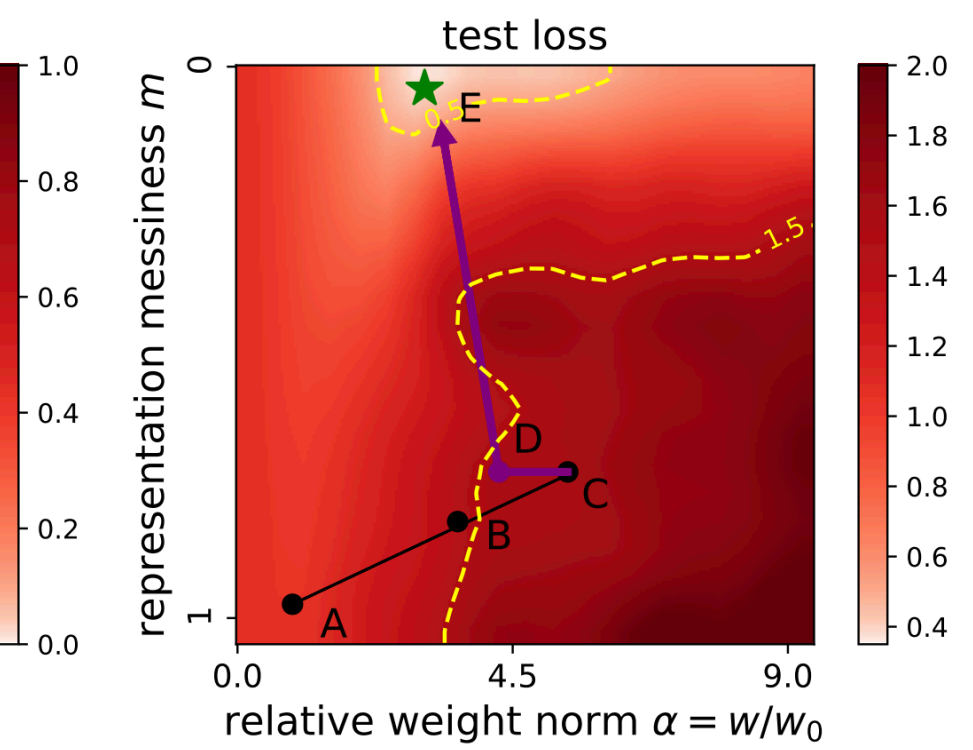
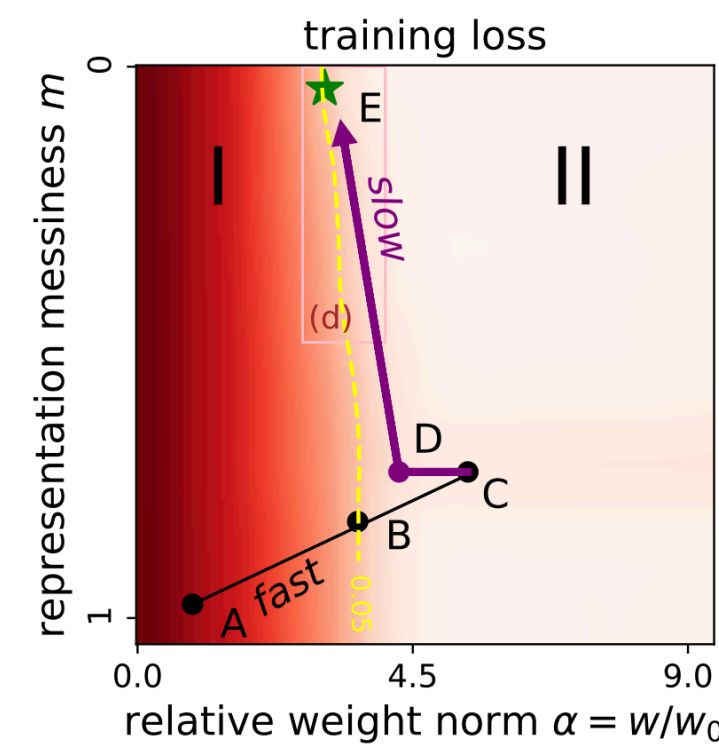
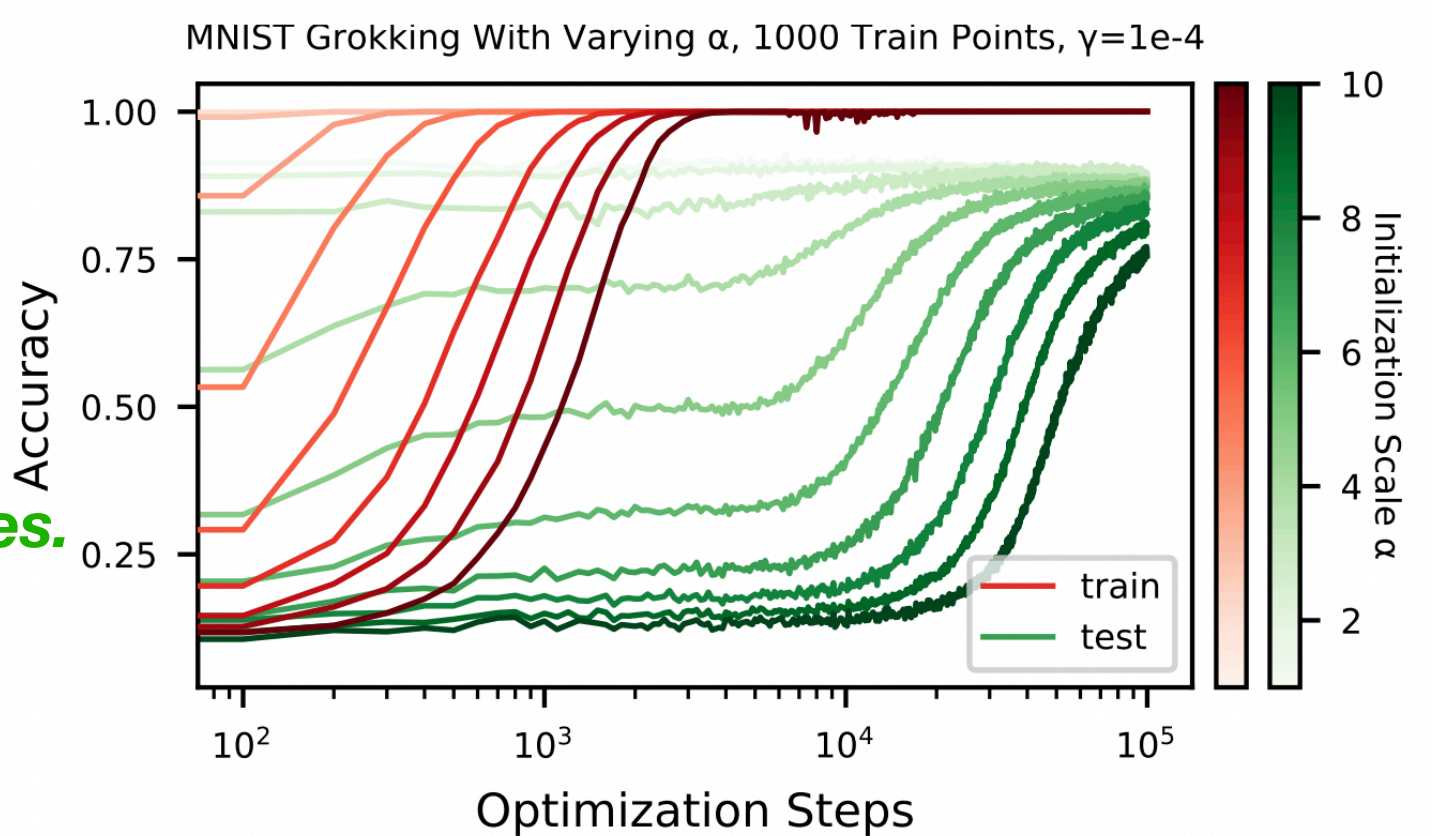
Computer Science > Machine Learning

[Submitted on 3 Oct 2022 (v1), last revised 23 Mar 2023 (this version, v2)]

Omnigrok: Grokking Beyond Algorithmic Data

Ziming Liu, Eric J. Michaud, Max Tegmark

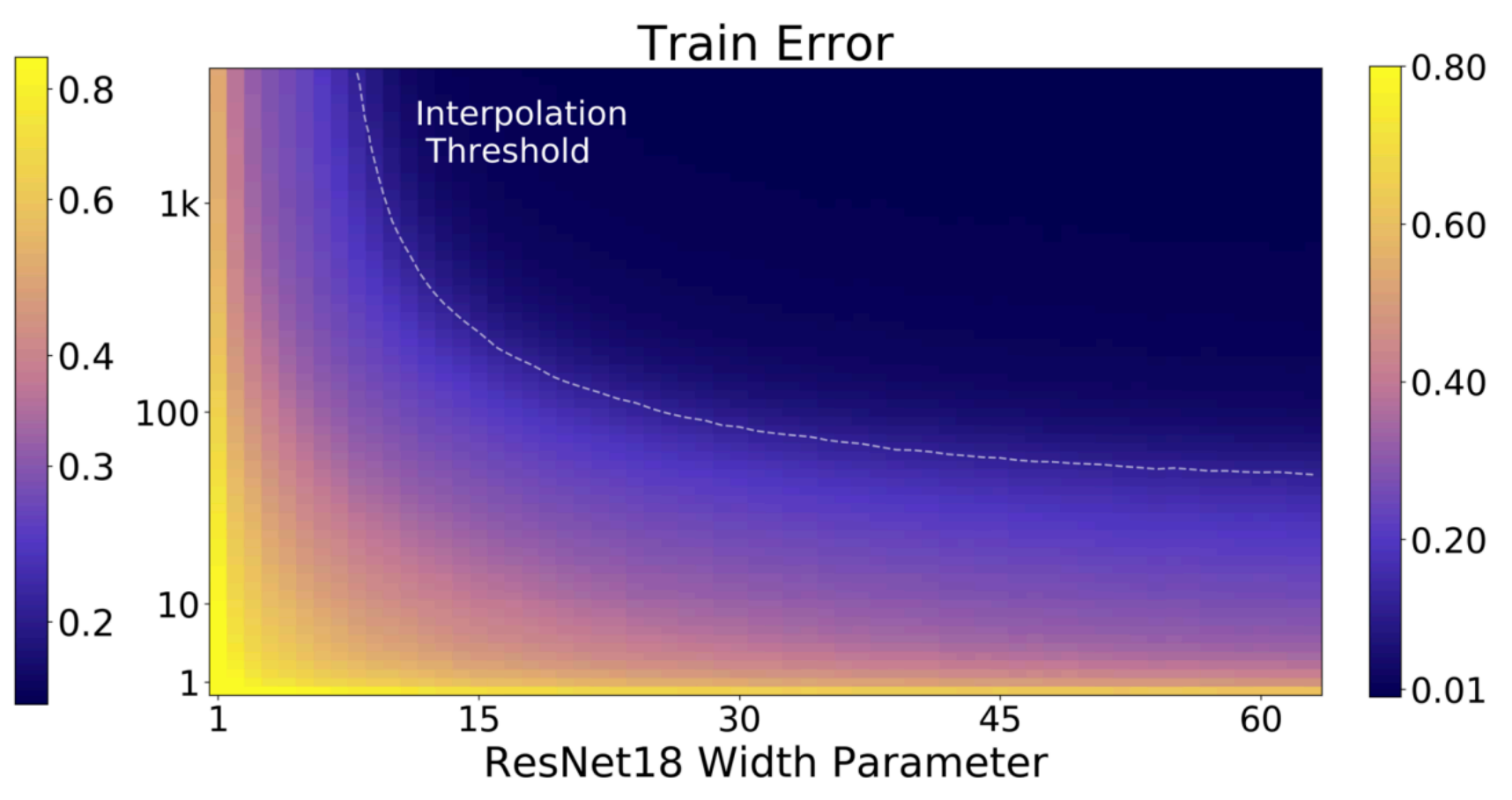
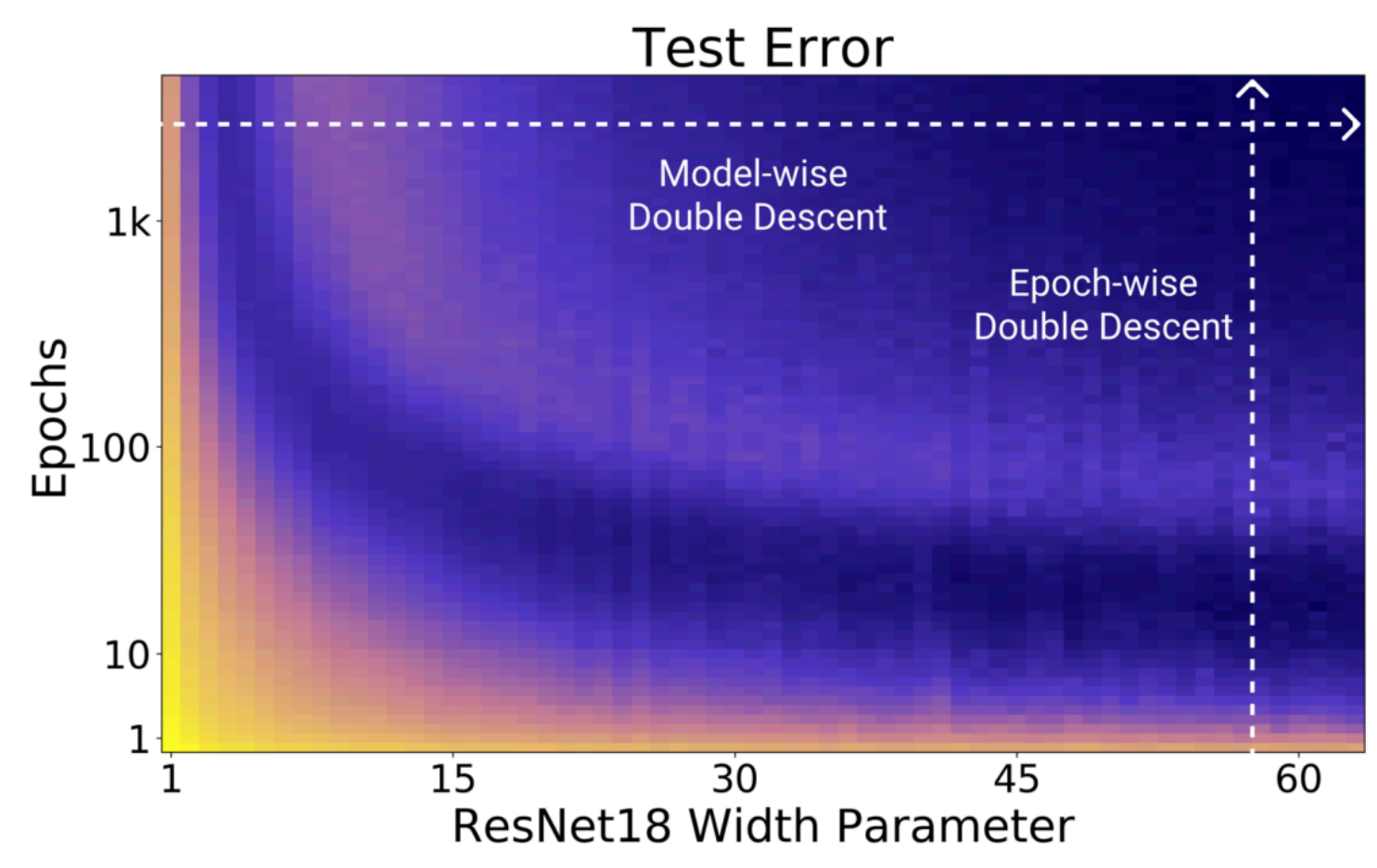
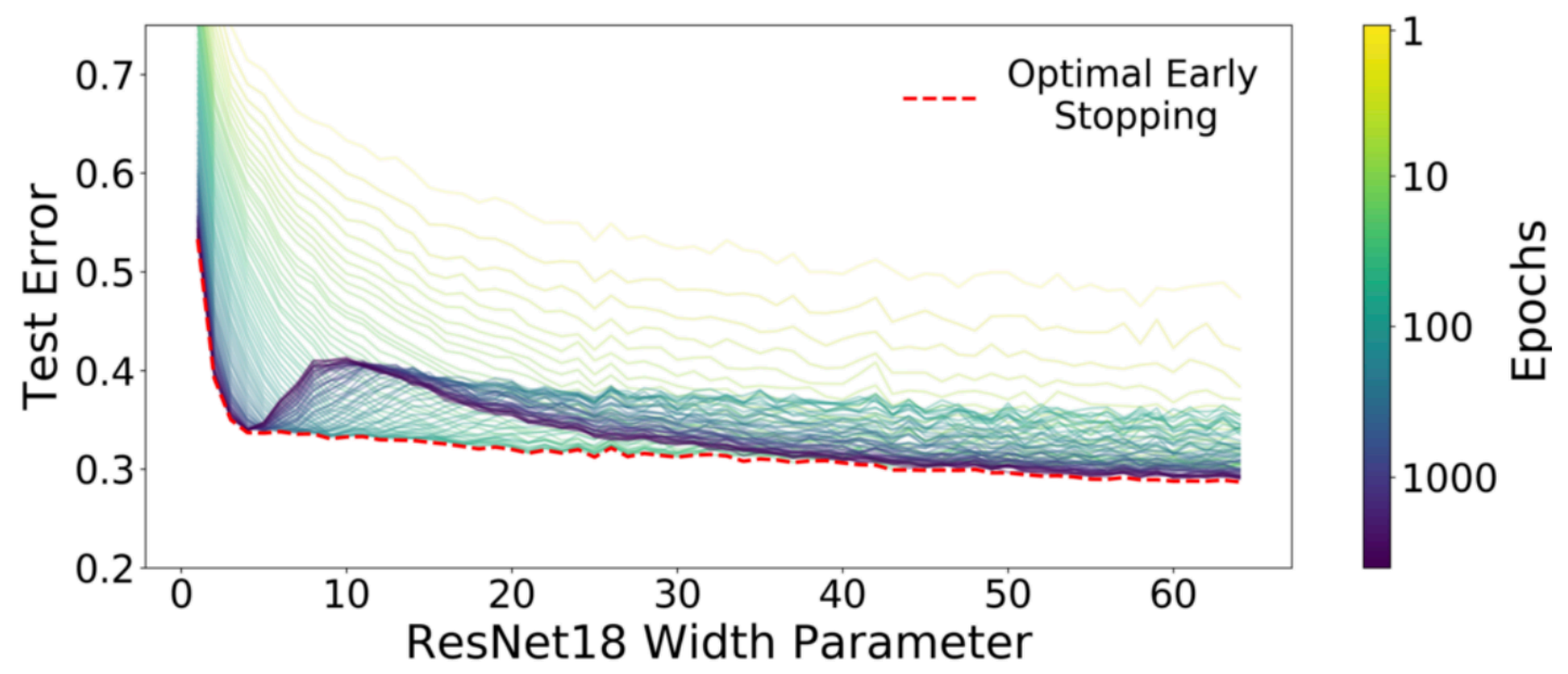
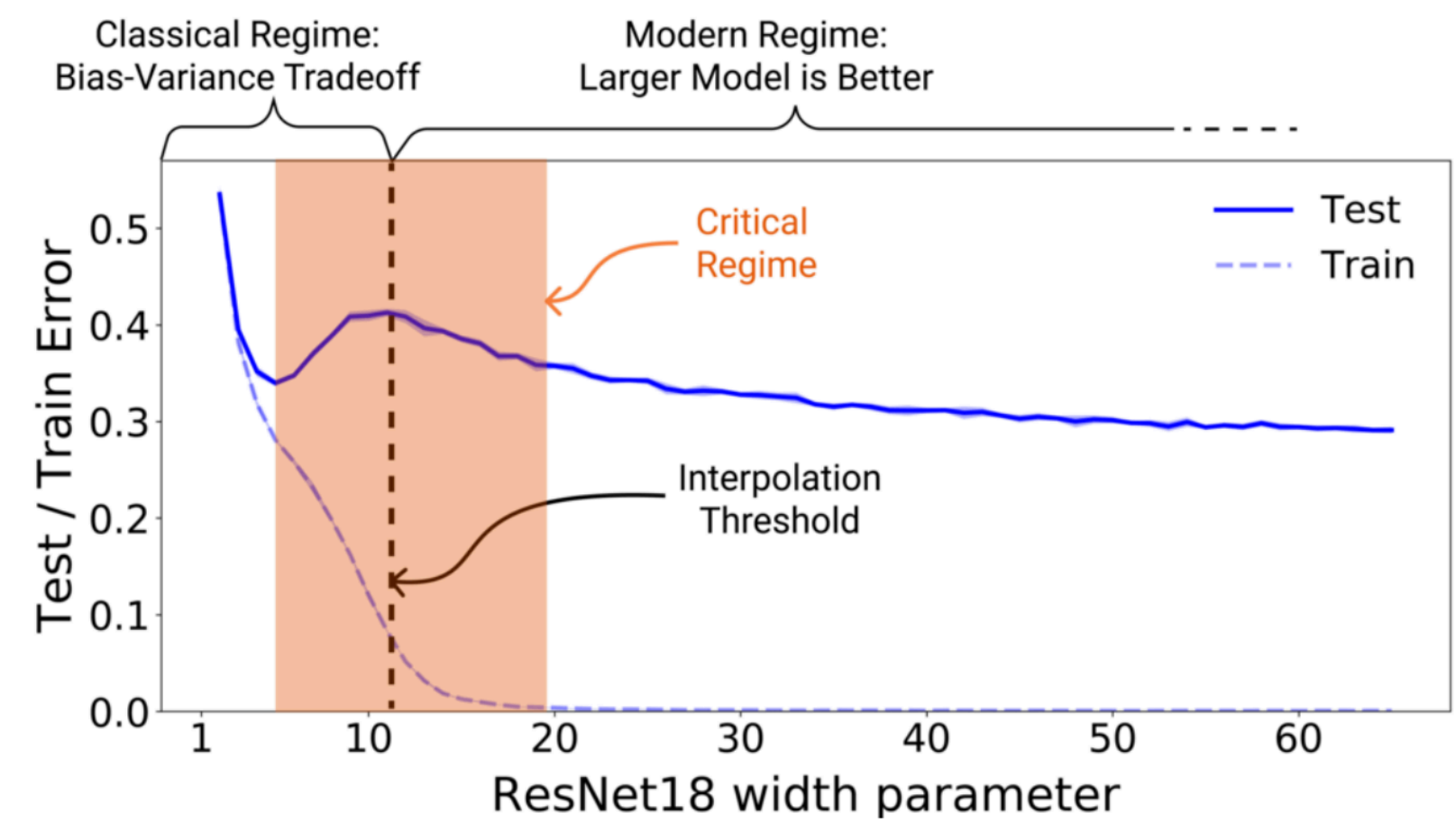
Grokking can be quite general.
 Grokking can be understood from neural loss landscapes.



2. Double descent

Deep Double Descent: Where Bigger Models and More Data Hurt

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, Ilya Sutskever

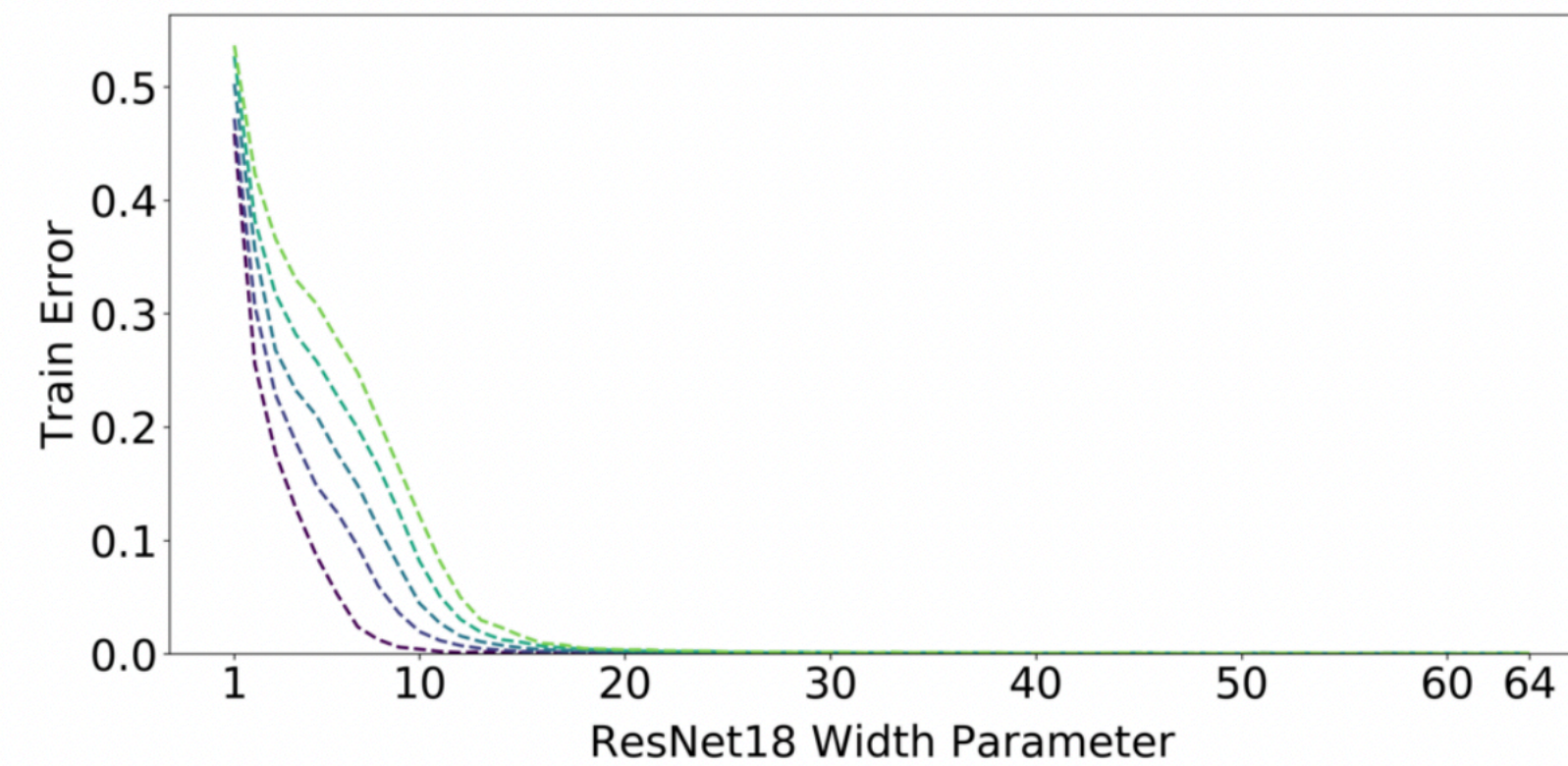
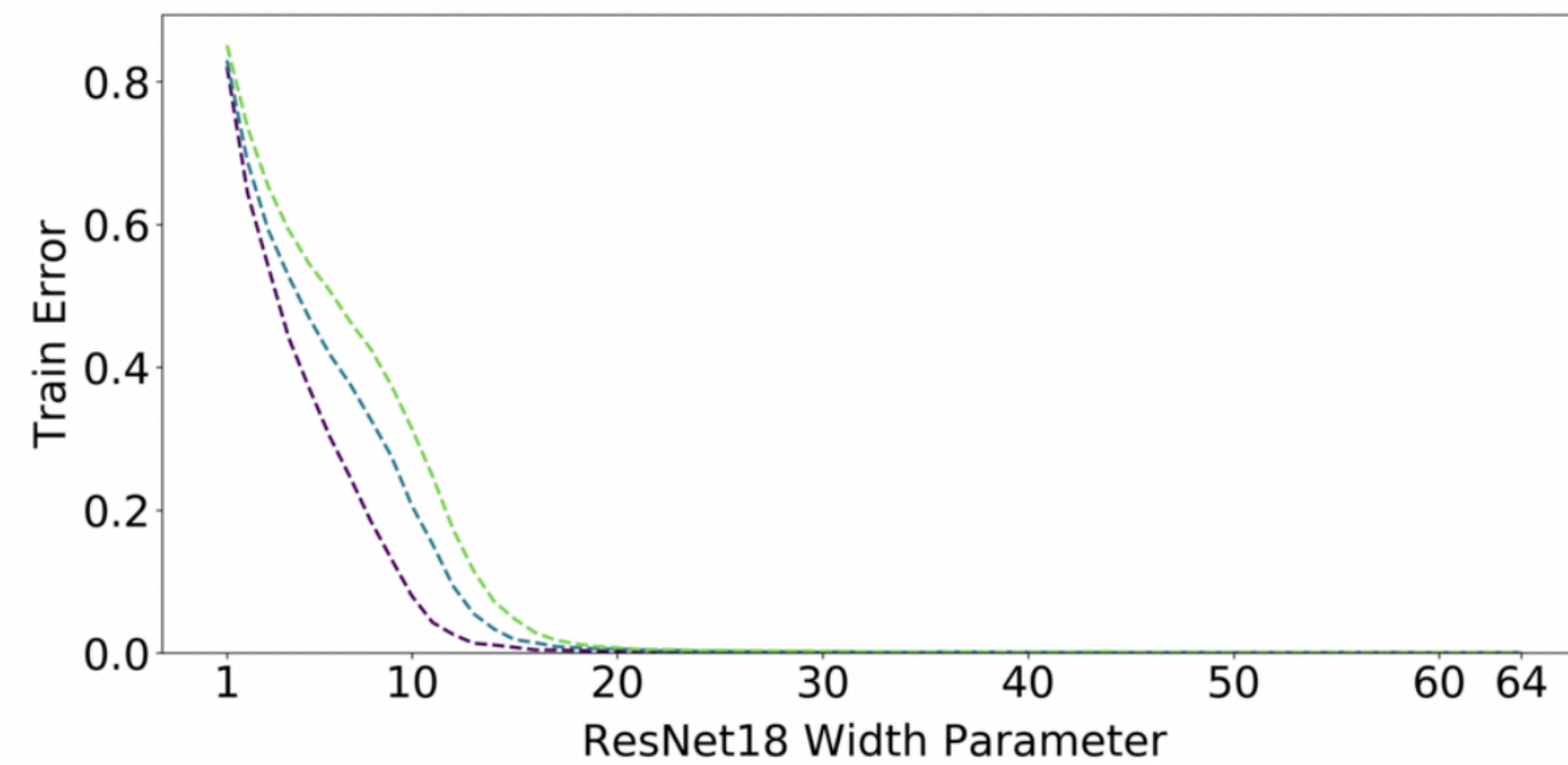
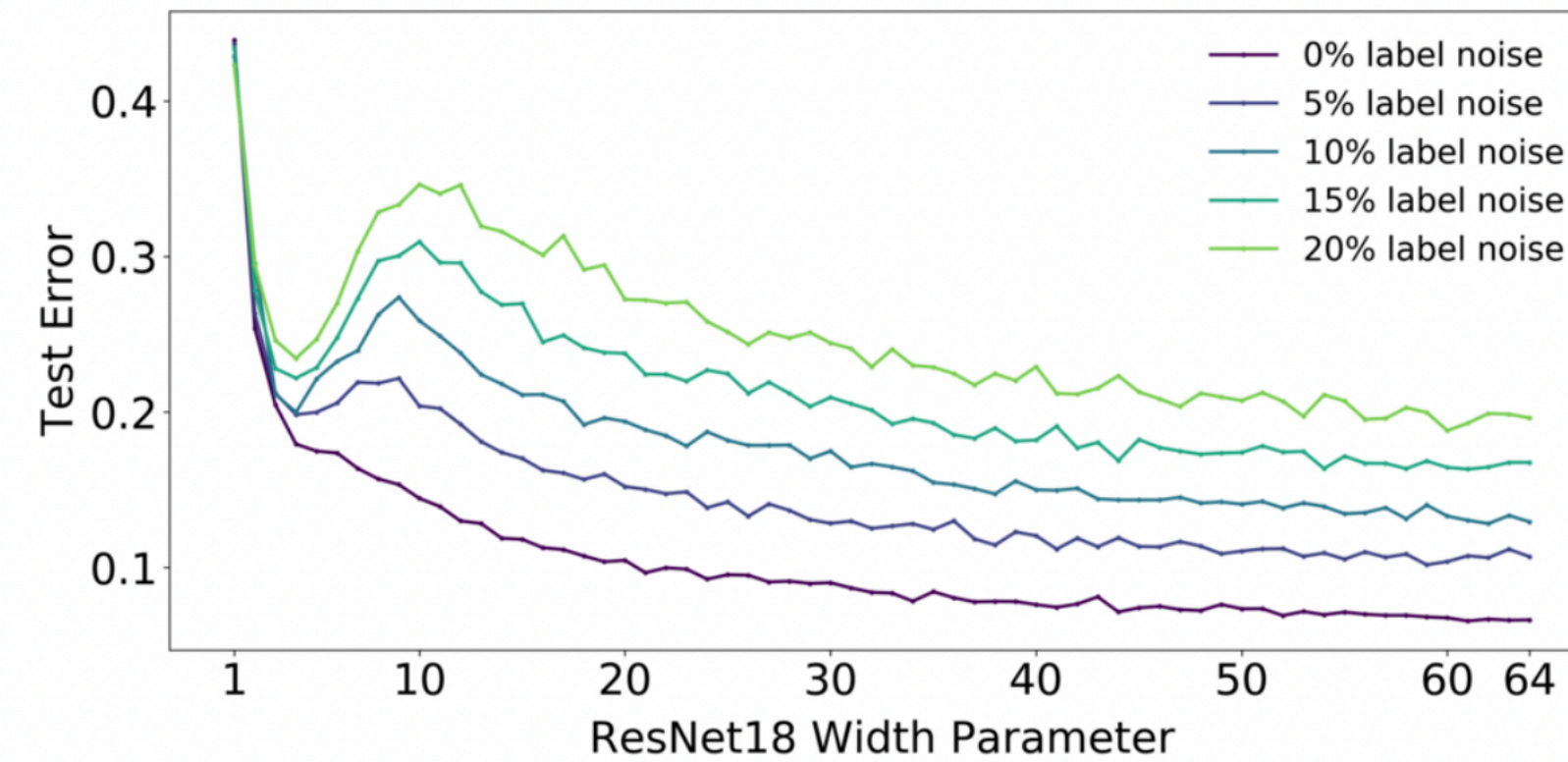
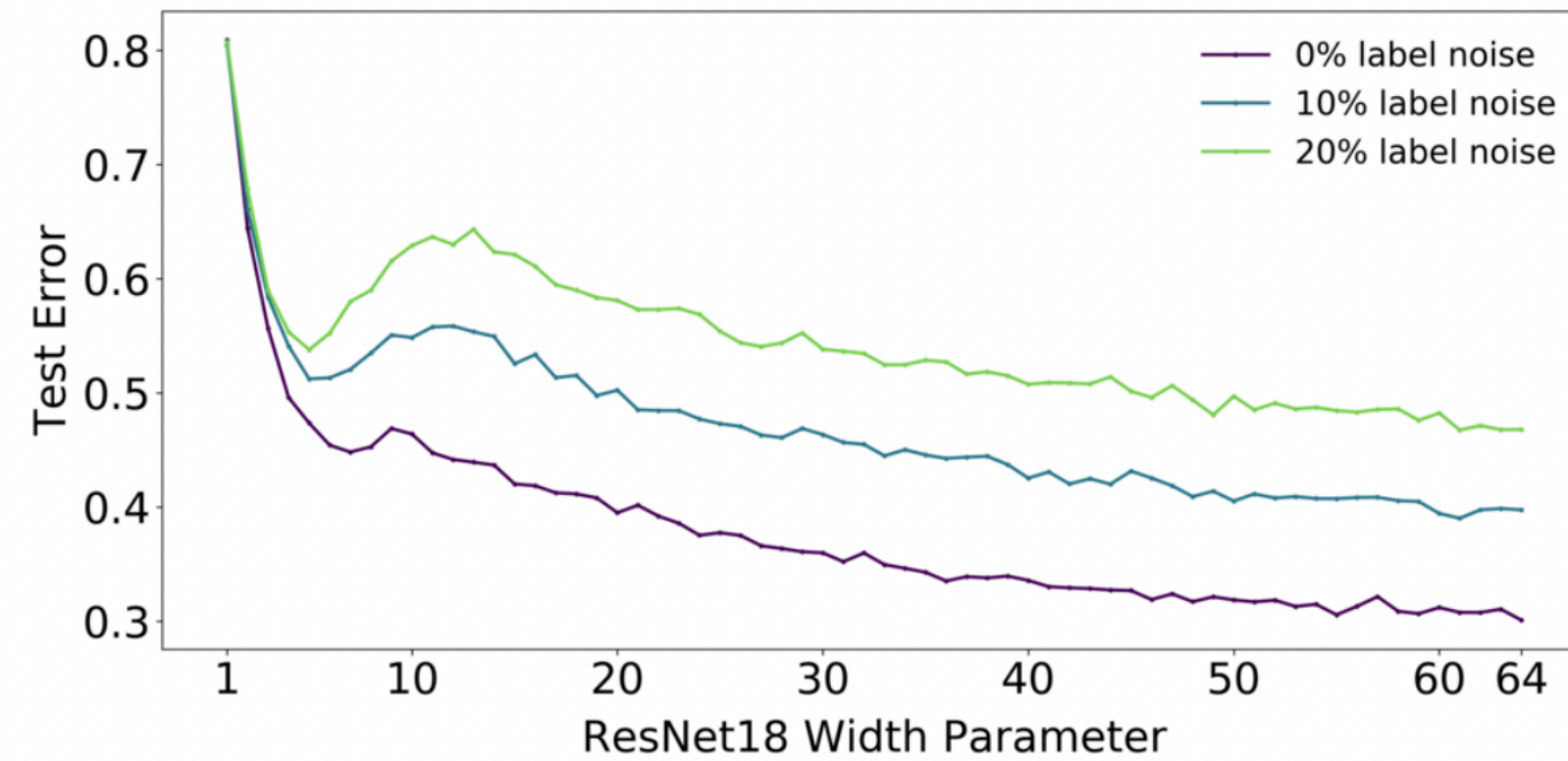


2. Double descent

Deep Double Descent: Where Bigger Models and More Data Hurt

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, Ilya Sutskever

You usually need noise to produce double descent.



(a) **CIFAR-100.** There is a peak in test error even with no label noise.

(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.

3. Broken Scaling Laws

Broken Neural Scaling Laws

Ethan Caballero, Kshitij Gupta, Irina Rish, David Krueger

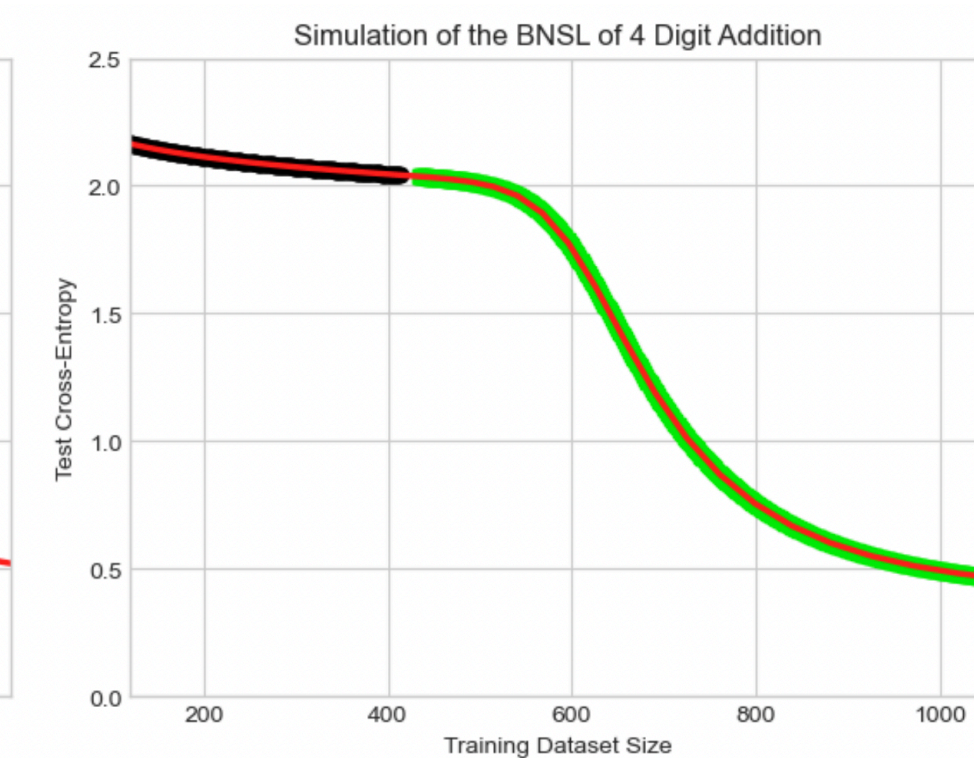
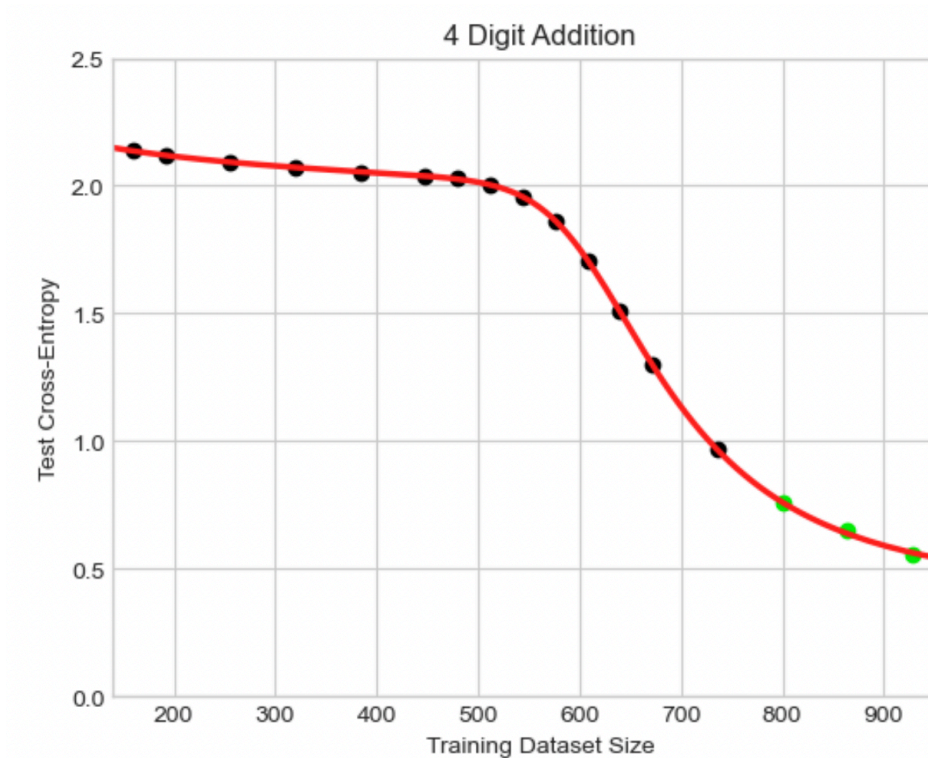
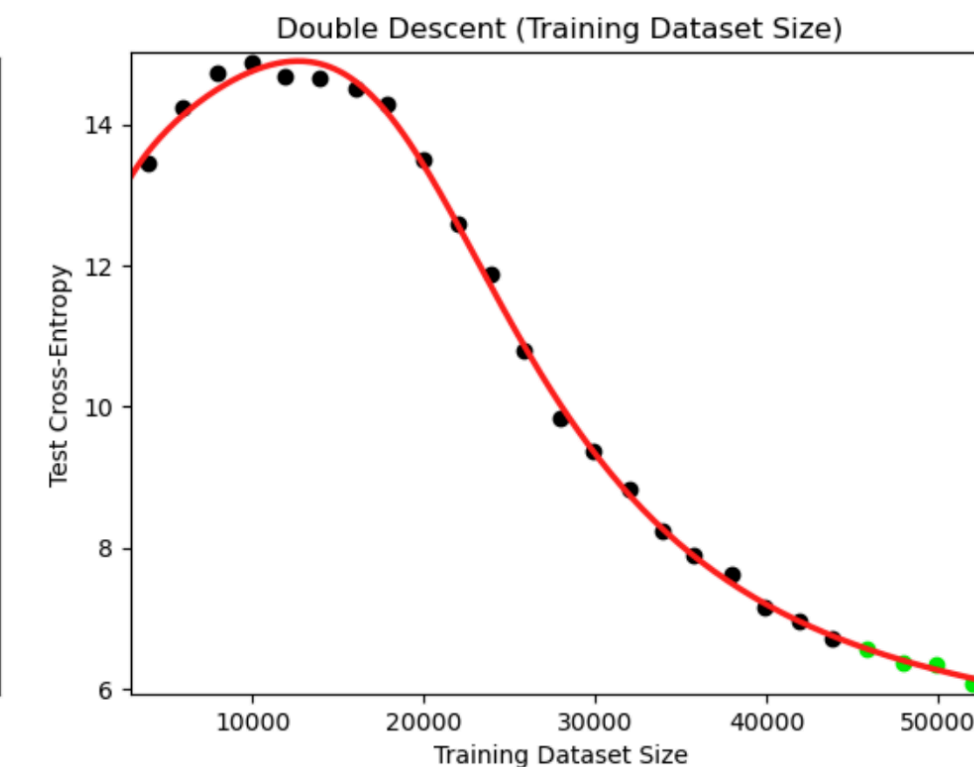
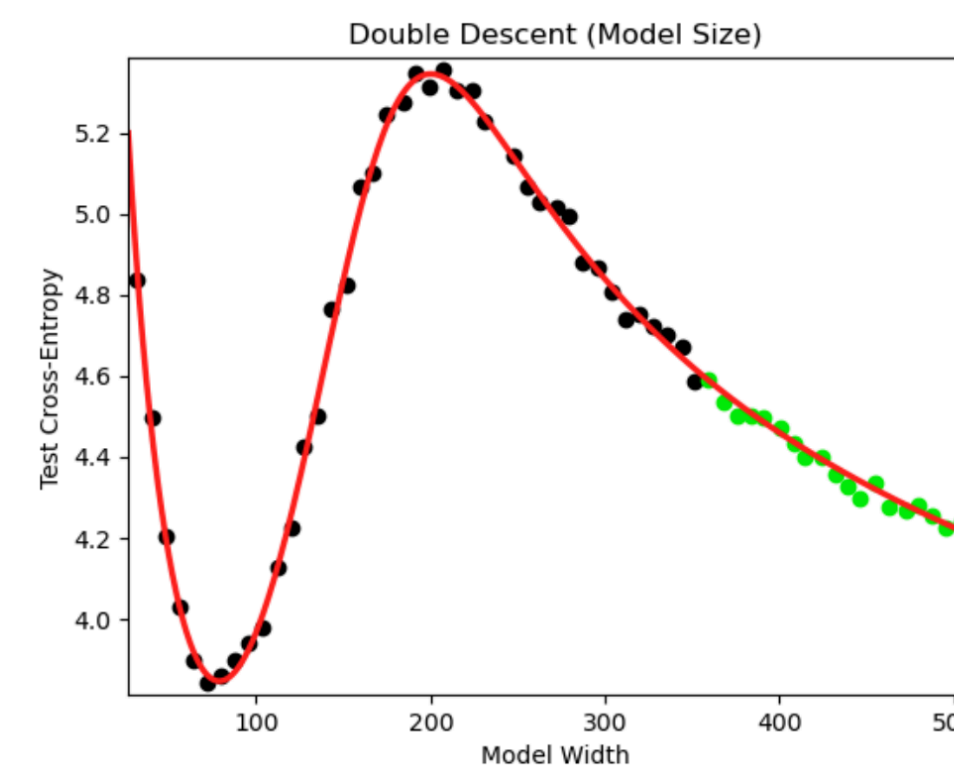
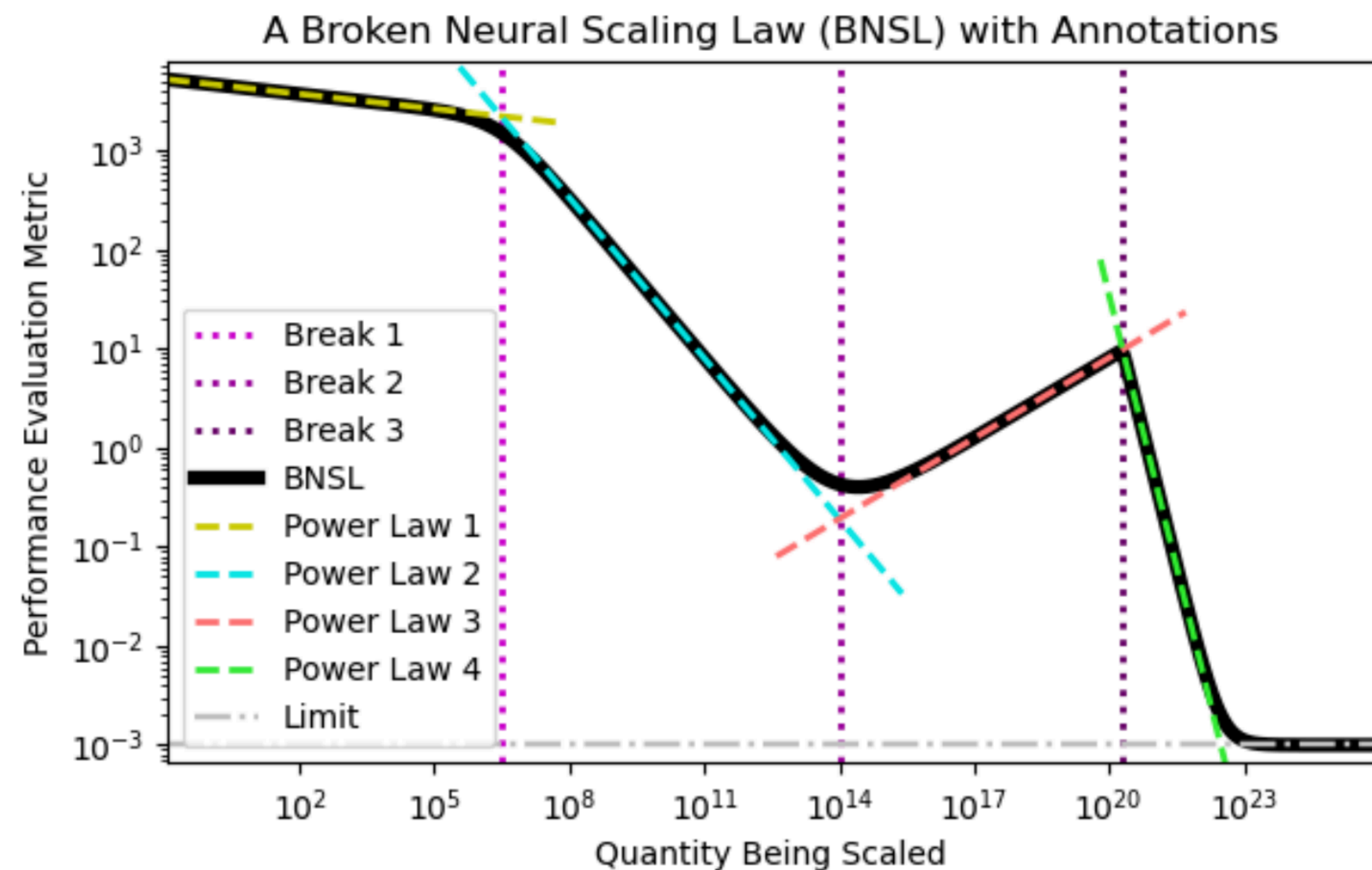
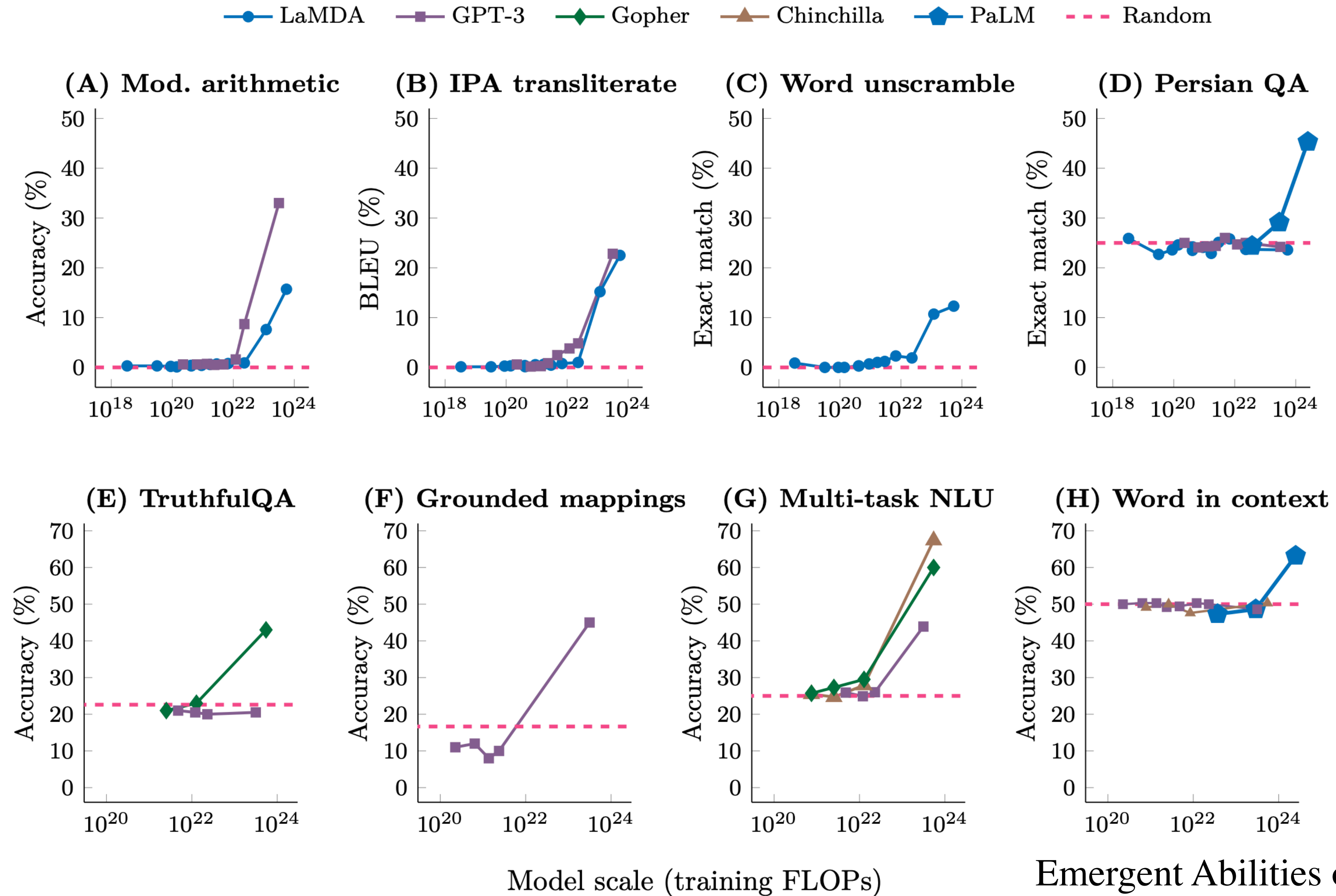


Figure 1: A Broken Neural Scaling Law (BNSL) (dark black solid line) (with 3 breaks where purple dotted lines intersect with dark black solid line) that contains 4 individual power law segments (where the dashed lines that are yellow, blue, red, and green overlap with the dark black solid line). The 1st and 2nd break are very smooth; the 3rd break is very sharp. See Section 2 for more details.

The general functional form of a broken neural scaling law (BNSL) is given as follows:

$$y = a + \left(bx^{-c_0} \right) \prod_{i=1}^n \left(1 + \left(\frac{x}{d_i} \right)^{1/f_i} \right)^{-c_i * f_i}, \quad (1)$$

4. Emergent abilities (EA)



arXiv: 2206.07682

Emergent Abilities of Large Language Models

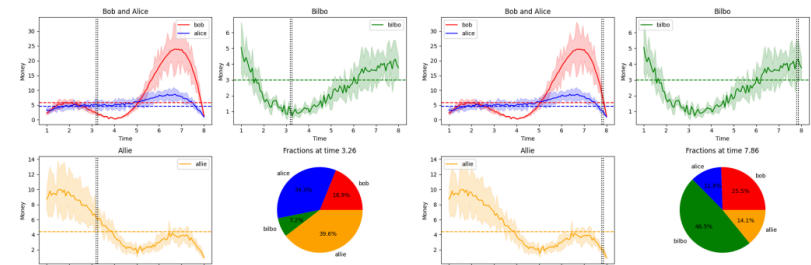
GPT-4 emergent abilities

Write python codes

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y_1 = [2, 3, 5, 18, 1]$, $y_2 = [3, 5, 6, 8, 1]$, $y_3 = [5, 1, 2, 3, 4]$, $y_4 = [9, 7, 2, 3, 1]$. I want y_1, y_2 to be in the same plot, but y_3 is in another plot next to that plot, y_4 is in below. I want the legend of y_1 to be "bob", y_2 to be "alice", y_3 to be "bilbo", y_4 to be "allie". I want the x-axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in "continuous time". Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: [We give two snapshots from the animation resulting from the produced code]



Compose music

GPT-4

Can you compose a short tune (say four to eight bars) using ABC notation?
There are many possible ways to compose a short tune in abc notation, but here is one example:

```
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2A2 c2B4 | G2A2 c2c2 | e4cB c2B4 | G4-G2c2 ||
|:c2B4 f2e4 | c2B4 c2c2 | g2fe d2B4 | c4-c2c2 ||
```

This tune has two four-bar phrases, each repeated once, forming an A part and a B part. The key is C major.

Can you describe the tune in musical terms?
Here is one possible description:
The tune starts with a rising arpeggio of the tonic chord (C major), followed by a descending scale that returns to the tonic. This pattern is repeated, but the second time the scale continues to the higher tonic note. The first phrase ends with a long tonic note.
The second phrase begins with a similar pattern, but using the subdominant chord (F major) instead of the tonic. The second time, the scale continues to the higher subdominant note. The phrase ends with a stepwise descent to the tonic, followed by a long tonic note.
The A part and the B part have a similar structure, but use different chords and ranges. The tune has a simple and catchy melody, with a clear sense of tonality and rhythm.

I like part A, but maybe part B is just a bit too similar to part A, making the song repetitive. How about beginning part B with a descending arpeggio instead of a rising one?
Sure, that would create some contrast and variation. Here is one possible way to do that:

```
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2A2 c2B4 | G2A2 c2c2 | e4cB c2B4 | G4-G2c2 ||
|:f2e4 c2B4 | f2e4 c2c2 | g2fe d2B4 | c4-c2c2 ||
```

That's much better. Now how about making this into a duet, can you add a second staff for this with a bass accompanying the melody?
One possible way to write a second staff for this tune is to use root notes and fifths of the chords, as well as some passing notes, to create a simple accompaniment. Here is an example in abc notation, with the second hand bass on the lower staff:

```
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2A2 c2B4 | G2A2 c2c2 | e4cB c2B4 | G4-G2c2 ||
|:f2e4 c2B4 | f2e4 c2c2 | g2fe d2B4 | c4-c2c2 ||
V:2 clef=bass
|:G,,2C,,2 G,,2C,,2 | C,,2E,,2 G,,2C,,2 | F,,2F,,2 E,,2E,,2 | C,,4-G,,2C,,2 ||
|:F,,2F,,2 G,,2C,,2 | F,,2F,,2 G,,2C,,2 | G,,2C,,2 F,,2F,,2 | E,,2C,,2 G,,2C,,2 ||
```

The sheet music of the final tune:

The image shows the sheet music for the composed tune. It consists of two staves. The top staff is the melody in C major, and the bottom staff is the bass accompaniment. The melody is written in 4/4 time and consists of two phrases, each repeated once. The bass accompaniment is written in 4/4 time and consists of two phrases, each repeated once.

arXiv > cs > arXiv:2303.12712

Search.
Help | A

Computer Science > Computation and Language

[Submitted on 22 Mar 2023 (v1), last revised 27 Mar 2023 (this version, v3)]

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

Write poems

Prompt: Can you write a proof that there are many primes, with every line that rhymes?

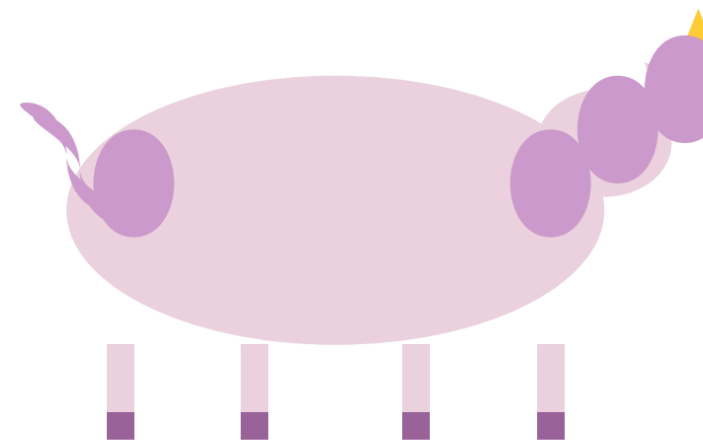
GPT-4:

Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

Write LaTeX codes

Prompt: Draw a unicorn in TikZ.

GPT-4: [Produces \LaTeX compiling to following picture.]



Do math problems

Prompt: Suppose $g(x) = f^{-1}(x)$, $g(0) = 5$, $g(4) = 7$, $g(3) = 2$, $g(7) = 9$, $g(9) = 6$ what is $f(f(f(6)))$?

GPT-4: There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g . That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g , and vice versa. Using this idea, we can try to find $f(6)$ by looking for the value of x such that $g(x) = 6$. From the given information, we see that $g(9) = 6$, so $f(6) = f(g(9)) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = f(g(7)) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(4) = 7$, so $f(7) = f(g(4)) = 4$. Therefore, $f(f(f(6))) = f(f(9)) = f(7) = 4$.

Explain jokes

Example of GPT-4 visual input:

User: What is funny about this image? Describe it panel by panel.



Source: <https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/>

GPT-4: The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Table 3. Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question about an image with multiple panels which GPT-4 is able to answer.

6. Exponential scaling

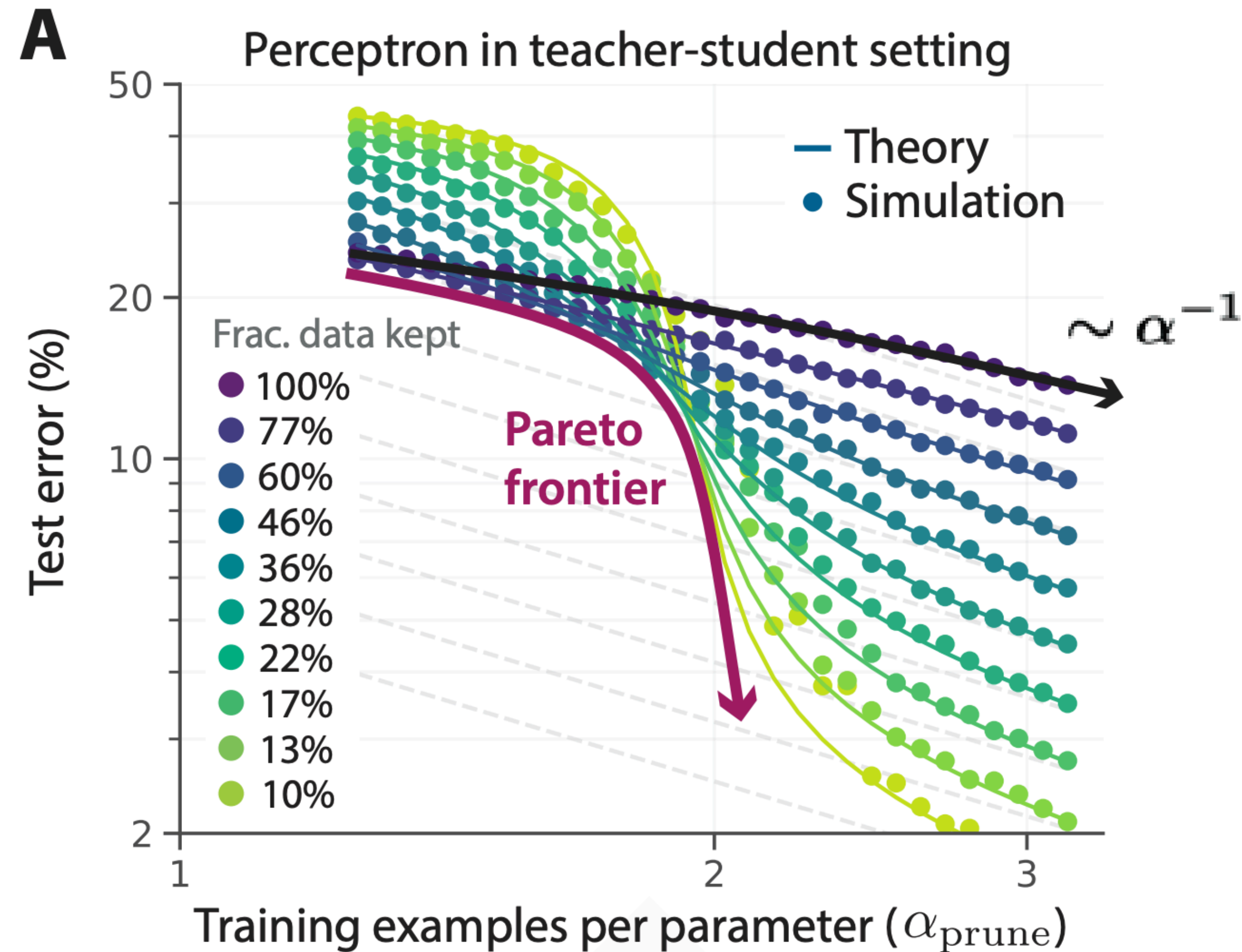
arXiv > cs > arXiv:2206.14486

Computer Science > Machine Learning

[Submitted on 29 Jun 2022 (v1), last revised 15 Nov 2022 (this version, v5)]

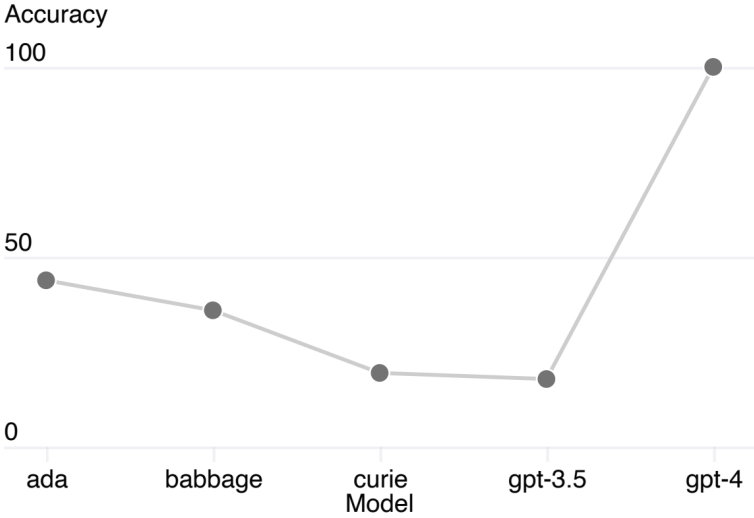
Beyond neural scaling laws: beating power law scaling via data pruning

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, Ari S. Morcos



A unified theory?

Inverse scaling prize, hindsight neglect



Inverse Scaling

arXiv > cs > arXiv:2303.08774
Computer Science > Computation and Language
[Submitted on 15 Mar 2023 (v1), last revised 27 Mar 2023 (this version, v3)]
GPT-4 Technical Report
OpenAI

arXiv > cs > arXiv:2303.13506
Computer Science > Machine Learning
[Submitted on 23 Mar 2023]
The Quantization Model of Neural Scaling
Eric J. Michaud, Ziming Liu, Uzay Girit, Max Tegmark

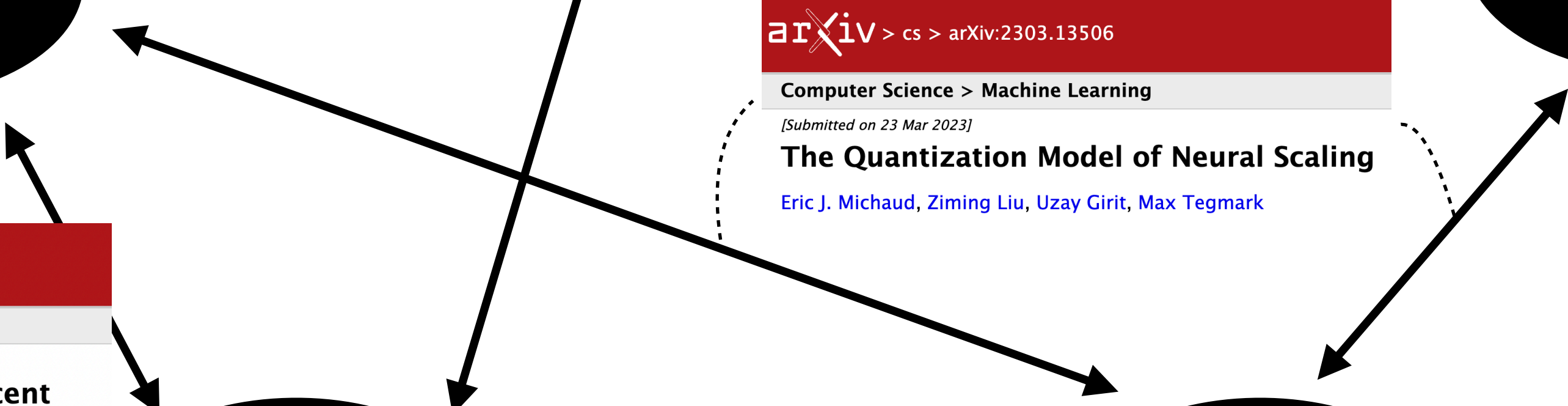
arXiv > cs > arXiv:2303.06173
Computer Science > Machine Learning
[Submitted on 10 Mar 2023]
Unifying Grokking and Double Descent
Xander Davies, Lauro Langosco, David Krueger

Grokking

Double Descent

Scaling Laws

Emergent Abilities



Two theories of Neural Scaling Laws

Continuous & Geometrical => “Classical physics”

arXiv > cs > arXiv:2004.10802

Computer Science > Machine Learning

[Submitted on 22 Apr 2020]

A Neural Scaling Law from the Dimension of the Data Manifold

Utkarsh Sharma, Jared Kaplan

Discrete & Statistical => “Quantum physics”

arXiv > cs > arXiv:2303.13506

Computer Science > Machine Learning

[Submitted on 23 Mar 2023]

The Quantization Model of Neural Scaling

Eric J. Michaud, Ziming Liu, Uzay Girit, Max Tegmark

Continuous & Geometrical => “Classical physics”

arXiv > cs > arXiv:2004.10802

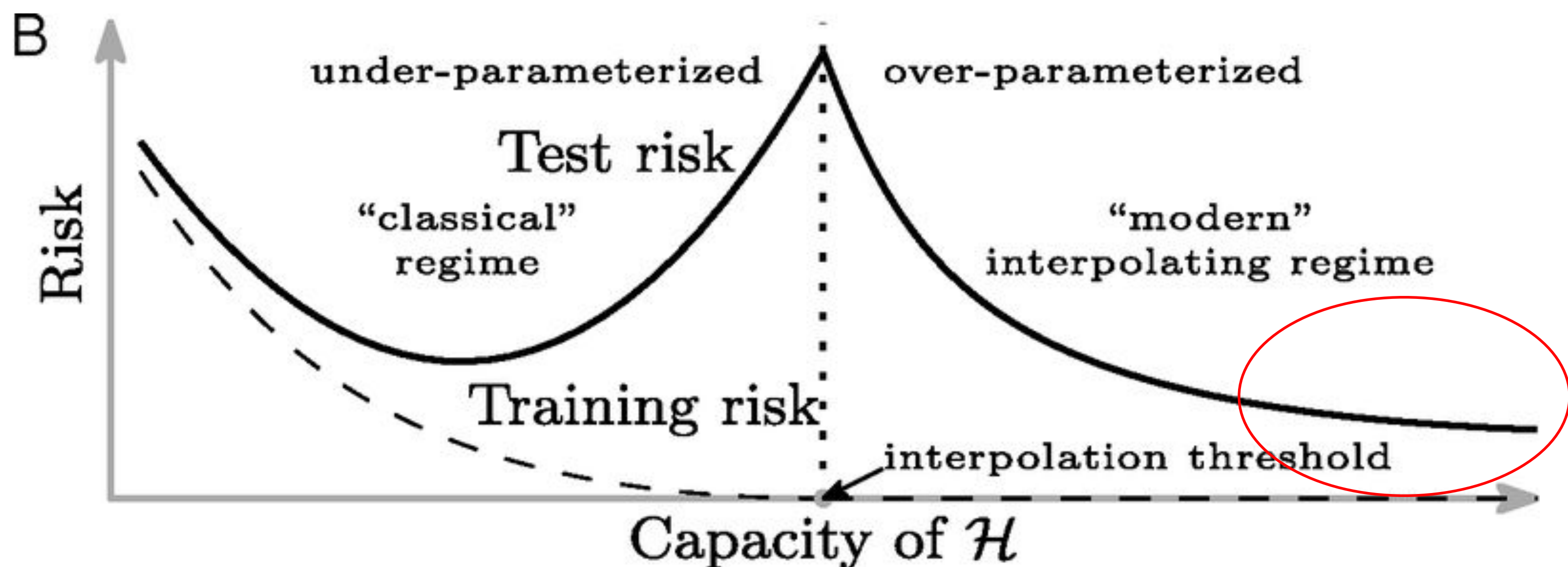
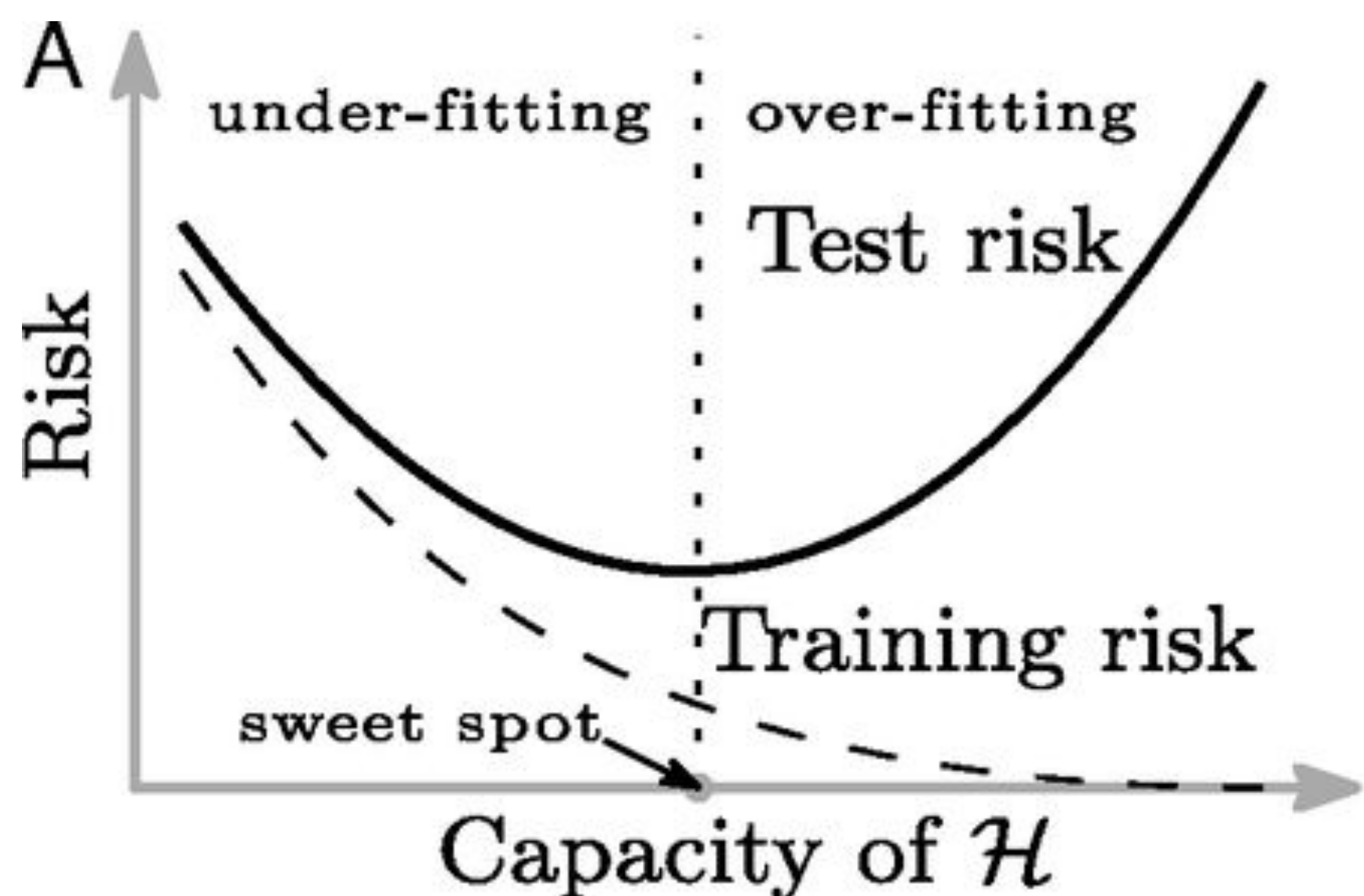
Computer Science > Machine Learning

[Submitted on 22 Apr 2020]

A Neural Scaling Law from the Dimension of the Data Manifold

Utkarsh Sharma, Jared Kaplan

Bias-Variance Tradeoff



What's the asymptotic behavior here?

arXiv:2004.10802

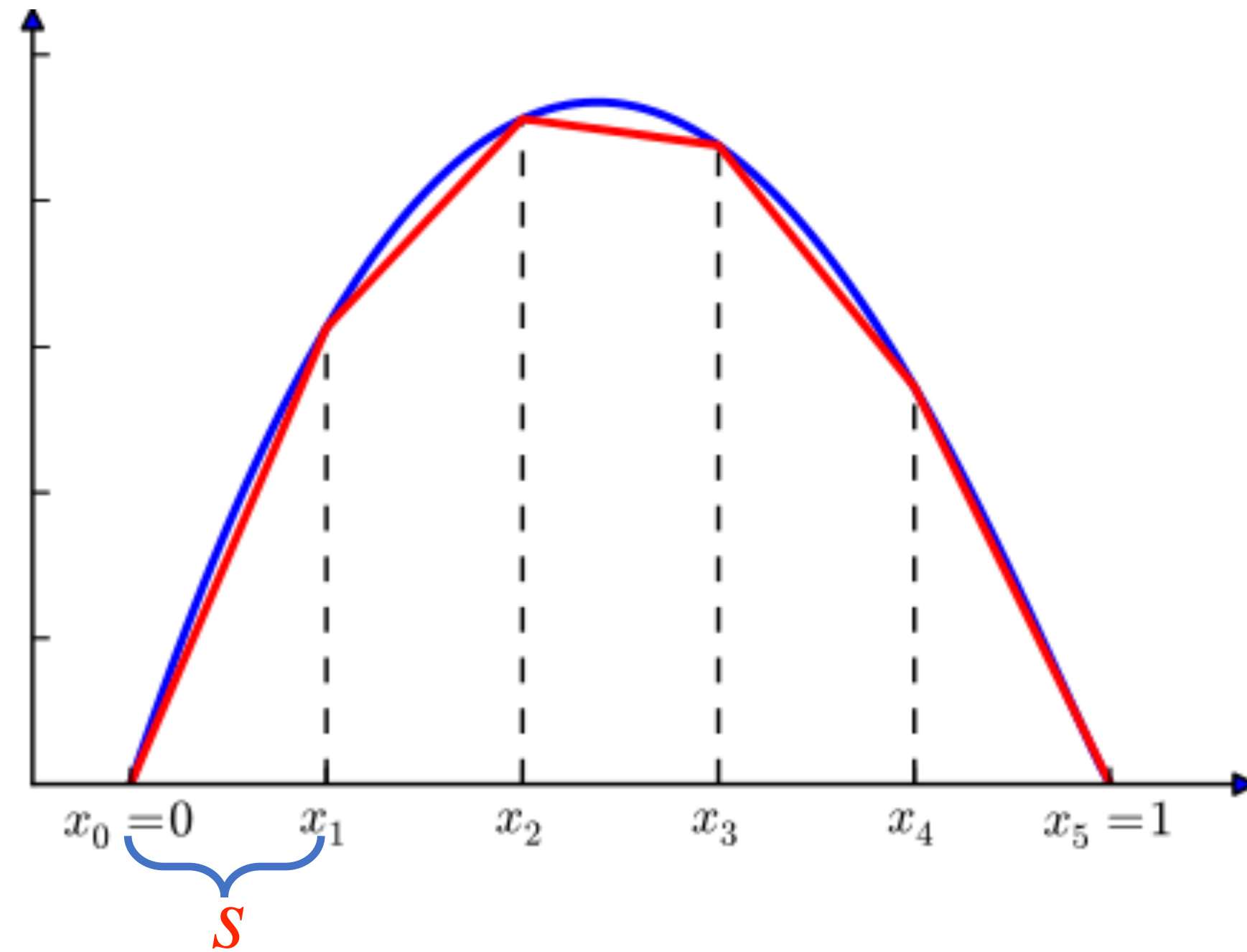
$$L \sim N^{-\alpha}, \alpha \approx \frac{4}{d} \text{ (MSE loss)}$$

Picture from "Reconciling modern machine-learning practice and the classical bias-variance trade-off"

<https://www.pnas.org/content/pnas/116/32/15849.full.pdf>

A simple theory of approximation

ReLU network = piecewise linear function wrt input



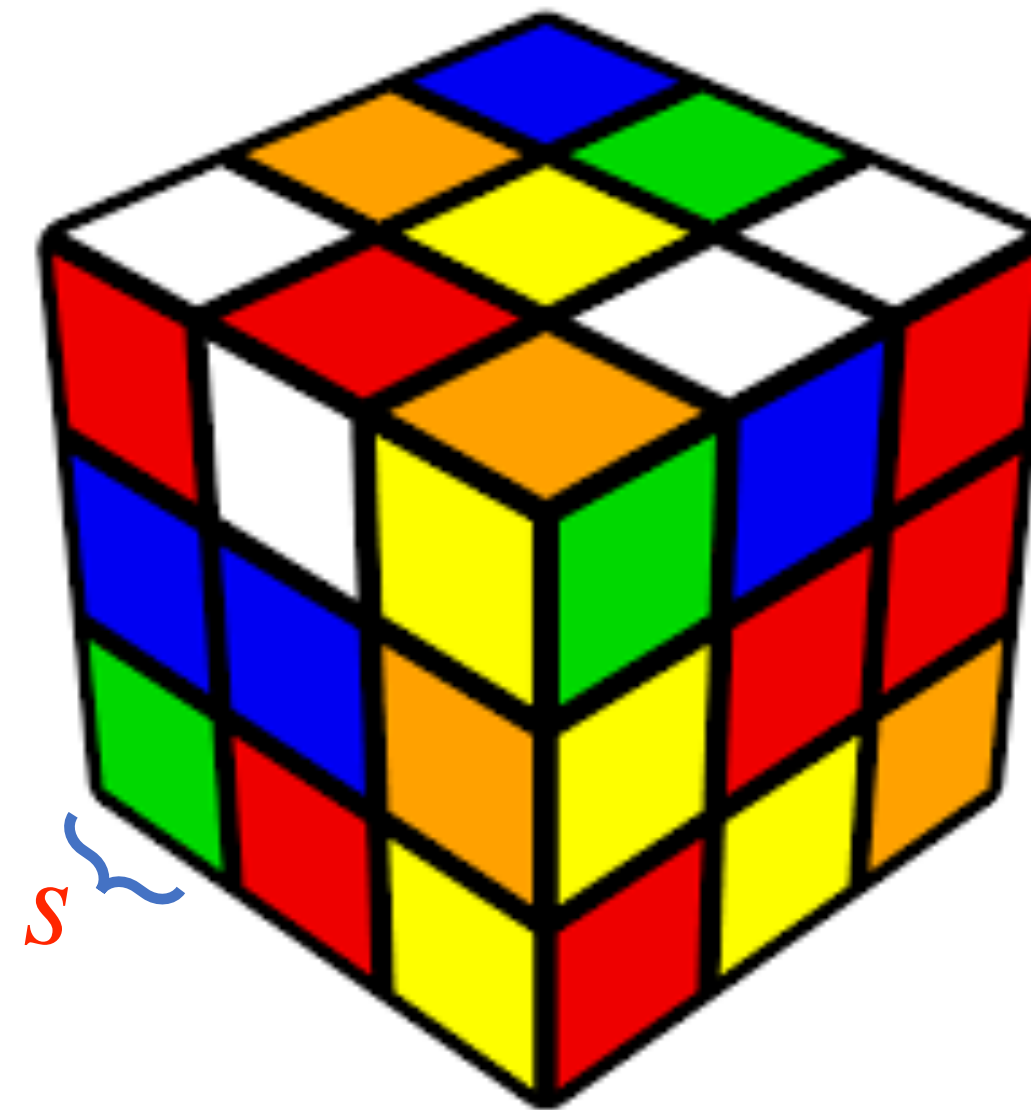
$N = \frac{1}{s}$ pieces ($s = \frac{1}{N}$), each piece need ~ 2 parameters

Totally $2N = O(N)$ parameters to describe the curve

$$L = \int_0^1 dx \left| f(x) - c(x) \right|^2 \leq \lambda^2 s^4 \sim N^{-4}$$

f is Lipschitz continuous: $|f(x+y) - f(x)| \leq \lambda|y|$

In general, p -loss and d -dimension:



$N = \frac{1}{s^d}$ pieces ($s = N^{-\frac{1}{d}}$), each piece need $\sim d + 1$ parameters

Totally $(d + 1)N = O(N)$ parameters to describe the curve

$$L = \int_0^1 dx^d \left| f(x) - c(x) \right|^p \leq \lambda^p s^{2p} \sim N^{-\frac{2p}{d}}$$

Implication

$$L_2 \sim N^{-\alpha}, \alpha \approx \frac{4}{d}$$

- (1) Larger d , slower convergence rate -> Curse of dimensionality!
- (2) d is the intrinsic dimension (dimension of data manifold), rather than full dimension e.g. number of pixels in an image
- (3) The other way around: knowing L curve, do some power law fitting, we obtain d !

Digression: Other methods to estimate the dimension of a manifold d

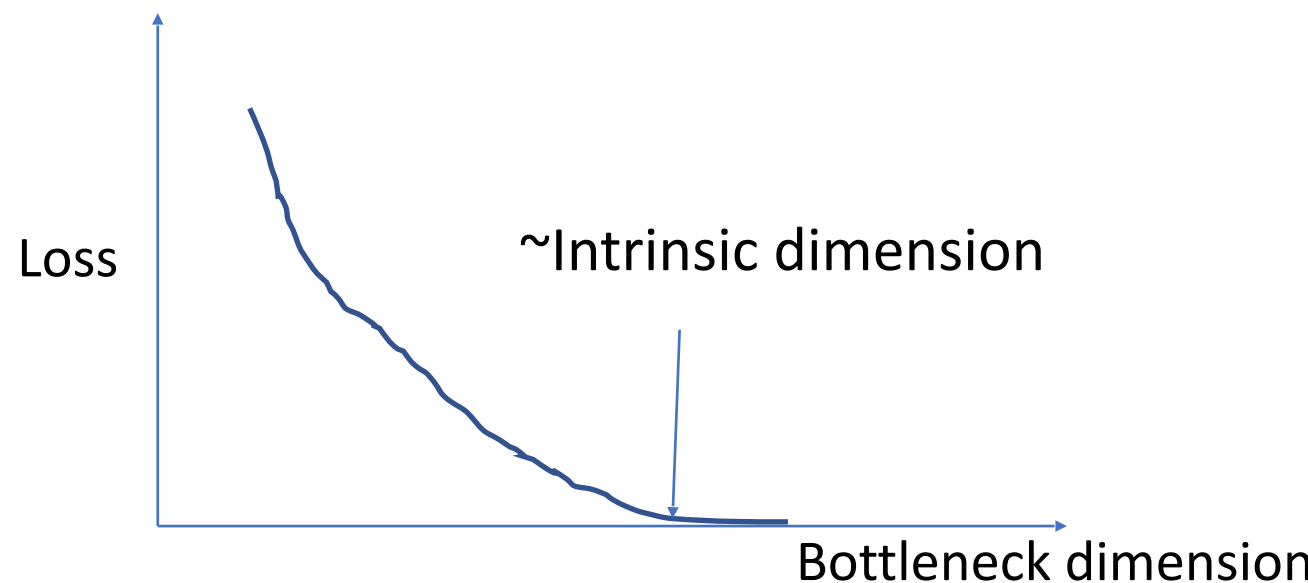
(1) Ideas from volume scaling law

Samples from uniform distribution

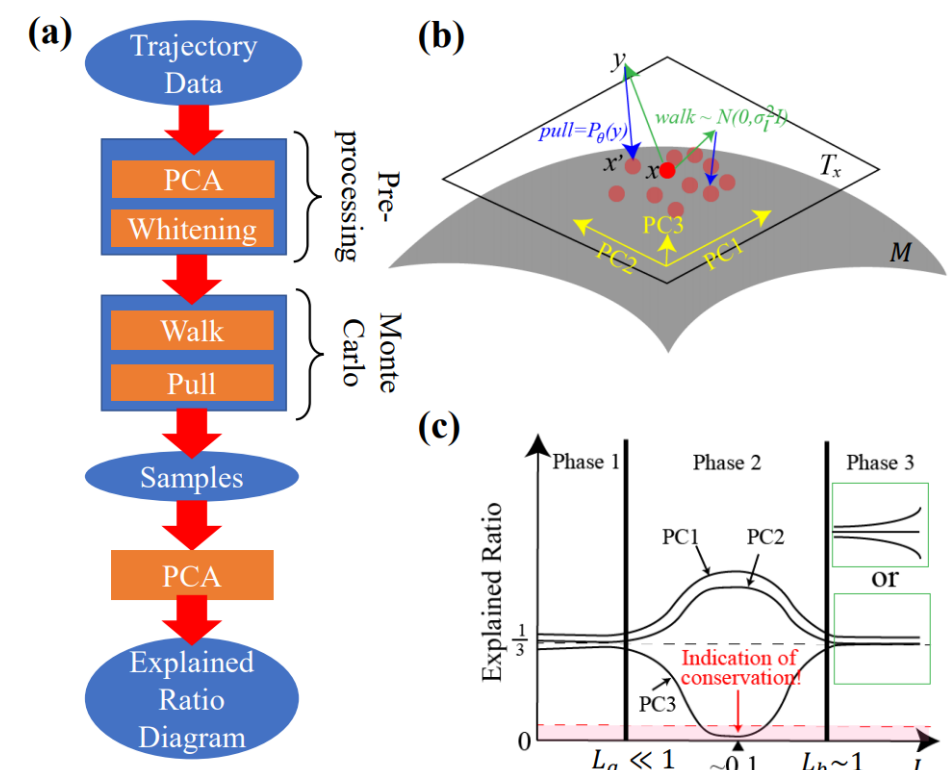
Number of Neighbors $\propto r^d$

$$d = \frac{\Delta \log(\# \text{ of Neighbors})}{\Delta \log r}$$

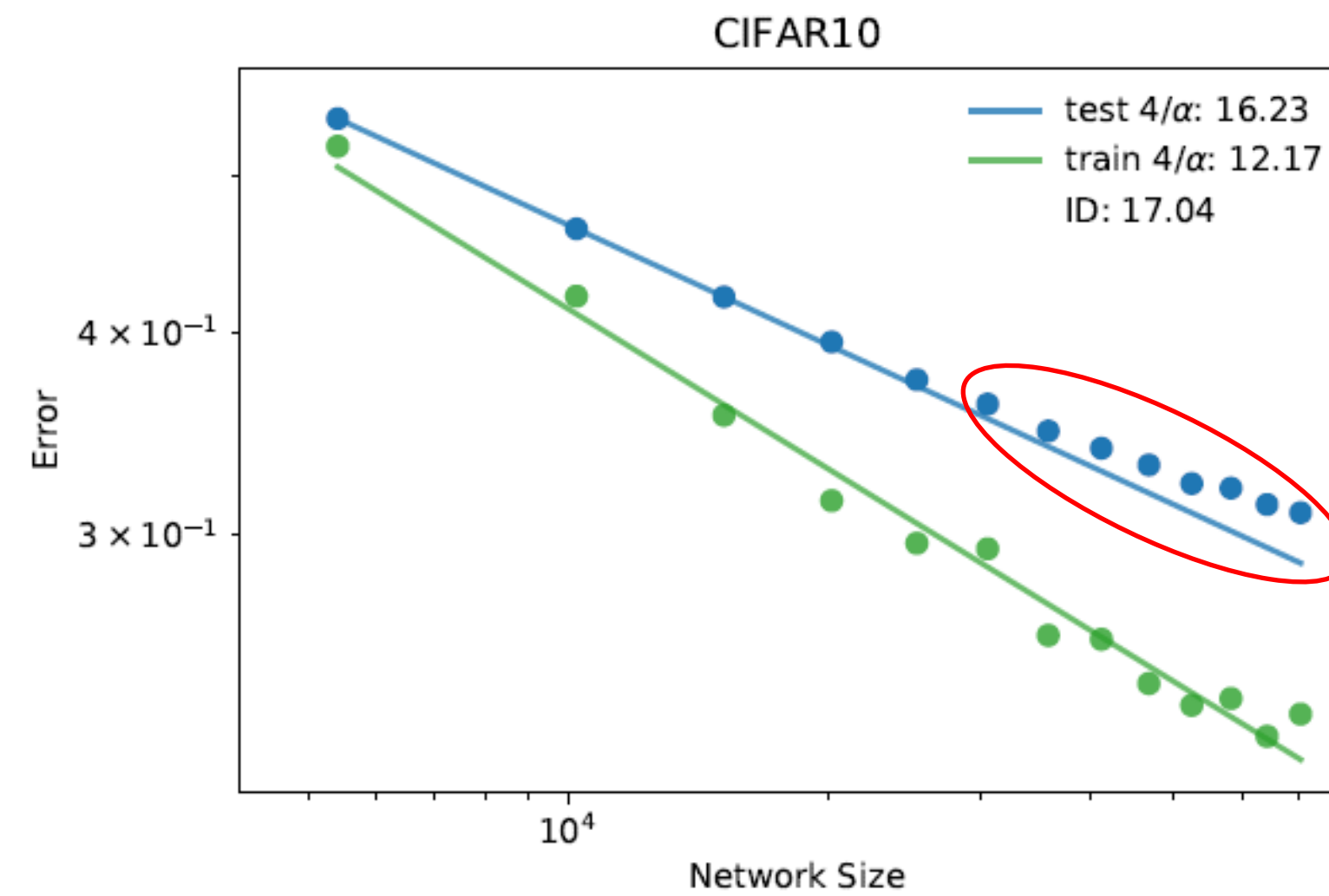
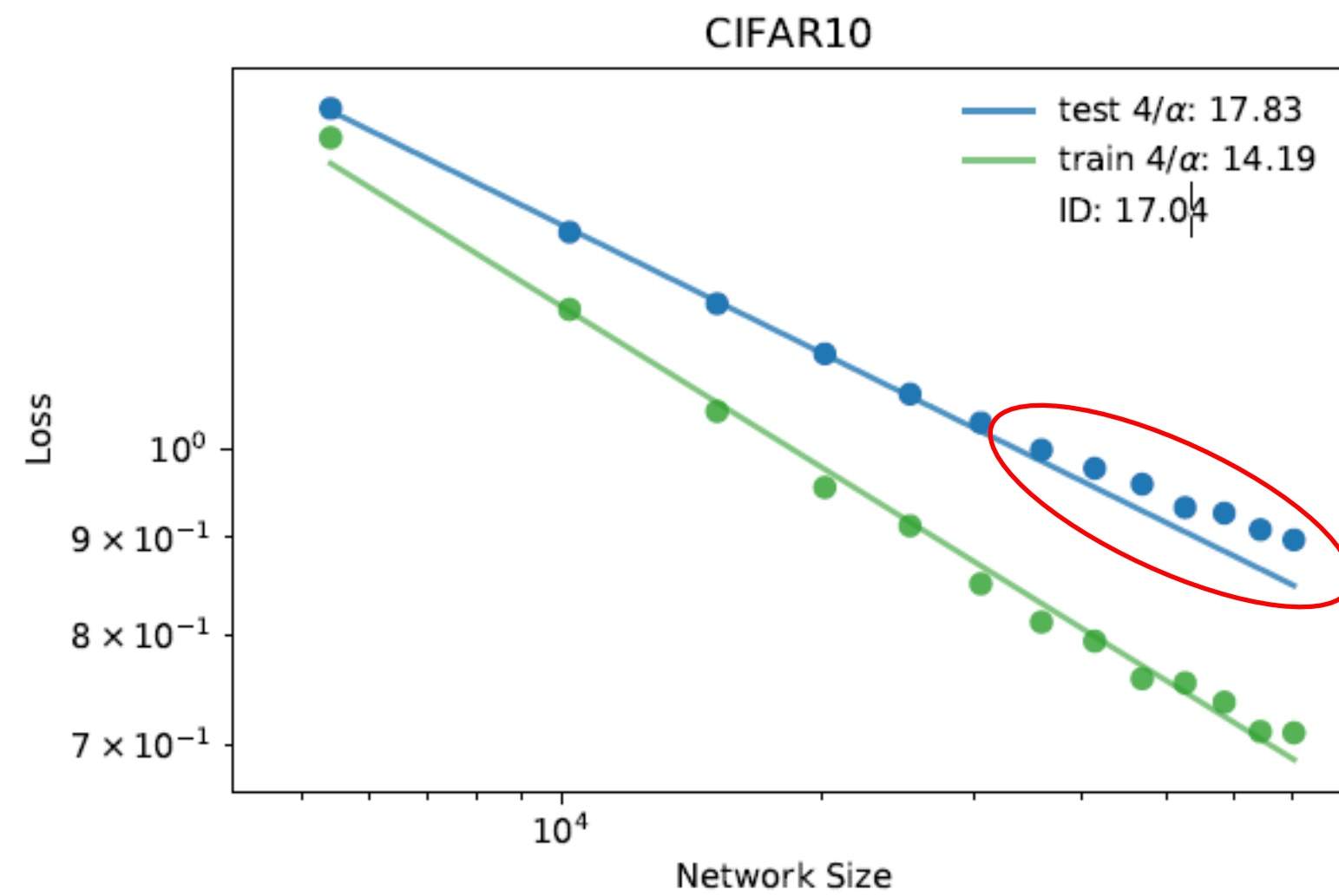
(2) (Variational) Autoencoders



(3) AI Poincare arXiv: 2011.04698



Power law breakdown



Deviation from power law!

Three possible reasons:

- (1) Overfitting
- (2) Data contains entropy or noise (Most probable)
- (3) Topological problem e.g. doughnuts

Discrete & Statistical => “Quantum physics”

arXiv > cs > arXiv:2303.13506

Computer Science > Machine Learning

[Submitted on 23 Mar 2023]

The Quantization Model of Neural Scaling

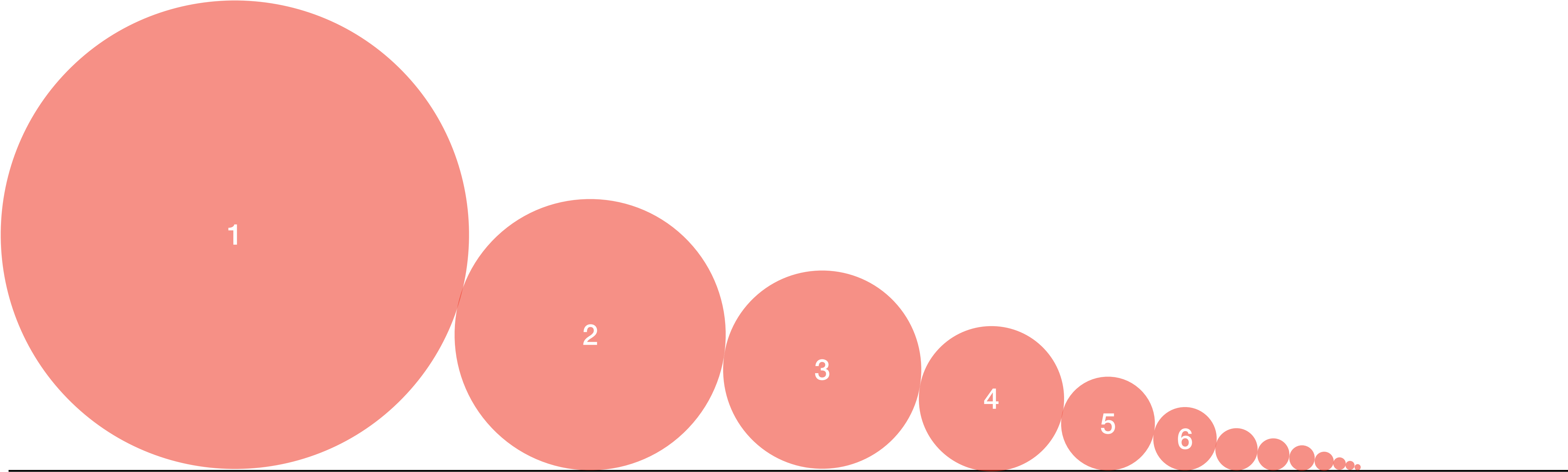
Eric J. Michaud, Ziming Liu, Uzay Girit, Max Tegmark

Motivation questions

1. Does the continuous manifold picture make sense for language modelling?
2. Is there a unified view for neural scaling laws and emergent abilities/grokking?
3. How to discover knowledge “quanta” in large language models?

Quantization Hypothesis

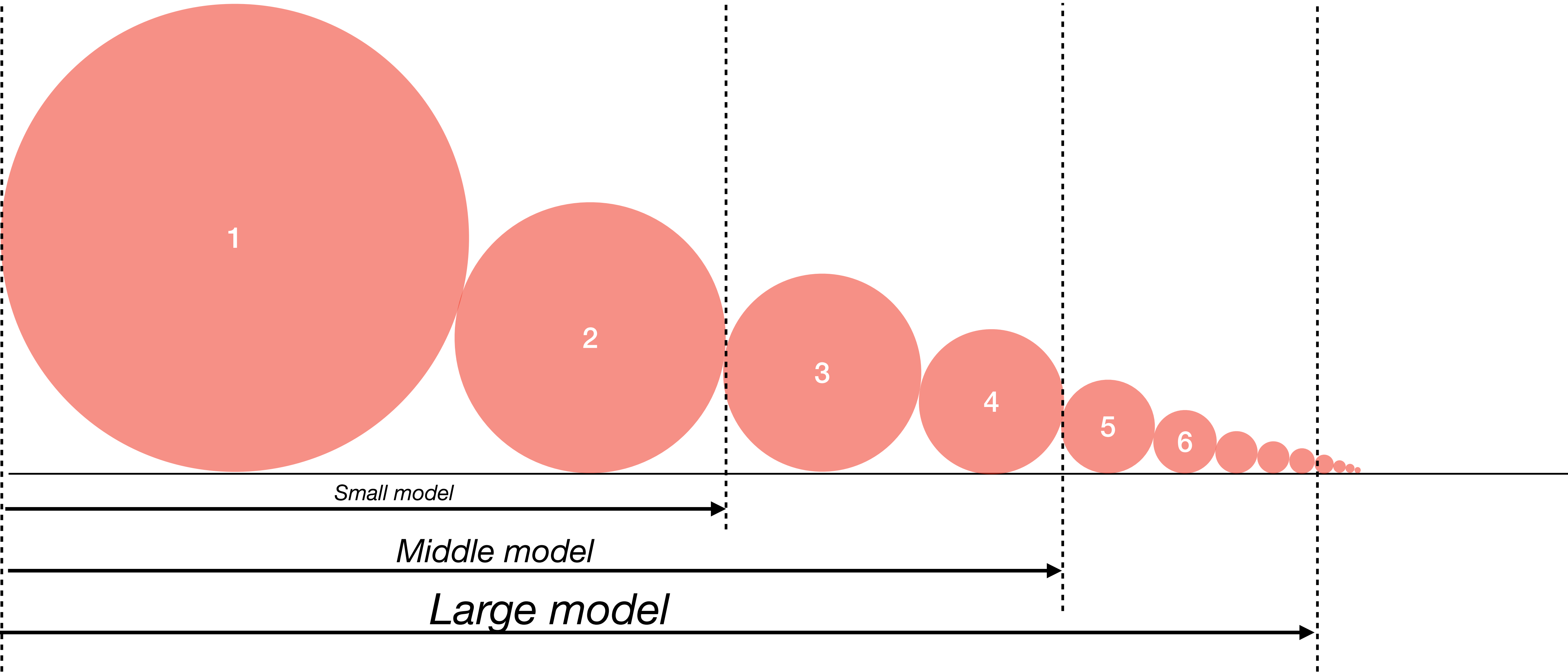
Knowledge quanta sequence



Size = Frequency (Importance)

Quantization Hypothesis

Knowledge quanta sequence



Quantization Hypothesis

In this paper, we conjecture the Quantization Hypothesis:

- QH1 Many natural prediction problems involve a discrete set of computations which are natural to learn and instrumental for reducing loss. We call these “quanta”. Model performance is determined by *which* quanta have been learned.
- QH2 Some abilities are more useful for reducing loss than others, leading to a natural ordering of the quanta. We call the ordered quanta the **Q Sequence**. Optimally trained networks should therefore learn the quanta in that order. The effect of scaling is to learn *more* of the quanta in the Q Sequence, so scaling performance is simply determined by *how many* quanta are successfully learned.
- QH3 The frequencies at which the quanta are used for prediction drop off as a power law.

Theory

We model the Quantization Hypothesis as follows. Let \mathbf{q} denote a bit string whose k^{th} bit $q_k = 1$ if the k^{th} quantum in the Q Sequence has been learned, and $q_k = 0$ otherwise. QH1 implies that the mean loss L is simply a function of \mathbf{q} . QH2 implies that when $n \equiv \sum_k q_k$ quanta have been learned, we have $q_k = 1$ for $k \leq n$. Let L_n denote the mean loss in this case.

From QH3, we have that the k^{th} quantum benefits prediction on a randomly chosen sample with probability

$$p_k = \frac{1}{\zeta(\alpha + 1)} k^{-(\alpha+1)} \propto k^{-(\alpha+1)} \quad (1)$$

for a Zipf power law $\alpha > 0$, where $\zeta(s) \equiv \sum_{k=1}^{\infty} k^{-s}$. Let us also assume that learning the k^{th} quantum reduces average loss from b_k before it is learned to a_k after it is learned on the samples where it is utilized.

If a_k and b_k are k -independent ($a_k = a$, $b_k = b$), then a model that has learned the first n quanta will have expected loss

$$\begin{aligned} L_n &= \sum_{k=1}^n a p_k + \sum_{k=n+1}^{\infty} b p_k = \sum_{k=1}^{\infty} a p_k + \sum_{k=n+1}^{\infty} (b - a) p_k \\ &\approx a + \frac{b - a}{\zeta(\alpha + 1)} \int_n^{\infty} k^{-(\alpha+1)} dk = a + \frac{b - a}{\alpha \zeta(\alpha + 1)} n^{-\alpha}. \end{aligned} \quad (2)$$

In other words, $L_{\infty} = a$ and $(L_n - L_{\infty}) \propto n^{-\alpha}$ is a power law.

Parameter scaling

Parameter scaling: In networks of finite size, only finitely many quanta can be learned – network capacity is a bottleneck. If we assume that all quanta require the same capacity of

C network parameters, and we have a network with N total parameters, roughly $n = N/C$ elements in the Q Sequence can be learned. We therefore expect loss to depend on the number of model parameters N like so:

$$L(N) = L_{N/C} \approx \frac{1}{\alpha \zeta(\alpha + 1)} \left(\frac{N}{C} \right)^{-\alpha} \propto N^{-\alpha}. \quad (3)$$

Data scaling (multi-epoch)

Data scaling (multi-epoch): For data scaling, we assume that a threshold of τ examples utilizing quantum k are needed in the training set in order for quantum k to be learned. τ can perhaps be thought of as the minimum number of examples on average requiring quantum k needed to uniquely specify its computation. Assuming network capacity is not a bottleneck, how many quanta will be learned? If we have a training set of D samples, then it will contain roughly Dp_1 samples utilizing quantum 1, Dp_2 samples utilizing quantum 2, and so on. If $p_k = \frac{1}{\zeta(\alpha+1)}k^{-(\alpha+1)}$, the last quantum n learned in the Q Sequence will then roughly be n such that $D\frac{1}{\zeta(\alpha+1)}n^{-(\alpha+1)} = \tau$ and so $n = (D/\tau\zeta(\alpha+1))^{1/(1+\alpha)}$. Under this model of how the training set size D influences which quanta are learned, we would therefore expect data scaling:

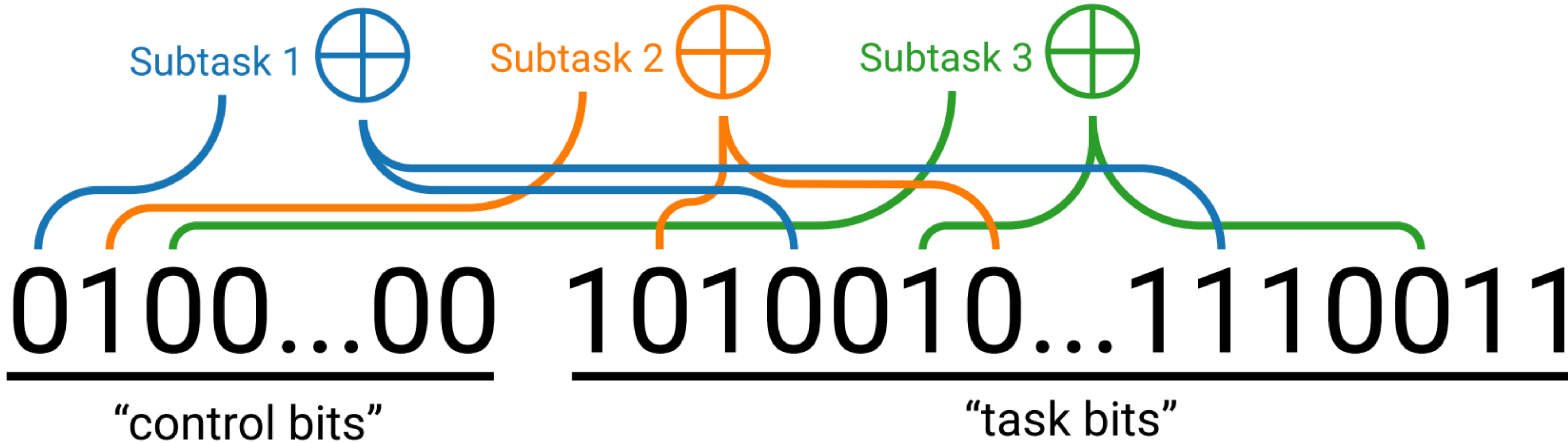
$$L(D) = L_{(D/\tau\zeta(\alpha+1))^{1/(1+\alpha)}} \approx \frac{1}{\alpha\zeta(\alpha+1)} \left(\frac{D}{\tau\zeta(\alpha+1)} \right)^{-\frac{\alpha}{\alpha+1}} \propto D^{-\frac{\alpha}{\alpha+1}}. \quad (4)$$

Data scaling (single-epoch)

Data scaling (single-epoch): In multi-epoch training, the information contained in the training dataset can bottleneck which quanta are learned. However, the rate of convergence of SGD can also bottleneck performance. For single-epoch training, a greater number of training samples allows one to train for longer. Assume that batches are large and that they contain effectively perfect gradient information. If quanta each reduce mean loss by an amount given by a power law, then the gradients incentivizing each quantum to form may also roughly follow a power law in magnitude. We might therefore expect that the number of training steps S to learn quantum k to be inversely proportional to use frequency p_k (more commonly useful quanta have larger gradients and are learned faster). Therefore if the first quantum requires T steps to be learned, then quantum n will require $Tn^{\alpha+1}$ steps to converge. As a function of the number of training steps S , the number of quanta learned is therefore $n = (S/T)^{1/(\alpha+1)}$, and so:

$$L(S) = L_{(S/T)^{1/(\alpha+1)}} \approx \frac{1}{\alpha \zeta(\alpha + 1)} \left(\frac{S}{T} \right)^{-\frac{\alpha}{\alpha+1}} \propto S^{-\frac{\alpha}{\alpha+1}}. \quad (5)$$

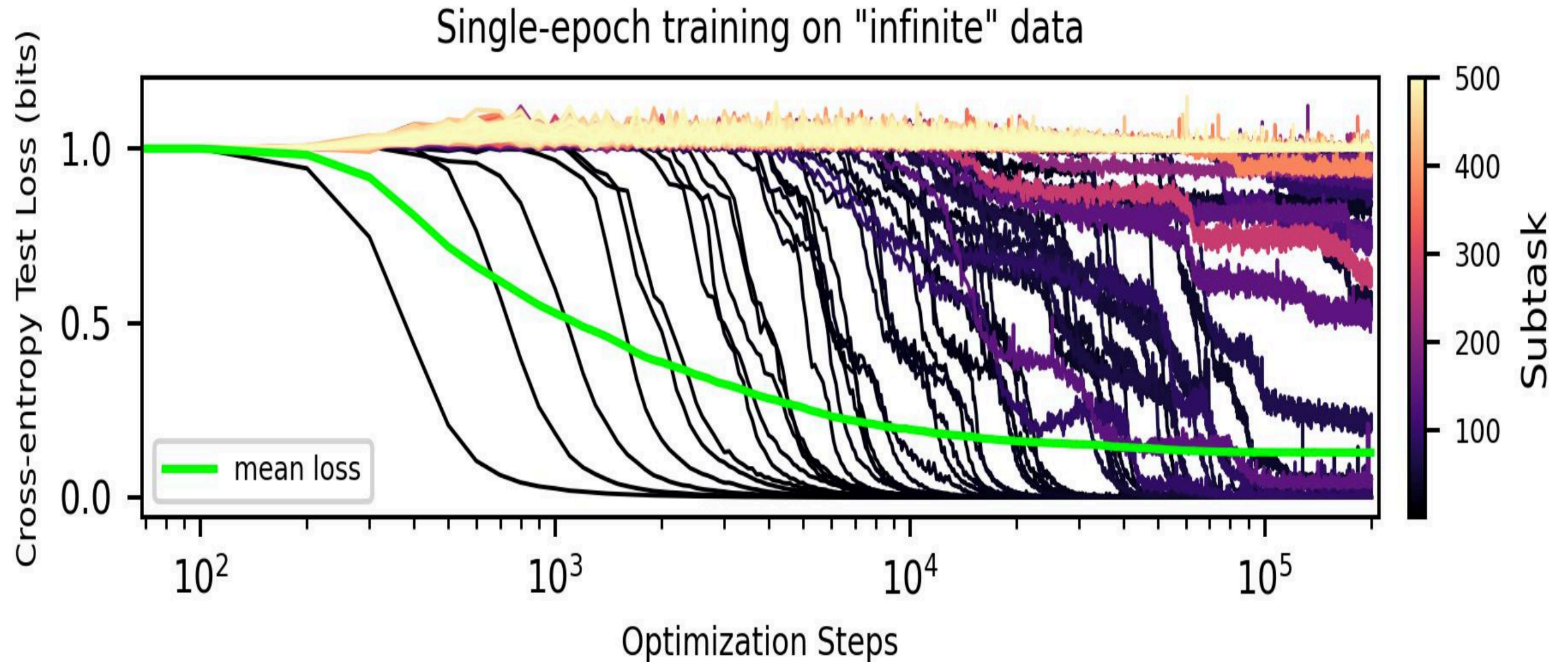
Toy example: Multitask sparse parity



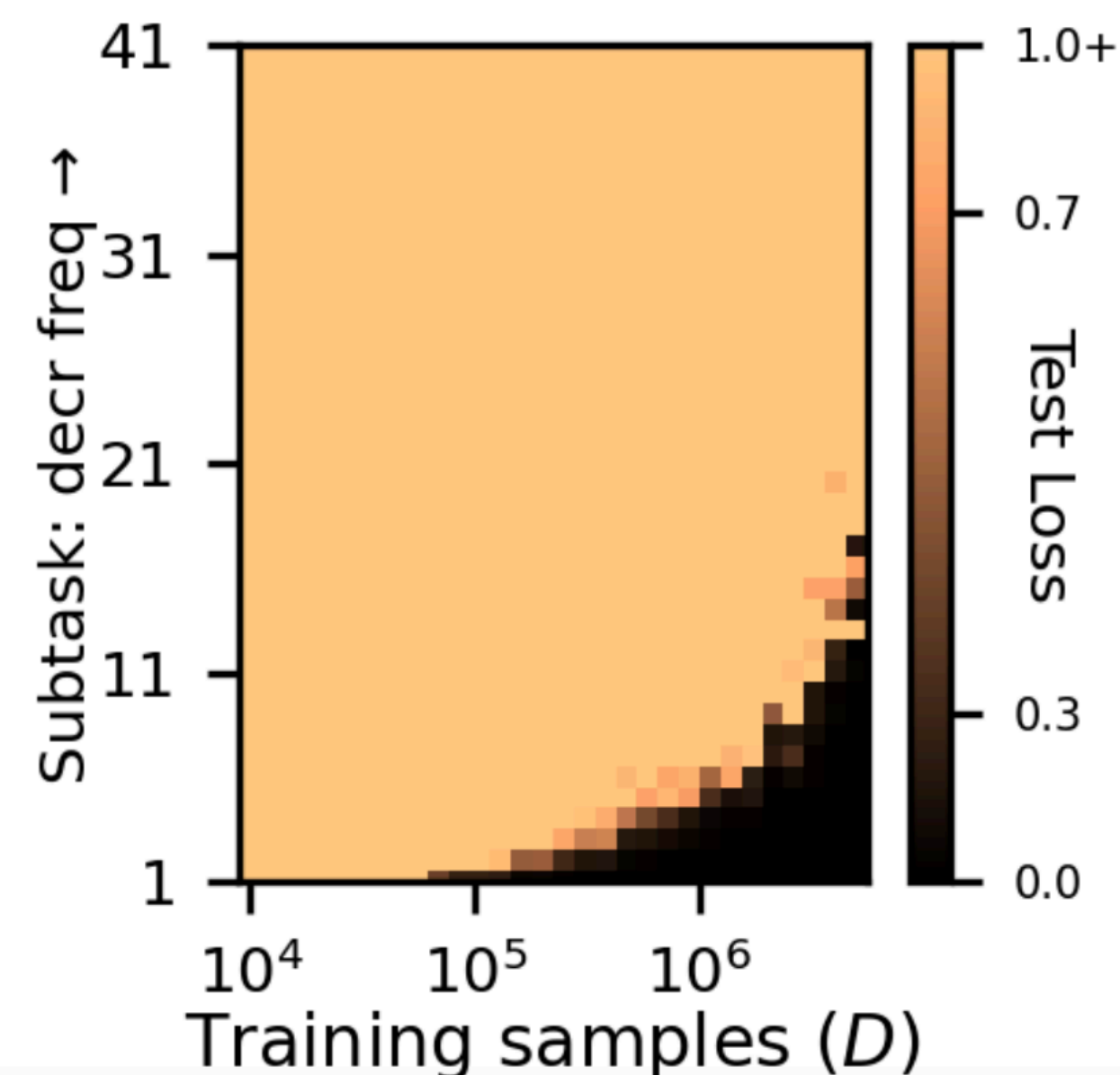
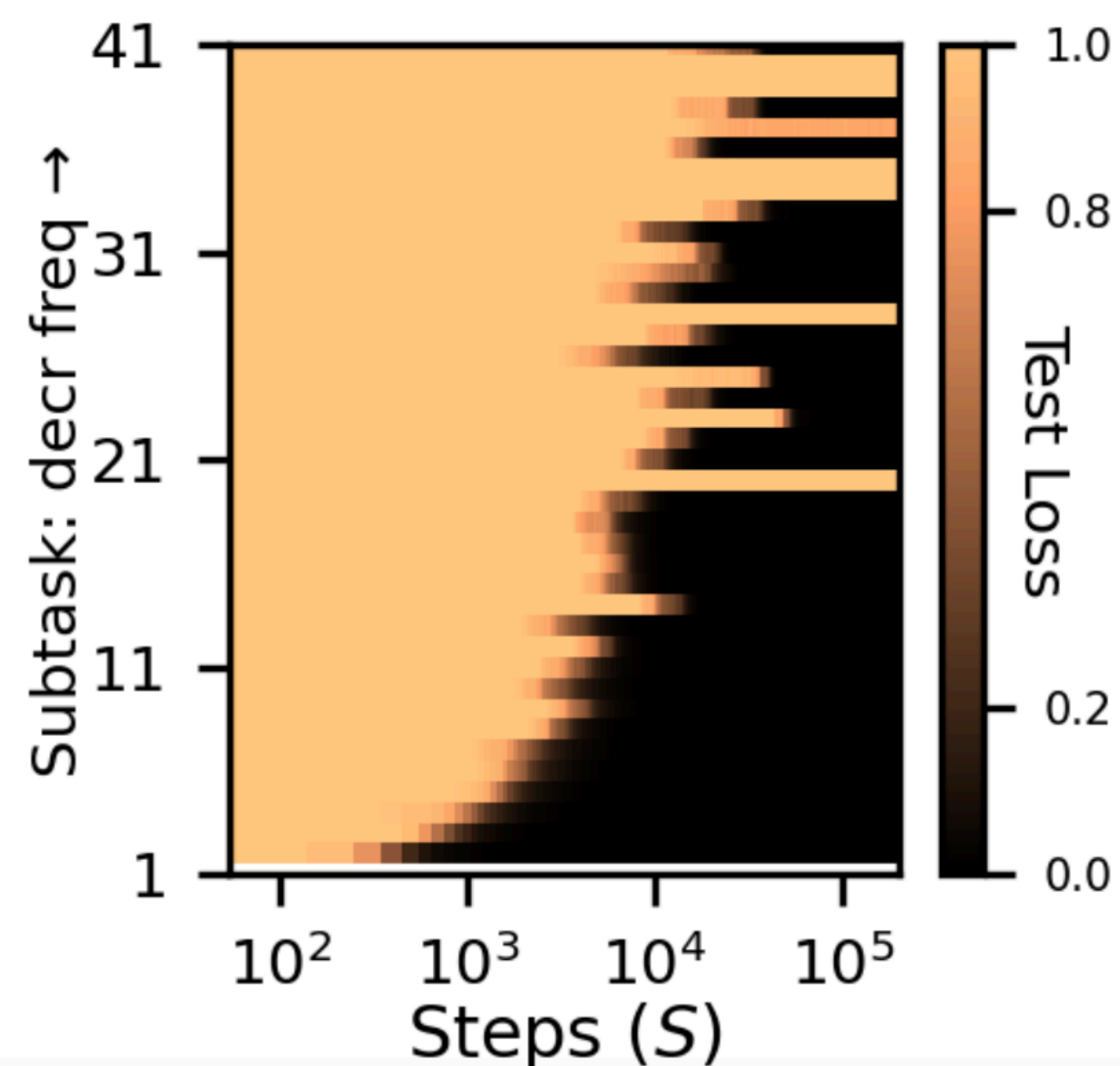
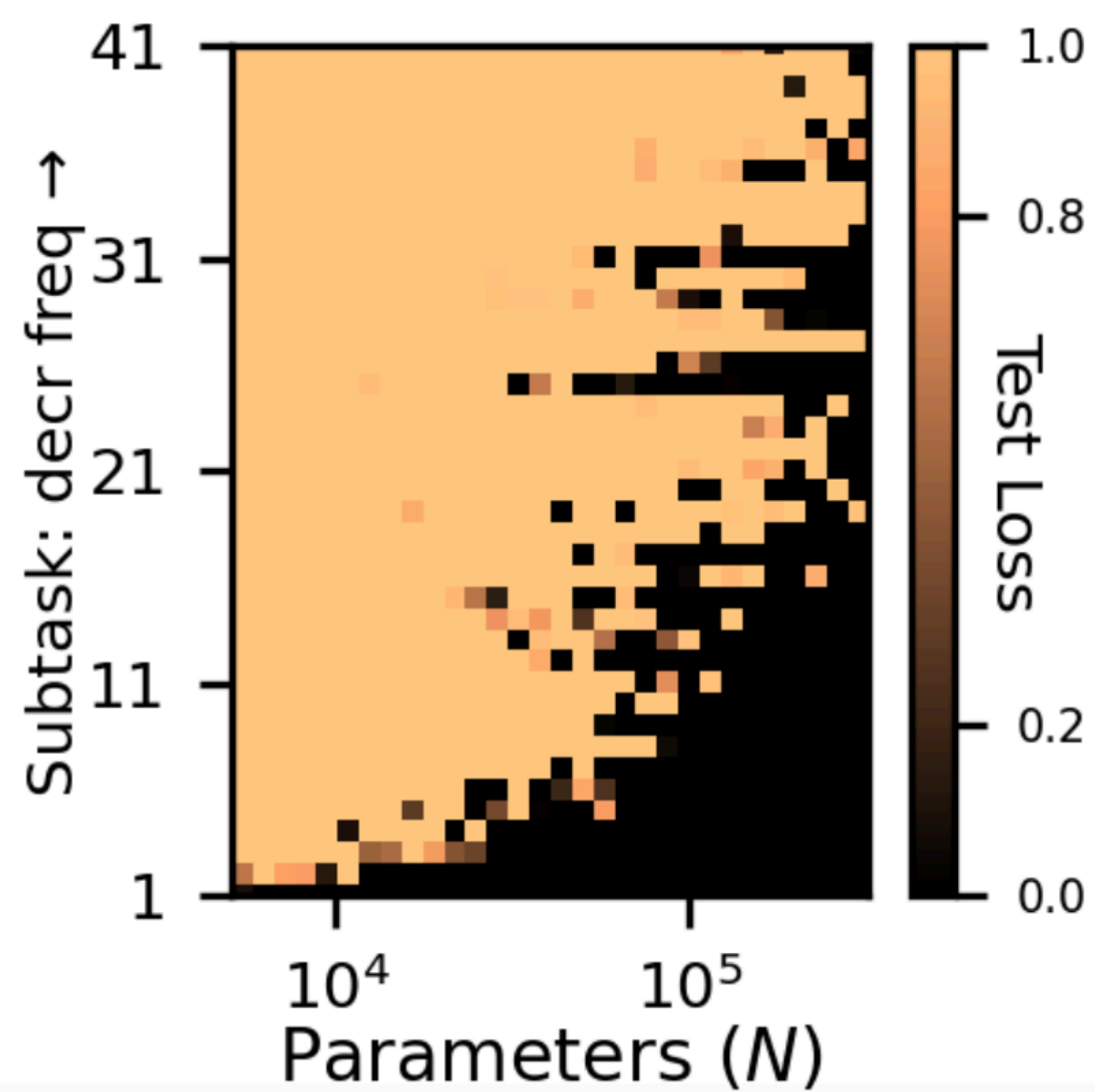
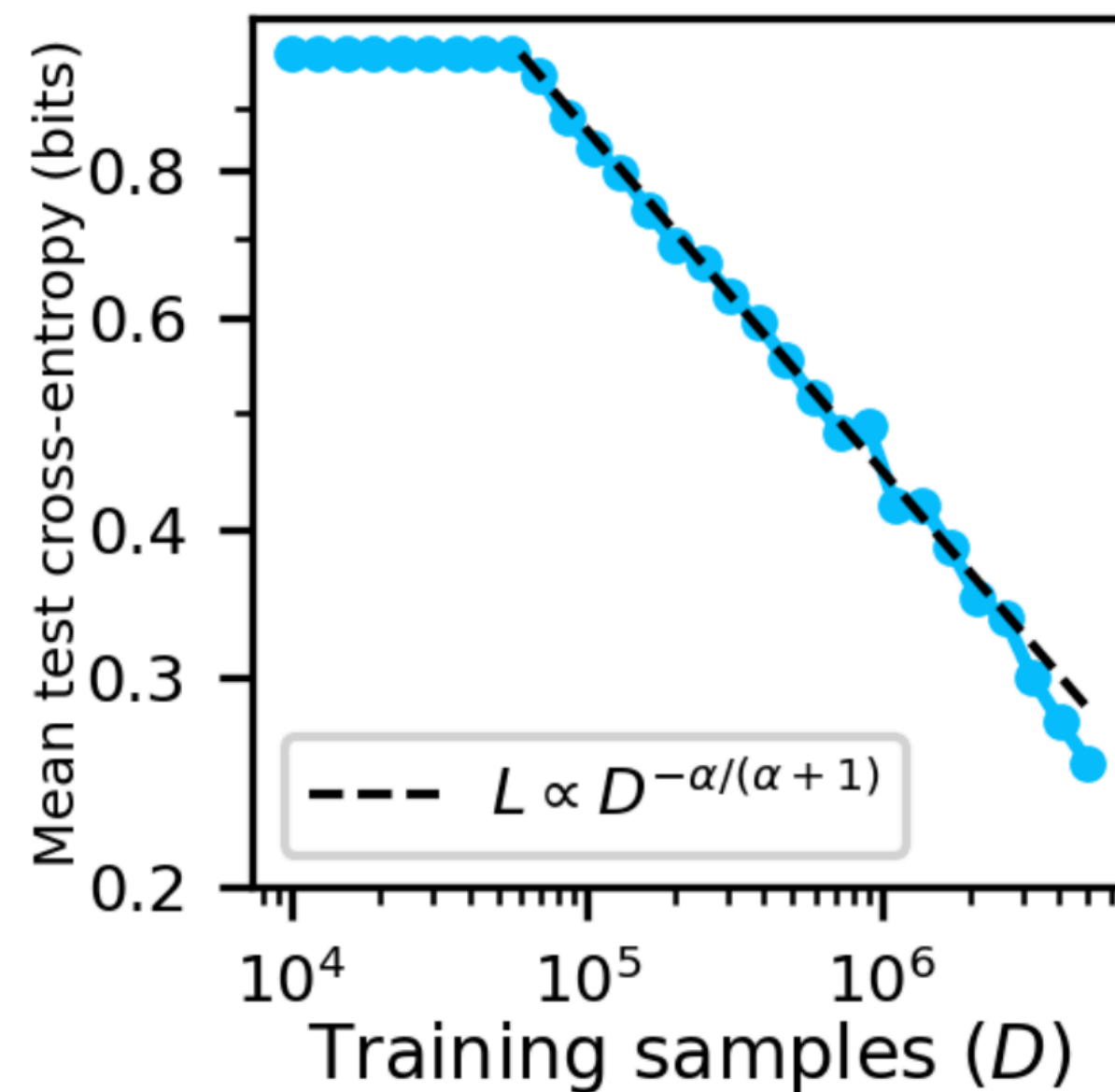
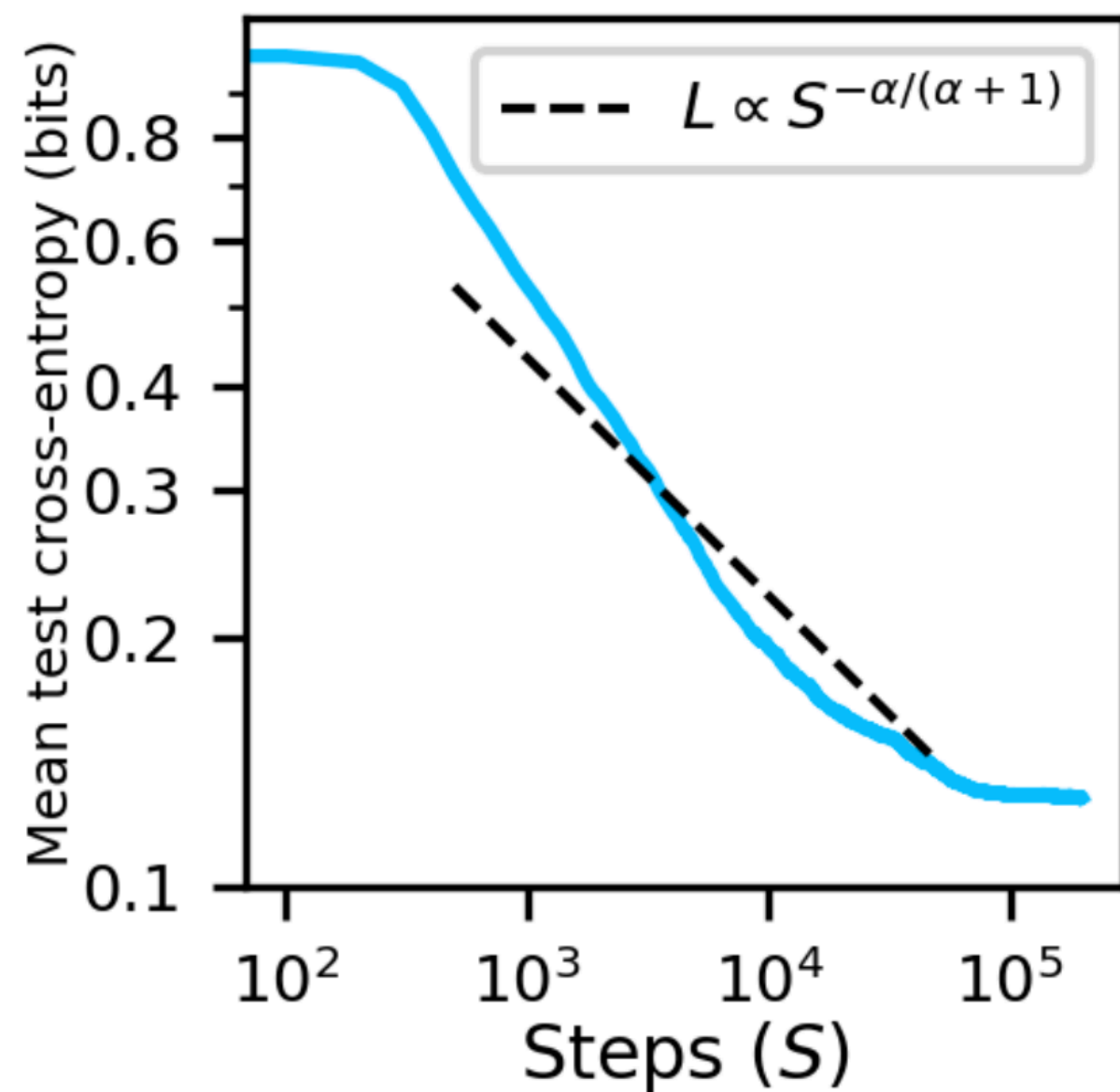
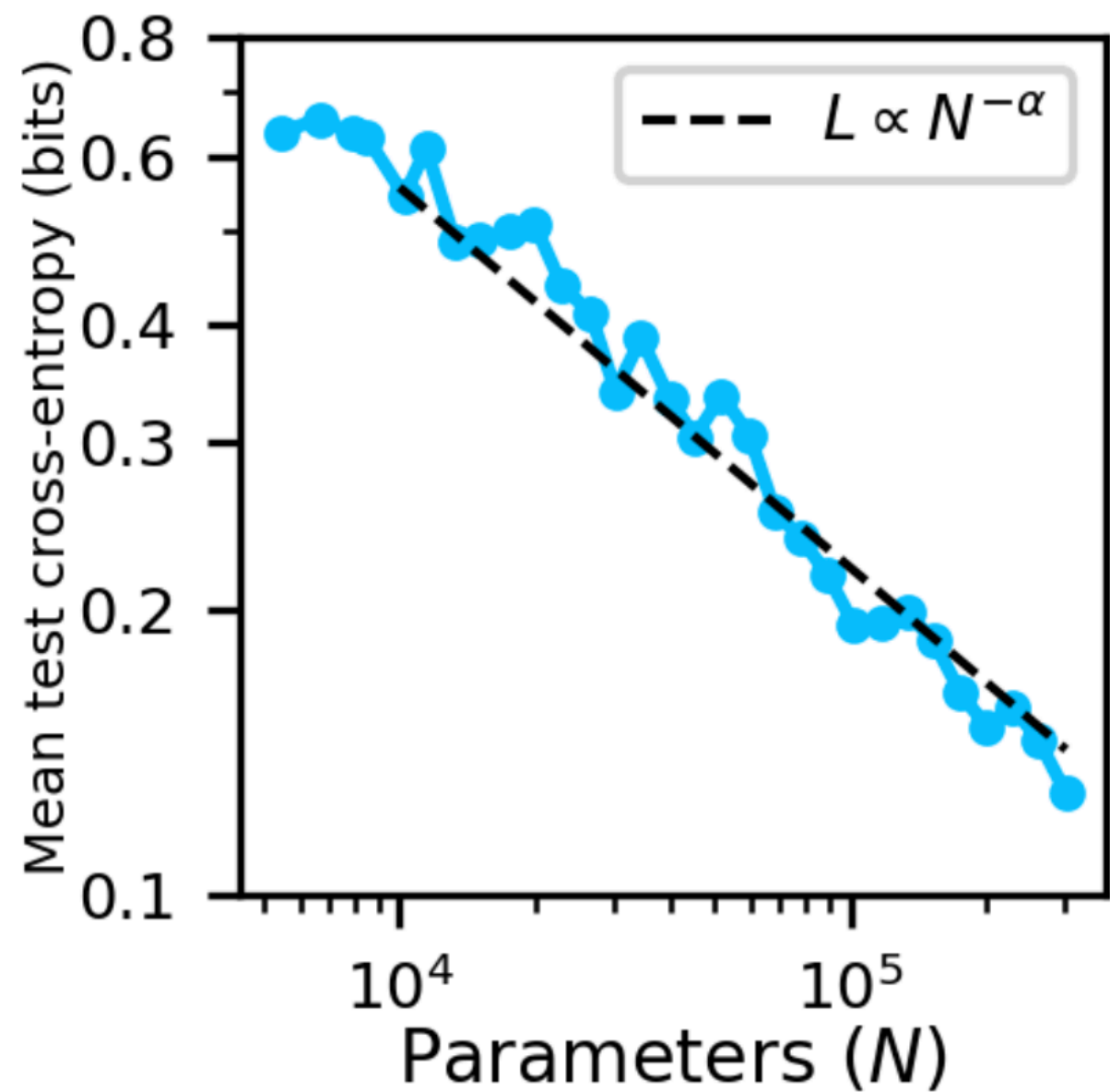
Toy example: dynamics

Individual task loss = grokking
Total loss = scaling law

A Quantization Model of Neural Scaling
arXiv: 2303.13506



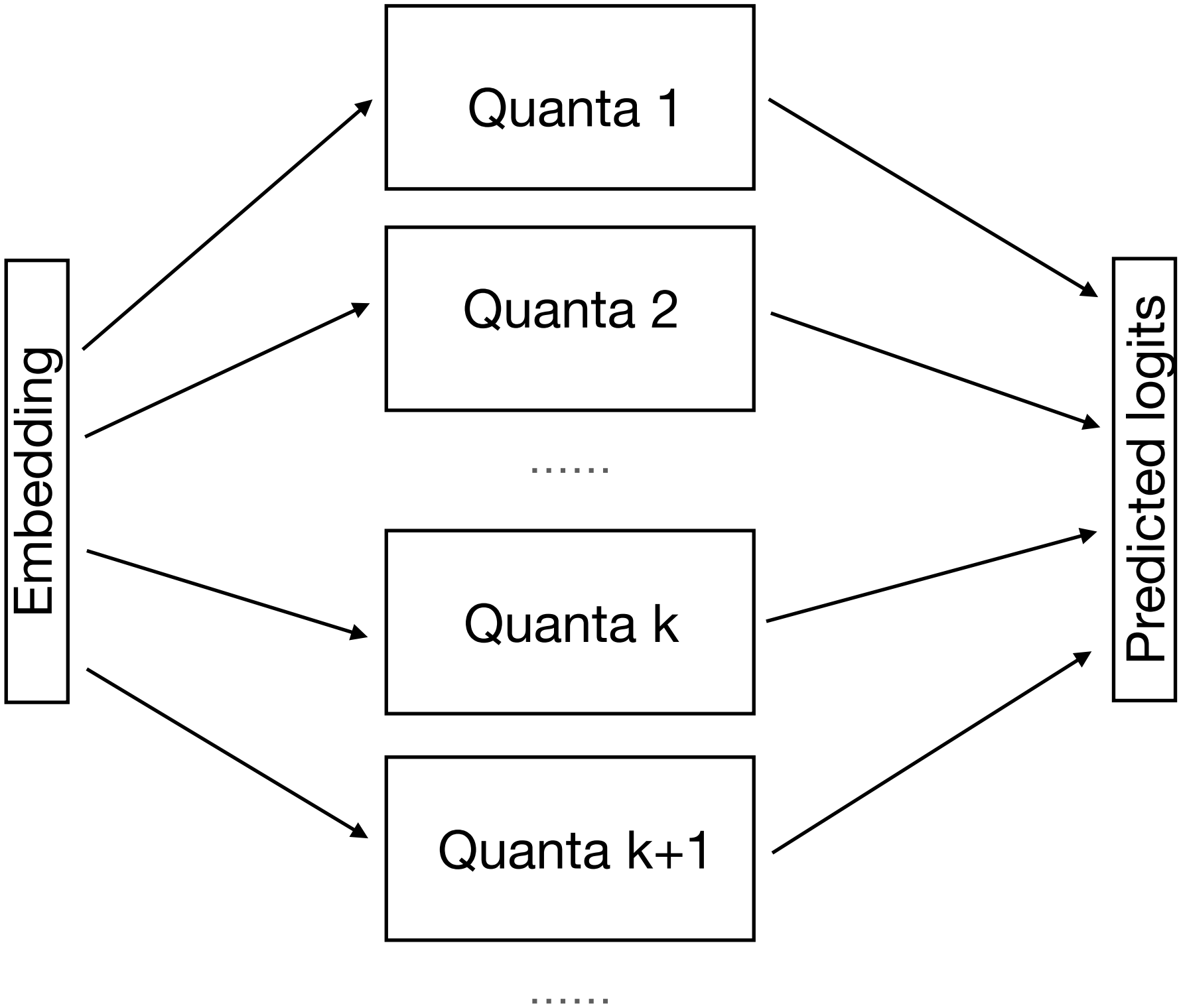
Toy example: scaling



Language Model

We now study how scaling curves for large language models decompose. For our experiments, we use the “Pythia” model sequence from Eleuther (EleutherAI 2023). These are decoder-only transformers of varying size trained on the same data in the same order – approximately 300 billion tokens of the train set of The Pile (Gao et al. 2020). Eleuther released 143 checkpoints for these models, spaced 1000 optimization steps apart. We can therefore study scaling w.r.t. model parameters N and training steps S . We evaluate the first seven models in the sequence, which range from 19m to 6.4b non-embedding parameters, on approximately 10 million tokens from the test set of The Pile. We record cross-entropy loss on every token. With this collection of loss values, we are able to study how neural scaling decomposes – rather than looking just at how mean test loss changes with scale, we can see how the distribution over losses changes with scale.

Quanta Discovery with Gradients (QDG)

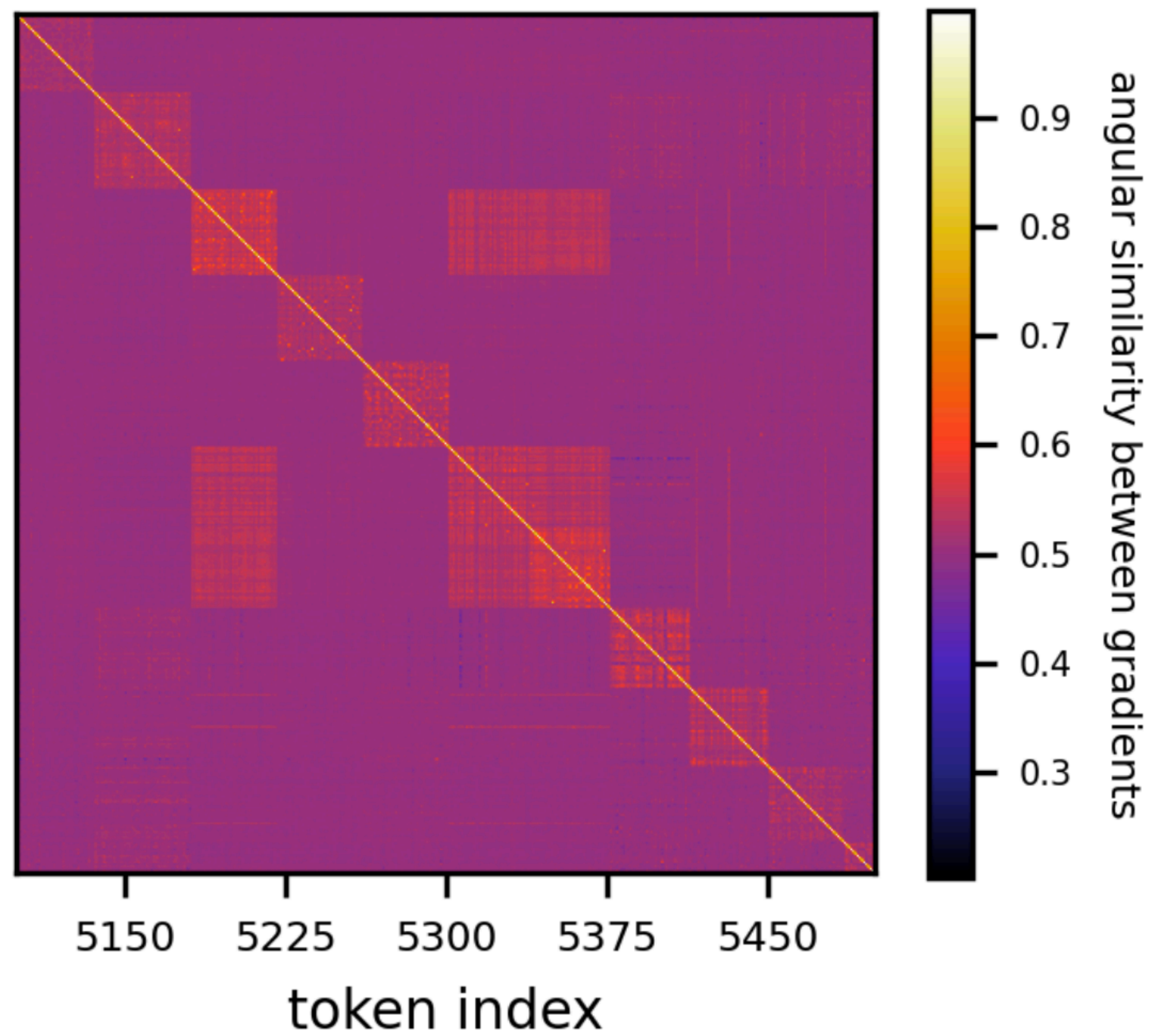


If two tokens belong to the same quanta, their activations/gradients should align.

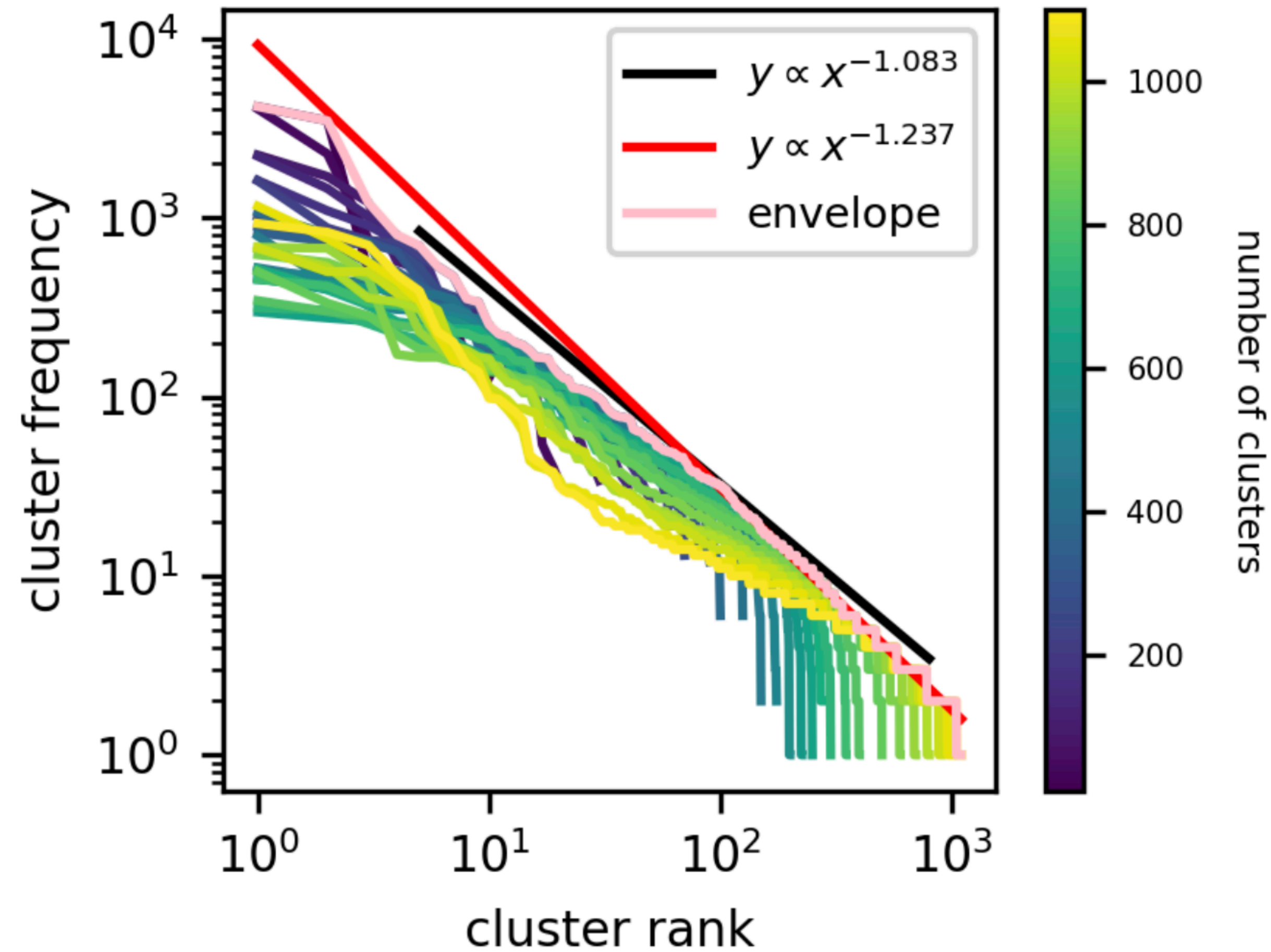
- QDG main idea:
- (1) Compute gradients for tokens.
 - (2) Clustering gradients. Each cluster is a quanta.

QDG results

Similarity Matrix



rank-frequency of clusters



Knowledge quanta

"Quanta" of LLM capabilities auto-discovered in natural text	
quantum for numerical sequence continuation (examples from cluster 50)	quantum for predicting newlines to maintain text width (examples from cluster 100)
<p>...ents his famous tonadas, a genre of the Venezuelan plains folk music.</p> <p>Track listing 01- Mi Querencia (Simón Díaz) 02- Tonada De Luna Llena (Simón Díaz) 03- Sabana (José Salazar/Simón Díaz) 04- Caballo Viejo (Simón Díaz) 05- Todo Este Campo Es Mío (Simón Díaz) 06- La Pena Del Becerrero (Simón Díaz) 07</p>	<p>...C REGRESSION. THE GOALS OF THIS VIDEO ARE TO PERFORM QUADRATIC REGRESSION ON THE TI84 GRAPHING CALCULATOR, DETERMINE HOW WELL THE REGRESSION MODEL FITS THE DATA, AND THEN MAKE PREDICTIONS USING THE REGRESSION EQUATION. IN STATISTICS, REGRESSION ANALYSIS INCLUDES ANY TECHNIQUES USED FOR MODELING \n</p>
<p>...sis supplied.) Appealing from that order, the city asserts (1) plaintiffs have no standing or right to maintain the action; (2) that the proposed road was in an undedicated part of the park; (3) that the proposed road was an access road and not a through street or part of the city's street system; (4</p>	<p>...ump is free software: you can redistribute it and/or modify # it under the terms of the GNU General Public License as published by # the Free Software Foundation, either version 3 of the License, or # (at your option) any later version. # # creddump is distributed in the hope that it will be useful,\n</p>
<p>... 4. _Introduction_ 5. Chapter 1: What Is Trust? 6. Chapter 2: Trust Brings Rest 7. Chapter 3: Who Can I Trust? 8. Chapter 4: The Folly of Self-Reliance 9. Chapter 5: Trust God and Do Good (Part 1) 10. Chapter 6: Trust God and Do Good (Part 2) 11. Chapter 7: At All Times 12. Chapter 8</p>	<p>... * Pursuant to 5TH CIR. R. 47.5, the court has determined that this opinion should not be published and is not precedent except under the limited circumstances set forth in 5TH CIR.\n</p>
<p>...gn of noncavitated lesion seen only when the tooth is dried; 2 = visible noncavitated lesion seen when wet and dry; 3 = microcavitation in enamel; 4 = noncavitated lesion extending into dentine seen as an undermining shadow; 5 = small cavitated lesion with visible dentine: less than 50% of surface; 6</p>	<p>... files (the // "Software"), to deal in the Software without restriction, including // without limitation the rights to use, copy, modify, merge, publish, // distribute, sublicense, and/or sell copies of the Software, and to permit // persons to whom the Software is furnished to do so, subject to the\n</p>
<p>...DynamicKey><Action>F1</Action><Label>F1</Label></DynamicKey> <DynamicKey><Action>F2</Action><Label>F2</Label></DynamicKey> <DynamicKey><Action>F3</Action><Label>F3</Label></DynamicKey> <DynamicKey><Action>F4</Action><Label>F4</Label></DynamicKey> <DynamicKey><Action>F5</p>	<p><!-- /** * Copyright (c) 2019, The Android Open Source Project * * Licensed under the Apache License, Version 2.0 (the "License"); * you may not use this file except in compliance with the License.\n</p>
<p>... GetPrepareVoteMsg = 0x07 PrepareVotesMsg = 0x08 GetQCBlockListMsg = 0x09 QCBlockListMsg = 0x0a GetLatestStatusMsg = 0x0b LatestStatusMsg = 0x0c PrepareBlockHashMsg = 0x0d GetViewChangeMsg = 0x0e PingMsg = 0x0f</p>	<p>...f maturity and an underdeveloped sense of responsibility, leading to recklessness, impul- sivity, and heedless risk-taking.... Second, children are more vulnerable... to negative influences and outside pressures, including from their family and peers; they have limited contro[1] over their own envi-\n</p>

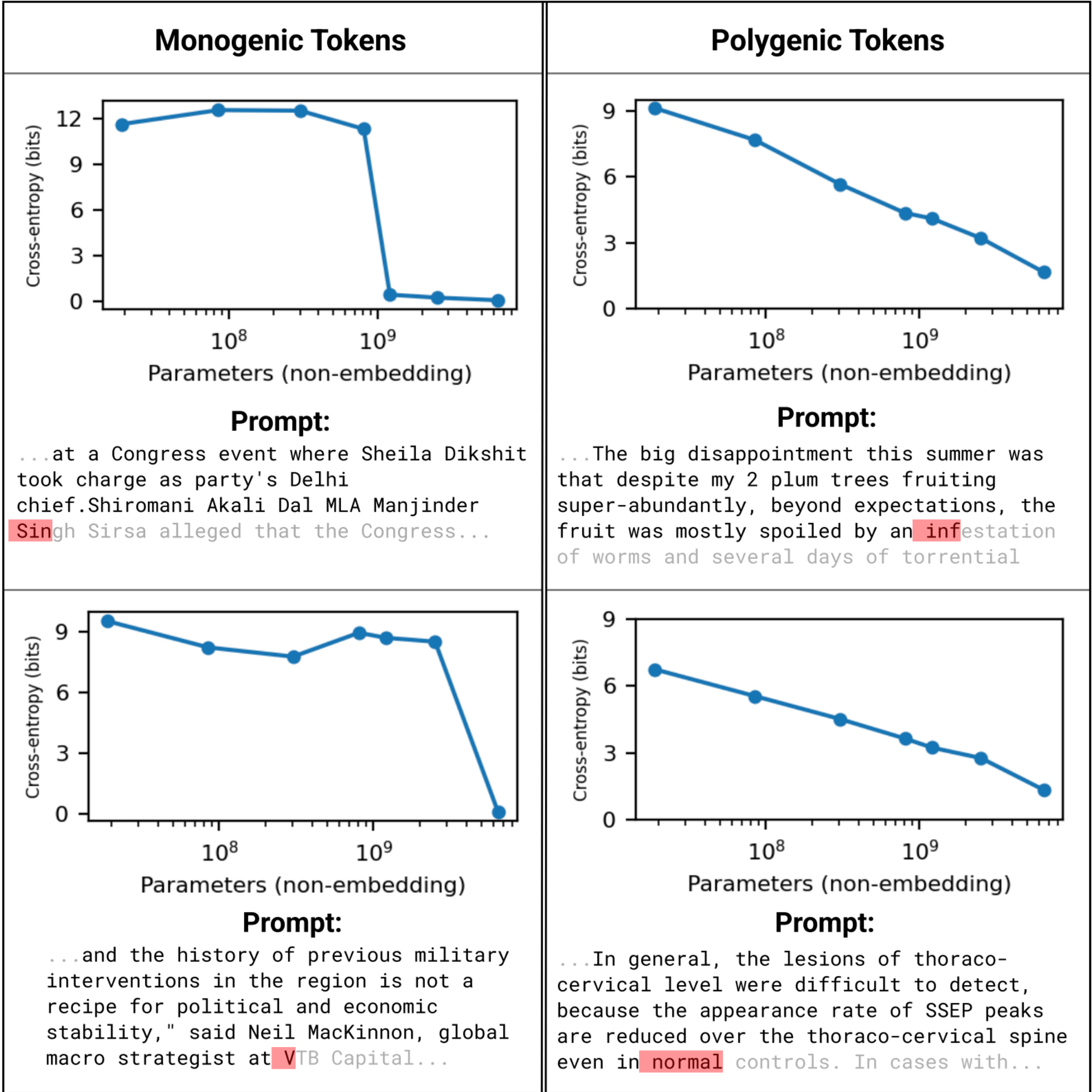
Knowledge quanta

Examples from Cluster 146: comma after day of month	Examples from Cluster 269: "s" after start year of decade
<p>After his tweet went viral Aslan apologized on Twitter saying "it's not like me" to use profanity.</p> <p>I should not have used a profanity to describe the President when responding to his shocking reaction to the #LondonAttacks. My statement: pic.twitter.com/pW69jjpoZy - Reza Aslan (@rezaaslan) June 4,</p>	<p>Romford Ice Arena</p> <p>Romford Ice Arena was an ice rink located in Romford in the London Borough of Havering, England. The venue was built in the 1980s</p>
<p>Sam Willard</p> <p>Samuel Steven Willard (born September 9,</p>	<p>...ownloadable formats: PDF</p> <p>The rings were stamped with a distinctive Kleinberg logo. Although the novel continues to be the dominant medium of the crime-mystery-detective narrative, short stories by these contemporary authors may be found in numerous anthologies of the genre published during the 1990s</p>
<p>215 U.S. 437 (1910) MECHANICAL APPLIANCE COMPANY v. CASTLEMAN. No. 48. Supreme Court of United States. Argued December 3, 1909. Decided January 3,</p>	<p>...as the Founder and First Director of the Institute of Atomic Physics (IFA) in Bucharest, Romania. He became a titular member of the Romanian Academy in 1946; stripped of membership by the new communist regime in 1948, he was restored to the Academy in 1955.</p> <p>University teaching During the early 1960s</p>
<p>485 F.2d 283 73-2 USTC P 9685, 179 U.S.P.Q. 450 GEORATOR CORPORATION, Appellee,v.UNITED STATES of America, Appellant. No. 73-1187. United States Court of Appeals,Fourth Circuit. Argued June 4, 1973.Decided Oct. 2,</p>	<p>...king down Ryan Farish's "Beautiful" CD after hearing "Full Sail" played during TWC's "Local On The 8's" segment. [Farish's music clips and a streaming Internet broadcast here] Yesterday, visitor Greg Davidson commented that he was searching for songs played on the local forecast back in the late '80s</p>
Examples from Cluster 278: colon after CSS property	Examples from Cluster 292: "://" after "http"
<pre>.rickshaw_graph.detail { pointer-events: none; position: absolute; top: 0; z-index: 2; background: rgba(0, 0, 0, 0.1); bottom: 0; width:</pre>	<pre>...##### # TeslaCrypt Ransomware Payment Sites domain blacklist (TC_PS_DOMBL) # # # # # # For questions please refer to: # # https:// #</pre>
<pre>@import '../.../assets/sass/spin'; .app-header { background-color: #282c34; min-height: 100vh; display:</pre>	<p>...to that document rather than overwrite it. If it does not exist, it should insert the new document to the collection.</p> <p>When I run the below code, I am getting an error: MongoError: The dollar (\$) prefixed field '\$push' in '\$push' is not valid for storage. I put this together based on the docs: https://</p>
<pre>...o work. I tried \$("#plane").toggle(".plane-right,.plane-left") inside the listener but that didn't do the trick. And the CSS class .plane-right { background-image: url("../img/zoomzoom.png"); background-position: center; background-repeat: no-repeat; background-size: 100%; height:</pre>	<p>Gruber, Martin A. Views of the National Zoological Park in Washington, DC, showing Exhibit. 1919. Retrieved from the Digital Public Library of America, http://</p> <p>... it be discontinued? I heard Java Swing is discontinued and no more future enhancements will be made. As a Beginner what should I learn.</p> <p>A:</p> <p>JavaFX is more recent and can be considered as the successor of Swing. There is many very useful features added in JavaFX. See here some key features : https://</p>

Monogenic vs Polygenic tokens

A Quantization Model of Neural Scaling

arXiv: 2303.13506



Neural Scaling Laws for Science

- For scientific problems, better inductive biases => better scaling law

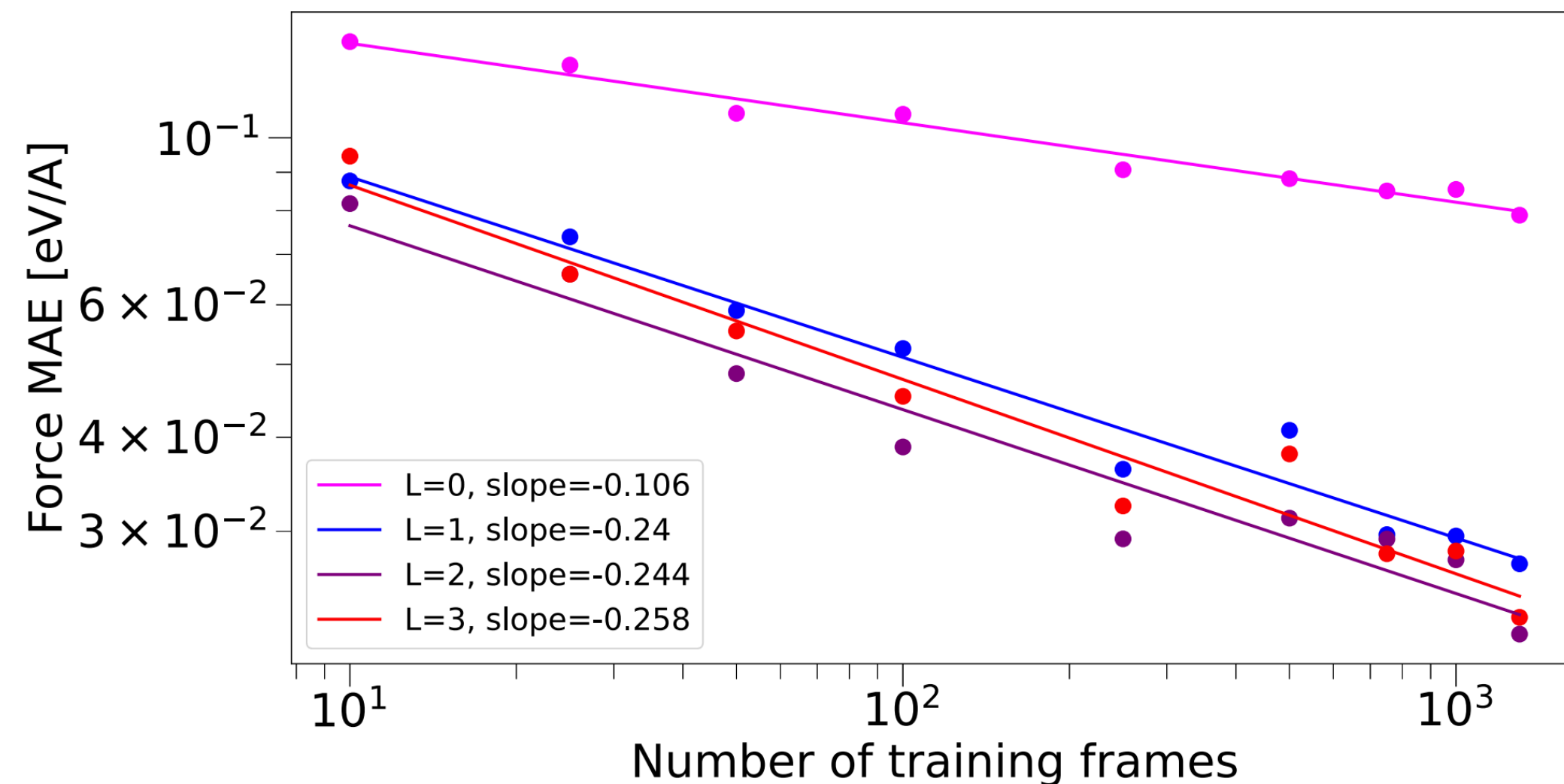
Article | [Open Access](#) | [Published: 04 May 2022](#)

E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials

[Simon Batzner](#), [Albert Musaelian](#), [Lixin Sun](#), [Mario Geiger](#), [Jonathan P. Mailoa](#), [Mordechai Kornbluth](#), [Nicola Molinari](#), [Tess E. Smidt](#) & [Boris Kozinsky](#)

Nature Communications **13**, Article number: 2453 (2022) | [Cite this article](#)

26k Accesses | 68 Citations | 86 Altmetric | [Metrics](#)



arXiv > cs > arXiv:2303.14269

Search...

Help | Advanced

Computer Science > Machine Learning

[Submitted on 24 Mar 2023]

The Exact Sample Complexity Gain from Invariances for Kernel Regression on Manifolds

[Behrooz Tahmasebi](#), [Stefanie Jegelka](#)

In practice, encoding invariances into models helps sample complexity. In this work, we tighten and generalize theoretical results on how invariances improve sample complexity. In particular, we provide minimax optimal rates for kernel ridge regression on any manifold, with a target function that is invariant to an arbitrary group action on the manifold. Our results hold for (almost) any group action, even groups of positive dimension. For a finite group, the gain increases the "effective" number of samples by the group size. For groups of positive dimension, the gain is observed by a reduction in the manifold's dimension, in addition to a factor proportional to the volume of the quotient space. Our proof takes the viewpoint of differential geometry, in contrast to the more common strategy of using invariant polynomials. Hence, this new geometric viewpoint on learning with invariances may be of independent interest.

Neural Scaling Laws for Science

- If the quantisation model holds widely, we will know the quanta of “science”.
A general method to extract knowledge from neural networks.

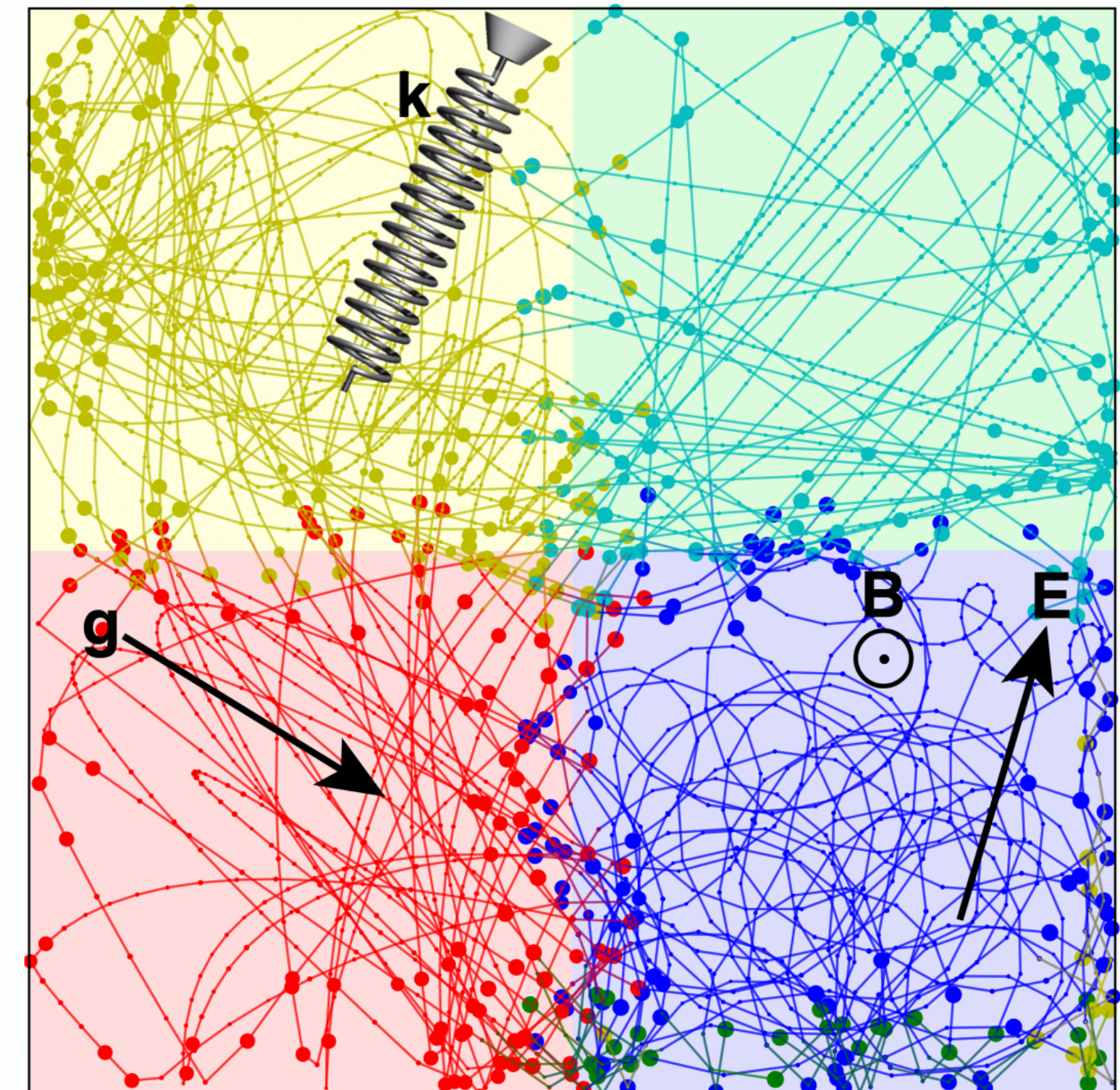
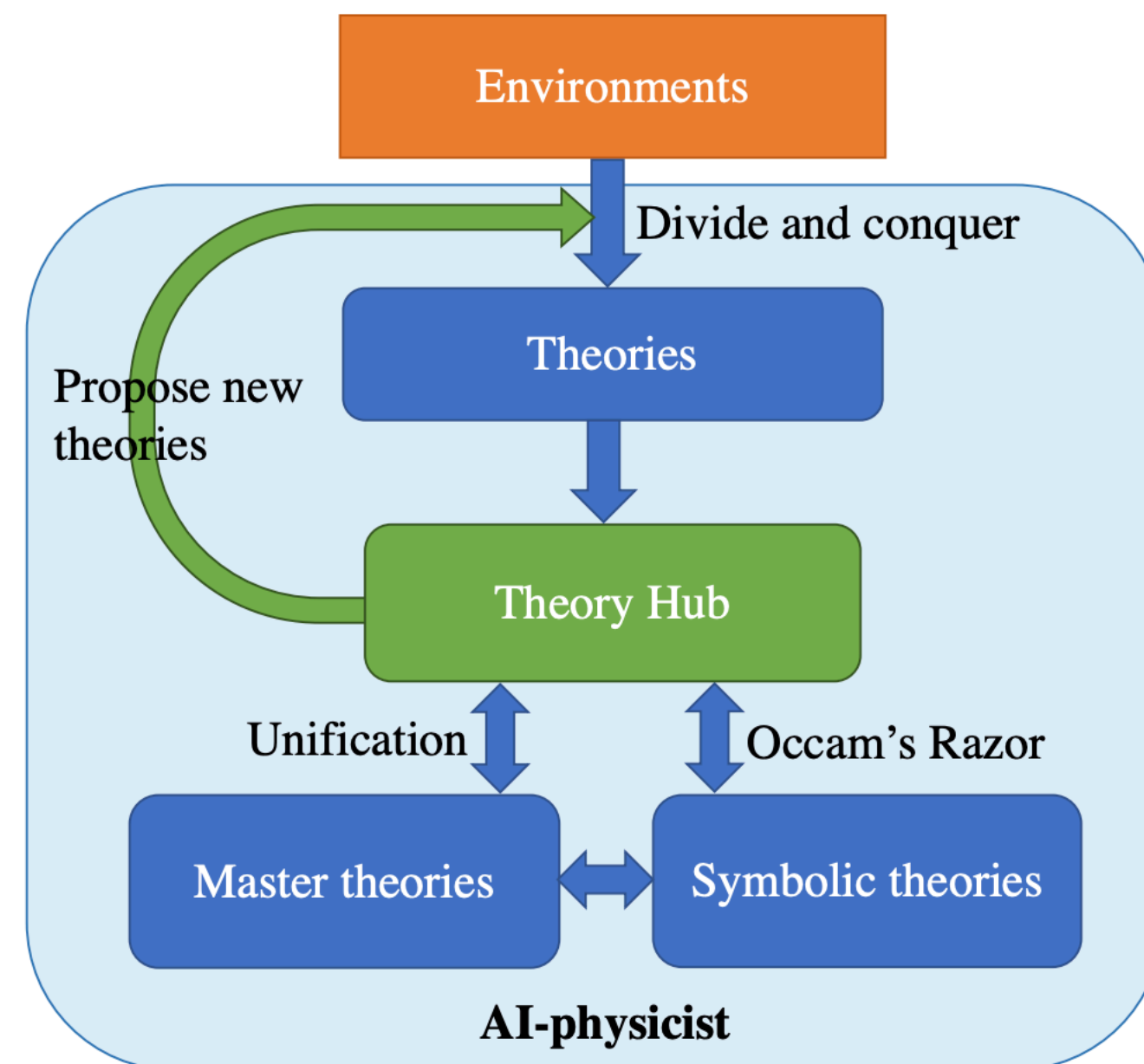
arXiv > physics > arXiv:1810.10525

Physics > Computational Physics

[Submitted on 24 Oct 2018 (v1), last revised 2 Sep 2019 (this version, v4)]

Toward an AI Physicist for Unsupervised Learning

Tailin Wu (MIT), Max Tegmark (MIT)



AI (LLM) and Science

圆桌讨论：人工智能（大语言模型）和科学的关系

圆桌主题1: AI (LLM) for Science

1. AI被称为“第五范式”（继经验范式、理论范式、模拟范式、数据驱动范式之后）。在你的研究中，AI是否已经改变了研究范式？AI有希望改变什么问题？
2. 在你的研究中，是否更大规模的AI模型会带来更好的效果？“Scale is all you need”有什么局限性？
3. 大模型能在科学发现上起什么作用？比如发现系统控制方程，发现系统重要概念和对称性等？

圆桌主题2:

Science (LLM) for AI, Science of AI

1. 你的研究中，AI出现过哪些有趣/奇怪的行为？比如相变，涌现，对Scaling Law的违反等等？
2. 希望针对这部分内容（神经标度律是大脑神经元数量与连接方式的变化会随着动物体型的增大或减小而发生调整，以维持相对稳定的神经系统功能）——探讨一些具体的事例，引发大家对于连接方式和涌现的关系的猜测
3. 幂律缩放除了能预测一些现象/结果，还能用来做什么？
4. 我们在大模型上发现的一些能力，真的在小模型上不可能复制成功吗？有没有可能我们只是没找到“撬动神经连接方式变化”的技巧？

圆桌主题3:

Interpretability/trustworthy of AI

1. 目前的AI大体上是黑盒。如何提高AI for Science的可解释性？你的研究有哪些工具/方法可以被借鉴？
2. 随着更先进的通用人工智能（AGI）出现，如何监管AI以保证结果可信？