Title: The Quantization Model of Neural Scaling

Speakers: Eric Michaud

Series: Machine Learning Initiative

Date: October 20, 2023 - 2:30 PM

URL: https://pirsa.org/23100100

Abstract: The performance of neural networks like large language models (LLMs) is governed by "scaling laws": the error of the network, averaged across the whole dataset, drops as a power law in the number of network parameters and the amount of data the network was trained on. While the mean error drops smoothly and predictably, scaled up LLMs seem to have qualitatively different (emergent) capabilities than smaller versions when one evaluates them at specific tasks. So how does scaling change what neural networks learn? We propose the "quantization model" of neural scaling, where smooth power laws in mean loss are understood as averaging over many small discrete jumps in network performance. Inspired by Max Planck's assumption in 1900 that energy is quantized, we make the assumption that the knowledge or skills that networks must learn are quantized, coming in discrete chunks which we call "quanta". In our model, neural networks can be understand as being implicitly a large number of modules, and scaling simply adds modules to the network. In this talk, I will discuss evidence for and against this hypothesis, its implications for interpretability and for further scaling, and how it fits in with a broader vision for a "science of deep learning".

Zoom link https://pitp.zoom.us/j/93886741739?pwd=NzJrcTBNS2xEUUhXajgyak94LzVvdz09

The Quantization Model of Neural Scaling

Eric J. Michaud

with **Ziming Liu**, **Uzay Girit**, and **Max Tegmark**

1

Background

Large Language Models (LLMs)
scaling laws
emergence

Review of what large language models (LLMs) do

Given some text...

The quick brown fox jumps over the lazy dog.

Split into "tokens"...

The	quick	brown	fox	jumps	over	the	lazy	dog	.

Which have numerical IDs...

```
[510, 3158, 8516, 30013, 27287, 689, 253, 22658, 4370, 15]
```

At each position in the sequence, the model outputs a probability distribution over the whole token vocabulary for the next token in the sequence

$$\text{loss} = \log \frac{1}{p_{\text{answer}}}$$

The average loss decreases smoothly and predictably

compute ∝ parameters × data

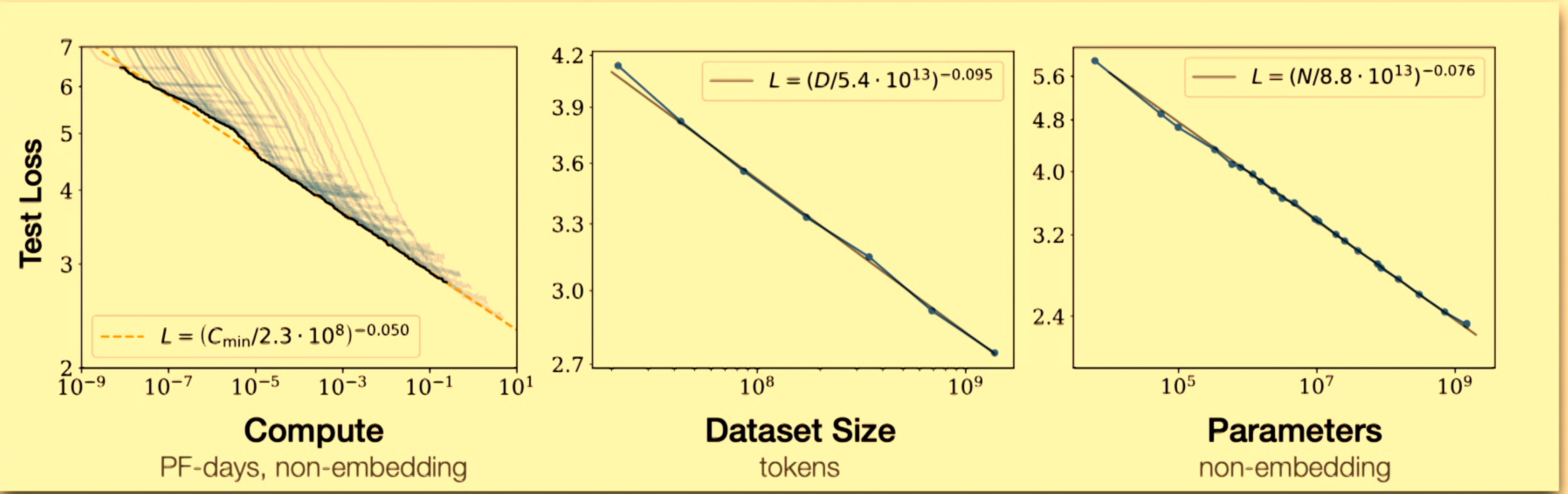


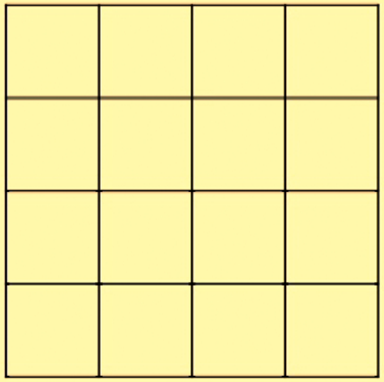
Figure 1 of Jared Kaplan, Sam McCandlish, et al. "Scaling Laws for Neural Language Models." *arXiv:2001.08361v1* (2020).

An existing model of neural scaling

Resolving a function on a manifold

Sharma and Kaplan, "Scaling Laws from the Data Manifold Dimension"

Approximating a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with a piecewise linear function has an error that drops off as a power law as the density of linear regions increases. The scaling exponent is $\alpha \leq 4/d$.



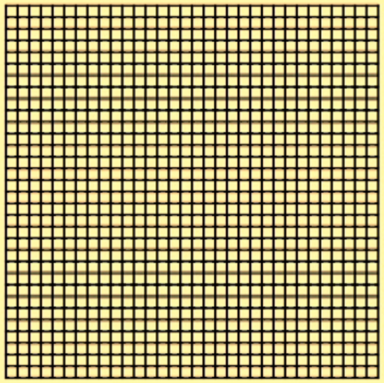
5

An existing model of neural scaling

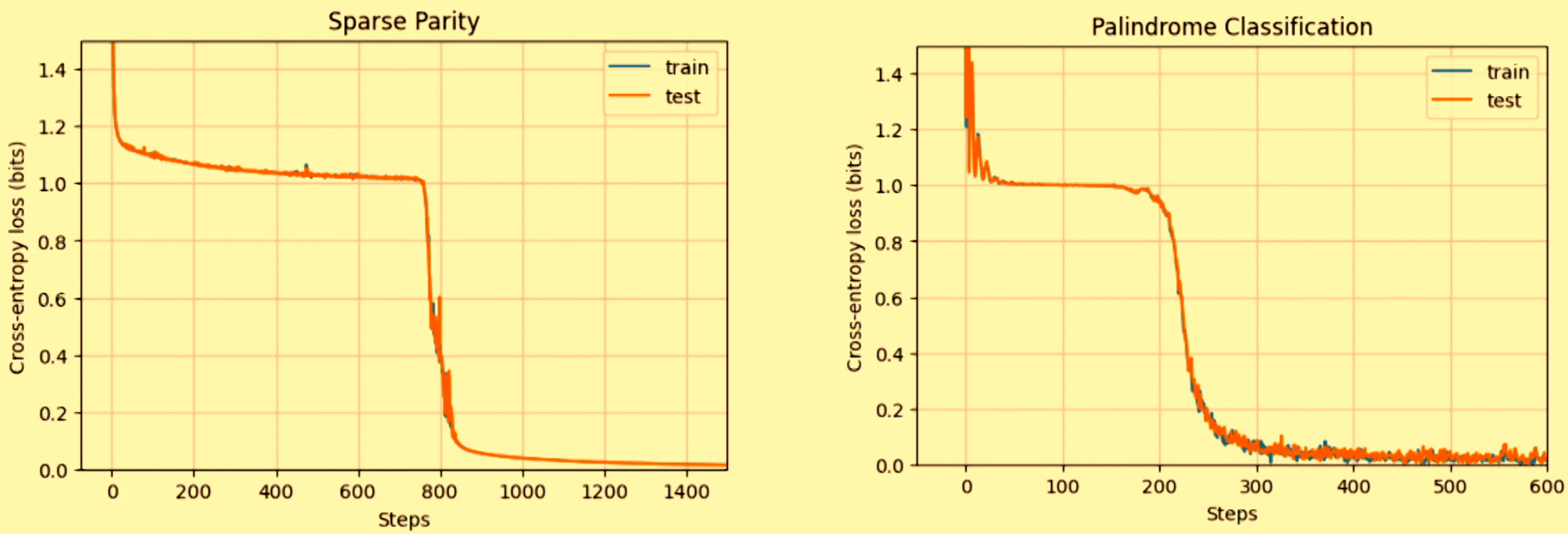
Resolving a function on a manifold

Sharma and Kaplan, "Scaling Laws from the Data Manifold Dimension"

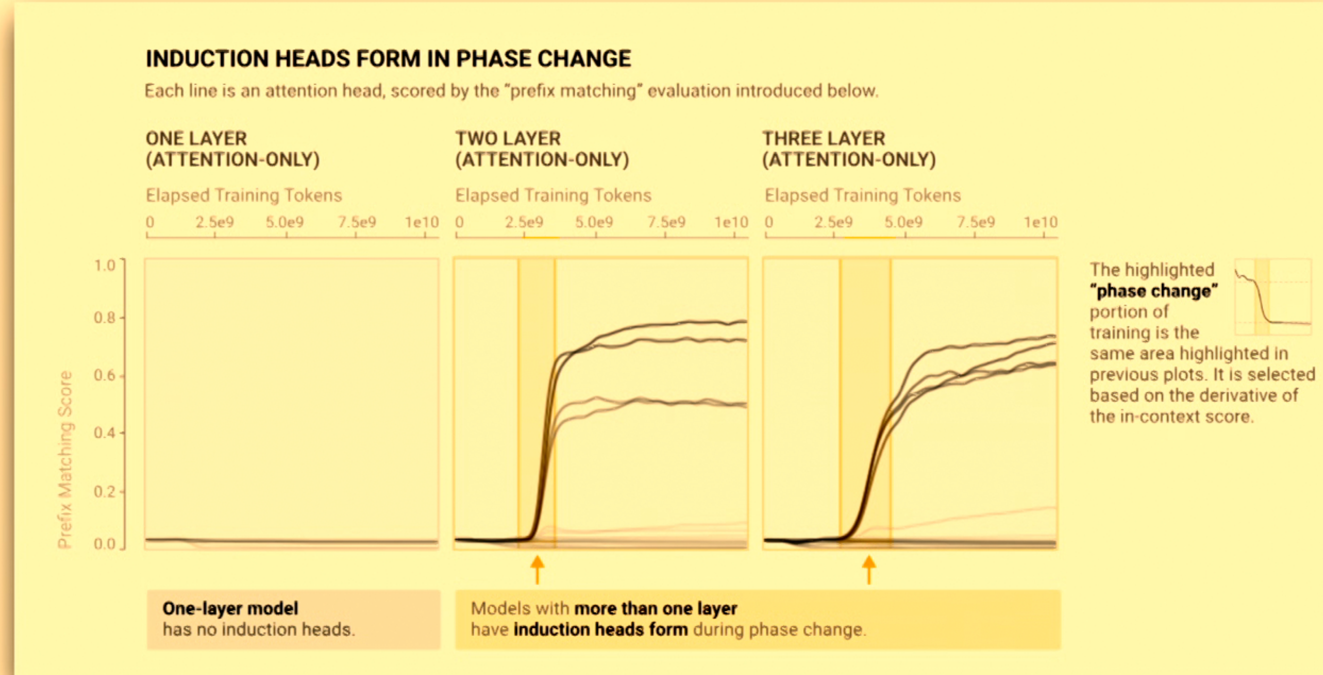
Approximating a function $f : \mathbb{R}^d \to \mathbb{R}$ with a piecewise linear function has an error that drops off as a power law as the density of linear regions increases. The scaling exponent is $\alpha \leq 4/d$.



But sometimes things seem more complicated



Discreteness during training: induction heads



From Olsson et al. "In-context Learning and Induction Heads", Transformer Circuits Thread, 2022

Particular capabilities seem to "emerge" in LLMs

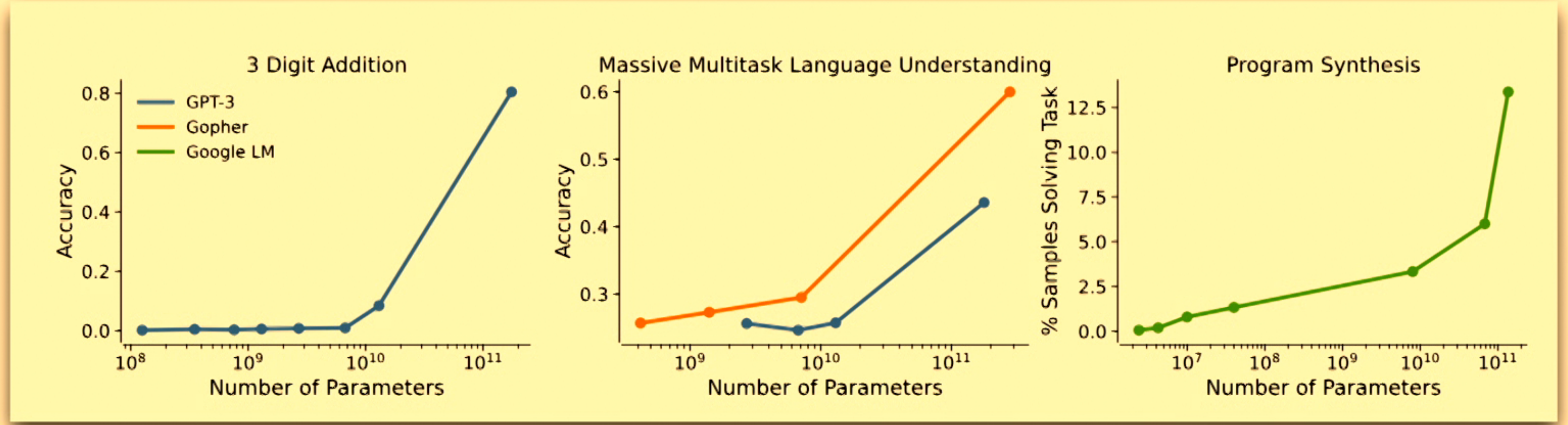


Figure 2 of Ganguli et al. "Predictability and Surprise in Large Generative Models"
2022 ACM Conference on Fairness, Accountability, and Transparency

"Emergence is when quantitative changes in a system result in qualitative changes in behavior"

-Wei (2022), Steinhardt (2022)

The average loss decreases smoothly and predictably

compute ∝ parameters × data

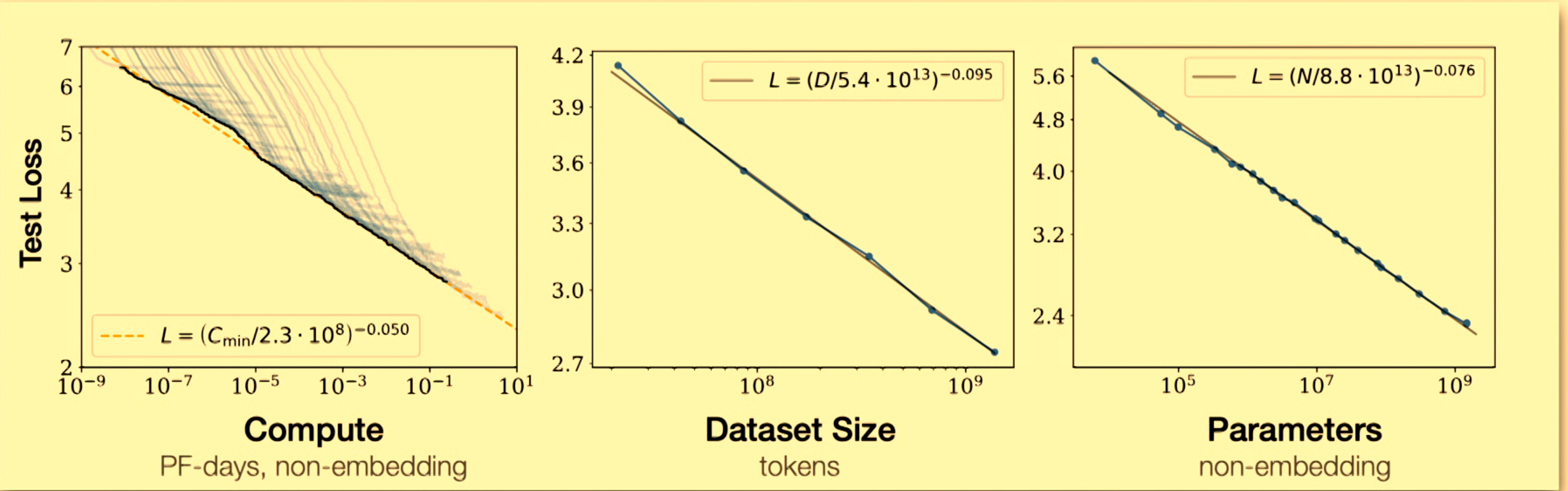


Figure 1 of Jared Kaplan, Sam McCandlish, et al. "Scaling Laws for Neural Language Models." *arXiv:2001.08361v1* (2020).

4

How does scaling change what neural networks learn?

Our key idea & result

Smooth scaling curves can average over many small discrete changes in model capabilities

Our key idea & result

Smooth scaling curves can average over many small discrete changes in model capabilities

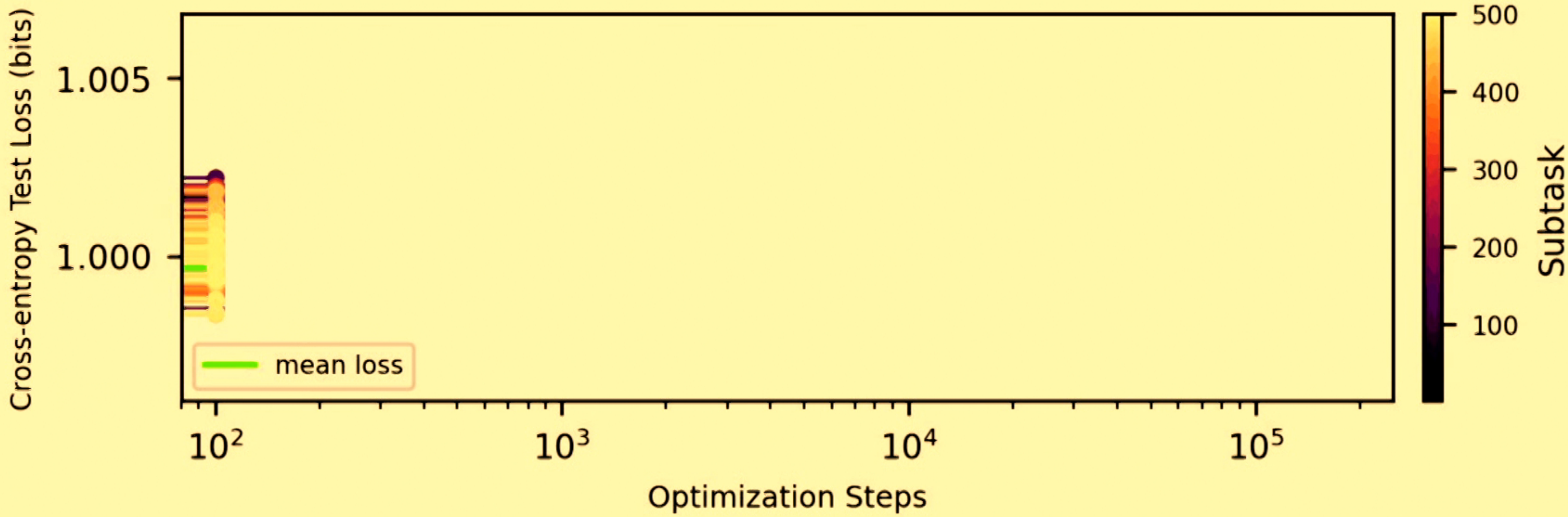
First: a demonstration on a toy dataset

Multitask sparse parity

A binary classification problem on binary strings

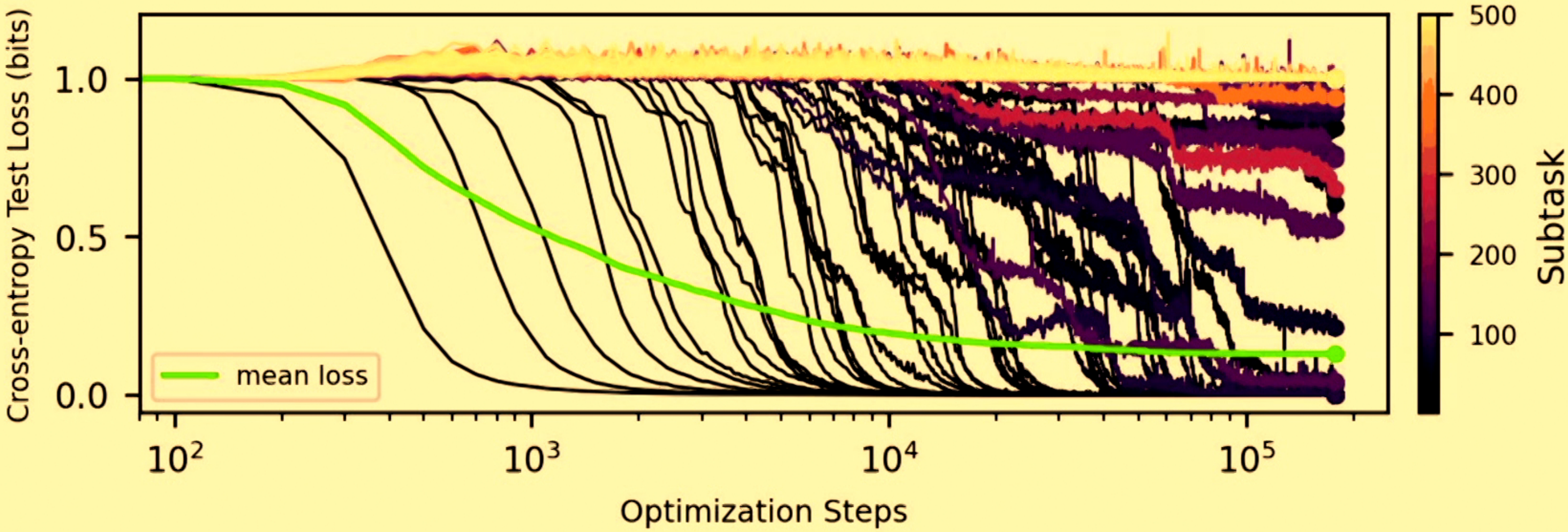
x		y
control bits	task bits	
00100000	010100011101000001010	0
01000000	000110101111001001001	1
10000000	111011101010010101100	0
01000000	110110001010101001110	1
10000000	101010101010010001010	0
00000010	100001101100011100011	0
00100000	111001011000101110101	1
00000100	010001110101100100010	0
00001000	000010101000111001000	1

Multitask sparse parity: training dynamics



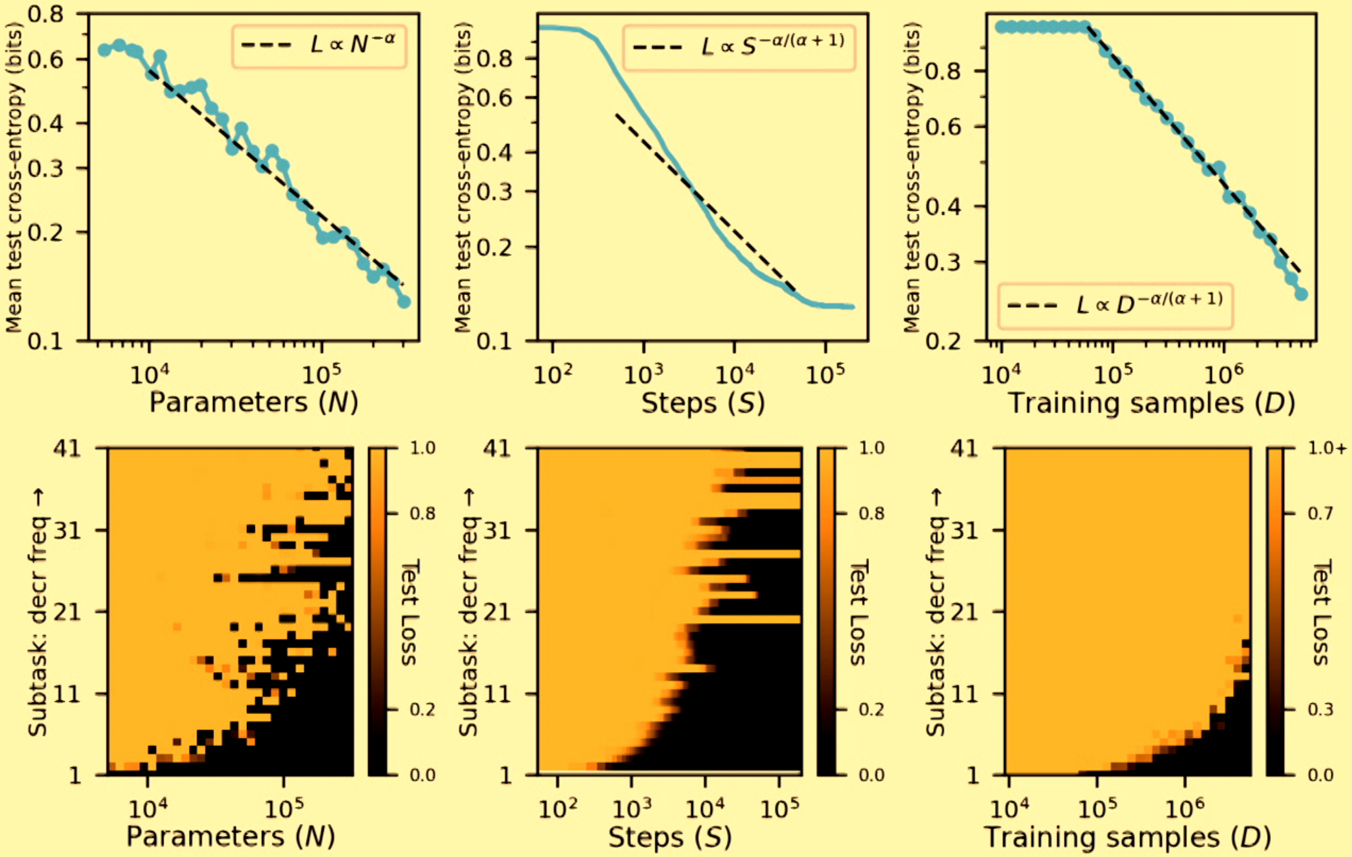
17

Multitask sparse parity: training dynamics



17

Multitask sparse parity: scaling



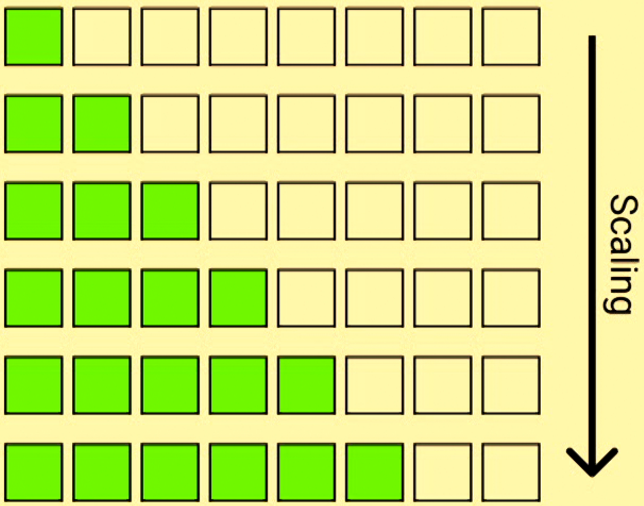
18

The Quantization Model

(1) There are a bunch of *things* which a neural network needs to learn to do prediction well. Let's assume these are *discrete* (either learned or not learned). We call these **"quanta"**.

(2) Some quanta are more useful for prediction (they lower the mean loss more) than others. We can order the quanta then into the **"Q Sequence"**. The effect of scaling is to learn more quanta in the Q Sequence.

(3) The **frequencies** at which the quanta are useful for prediction follow a **power law**.



The Quantization Model: inspiration

Max Planck's resolution of the ultraviolet catastrophe (1900)

- Energy quantized into discrete chunks
- Energy chunks called "quanta"

Quantization Model of Neural Scaling

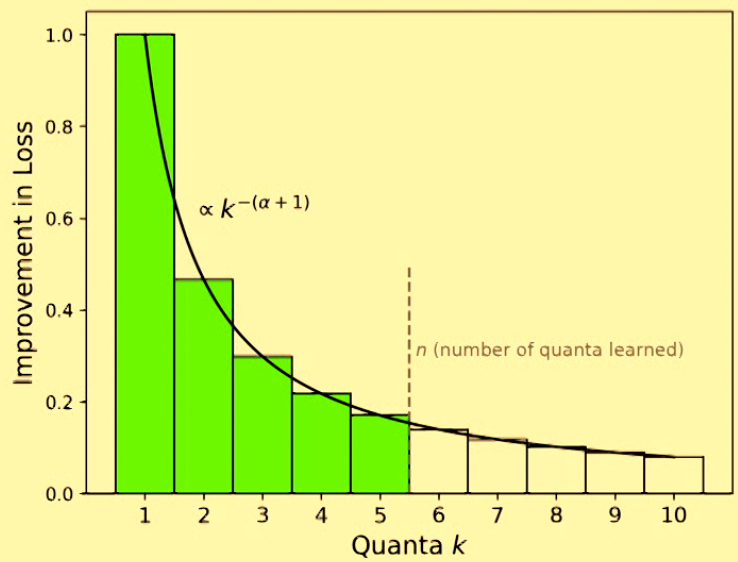
- Neural network knowledge/ capabilities quantized into discrete chunks
- Basic capabilities called "quanta"

How does this lead to power law scaling in mean loss?

How much does learning quantum k reduce the model's mean loss by?

If quantum k improves the model's performance on a fraction f_k of samples, and it lowers loss on those samples on average by some amount δ, then learning it reduces mean loss by δf_k.

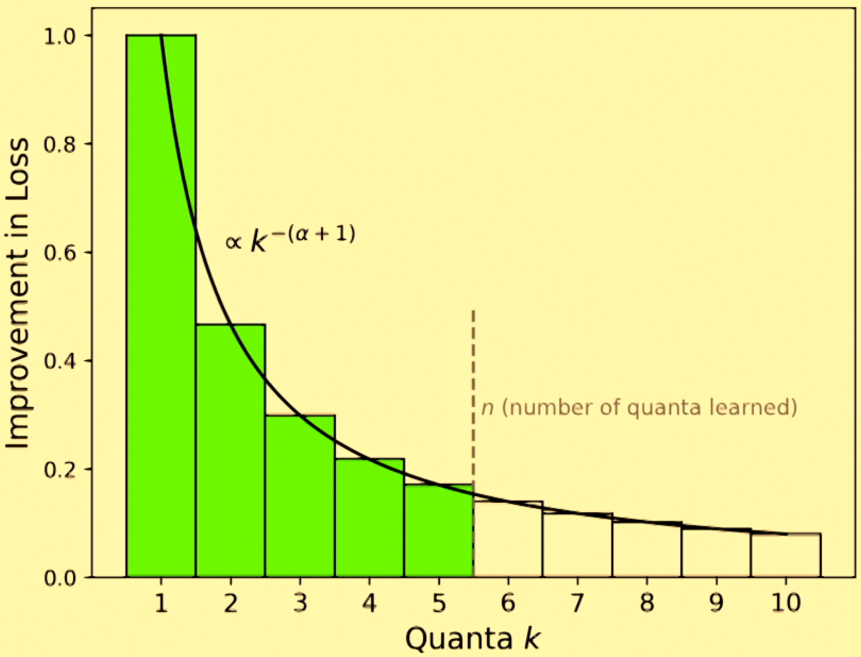
If $f_k \propto k^{-(\alpha+1)}$, then the reduction in mean loss from learning quantum k also follows a power law.



$$L(n) = L_0 - \delta \sum_{k=1}^{n} k^{-(\alpha+1)}$$

$$L(n) \approx C_0 + C_1 n^{-\alpha}$$

How does this lead to power law scaling in mean loss?



$$L(n) = L_0 - \delta \sum_{k=1}^{n} k^{-(\alpha+1)}$$

$$L(n) \approx L_0 - \delta \int_1^n k^{-(\alpha+1)} dk$$

$$L(n) \approx C_0 + C_1 n^{-\alpha}$$

22

Translating scaling in quanta n to scaling in parameters N

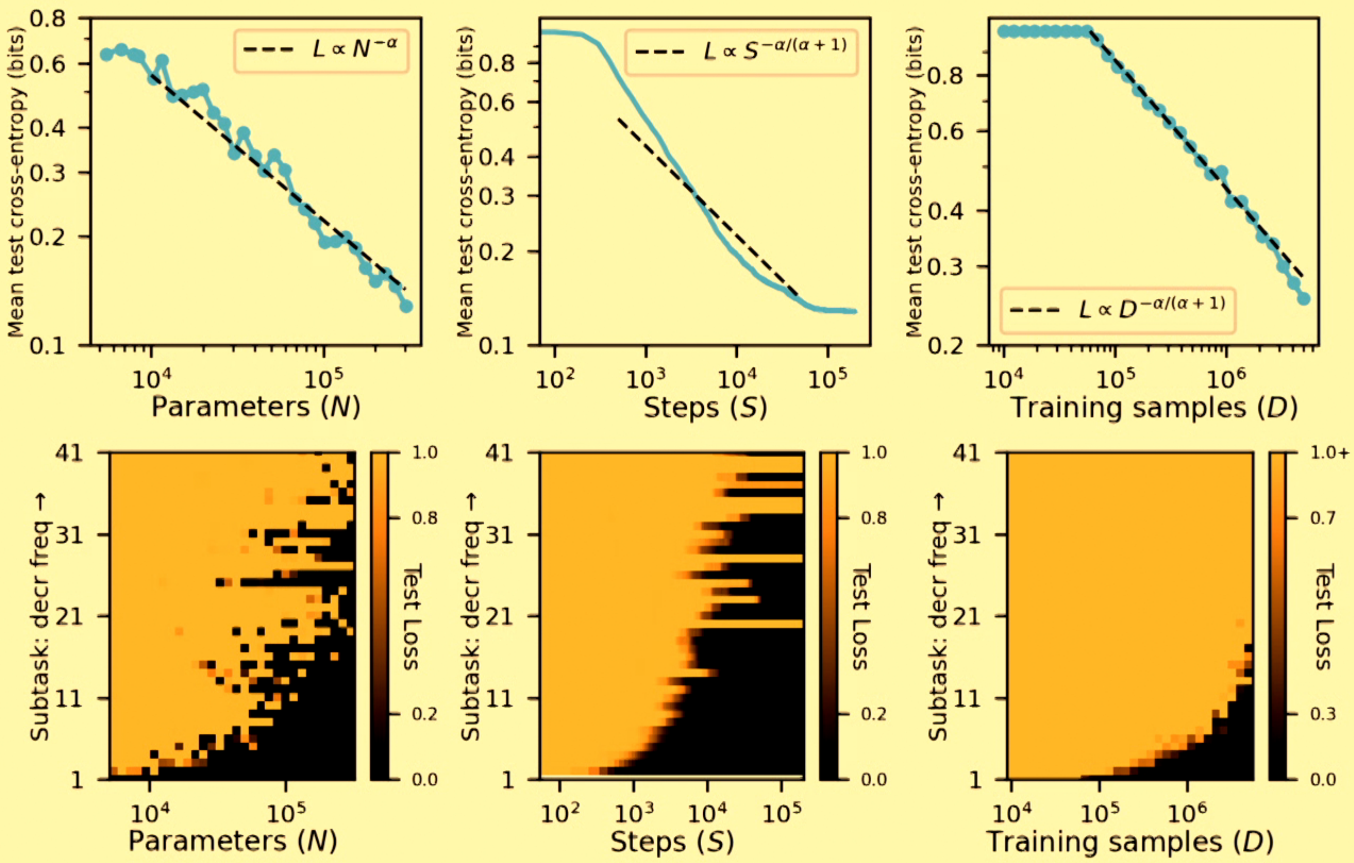
Quanta (n) scaling: $L(n) \approx C_0 + C_1 n^{-\alpha}$

Parameter (N) scaling: Assume we are bottlenecked not by data or training time, but just by model capacity. If quanta on average take up a constant number of network parameters, then $n \propto N$, and so:

$$L(N) \approx C_0 + C_2 N^{-\alpha}$$

23

Multitask sparse parity: scaling



18

Translating scaling in quanta n to scaling in data samples D

Quanta (n) scaling: $L(n) \approx C_0 + C_1 n^{-\alpha}$

Data (D) scaling: Assume that on average a constant threshold τ of examples involving a quantum are needed for the network to learn that quantum. In a given training dataset with D samples, the number of samples involving quantum k is $\propto Dk^{-(\alpha+1)}$. One can work out that $n \propto D^{1/(\alpha+1)}$

$$L(D) \approx C_0 + C_3 D^{-\alpha/(\alpha+1)}$$

Scaling laws from the Quantization Model

Power law over quanta: $f_k \propto k^{-(\alpha+1)}$

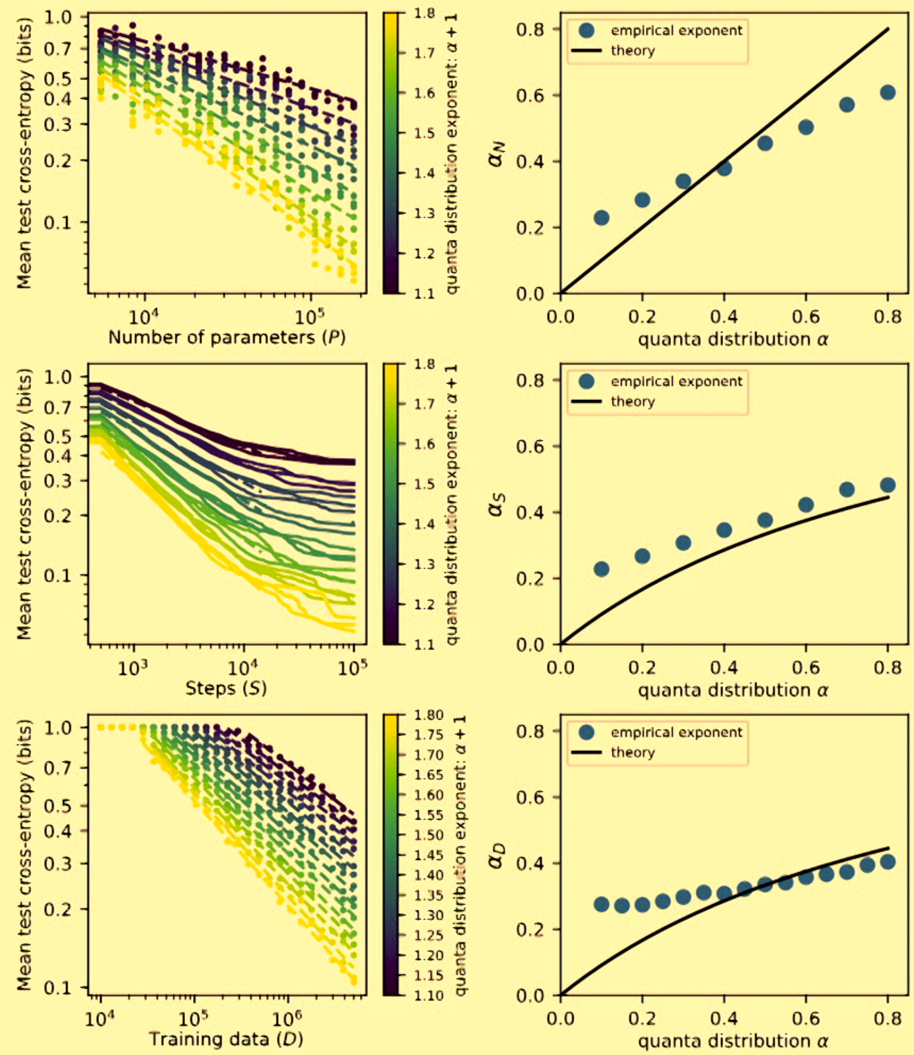
Quanta (n) scaling: $L(n) \approx C_0 + C_1 n^{-\alpha}$

Parameter (N) scaling: $L(N) \approx C_0 + C_2 N^{-\alpha}$

Data (D) scaling: $L(D) \approx C_0 + C_3 D^{-\alpha/(\alpha+1)}$

Training steps (S) scaling: $L(S) \approx C_0 + C_4 S^{-\alpha/(\alpha+1)}$

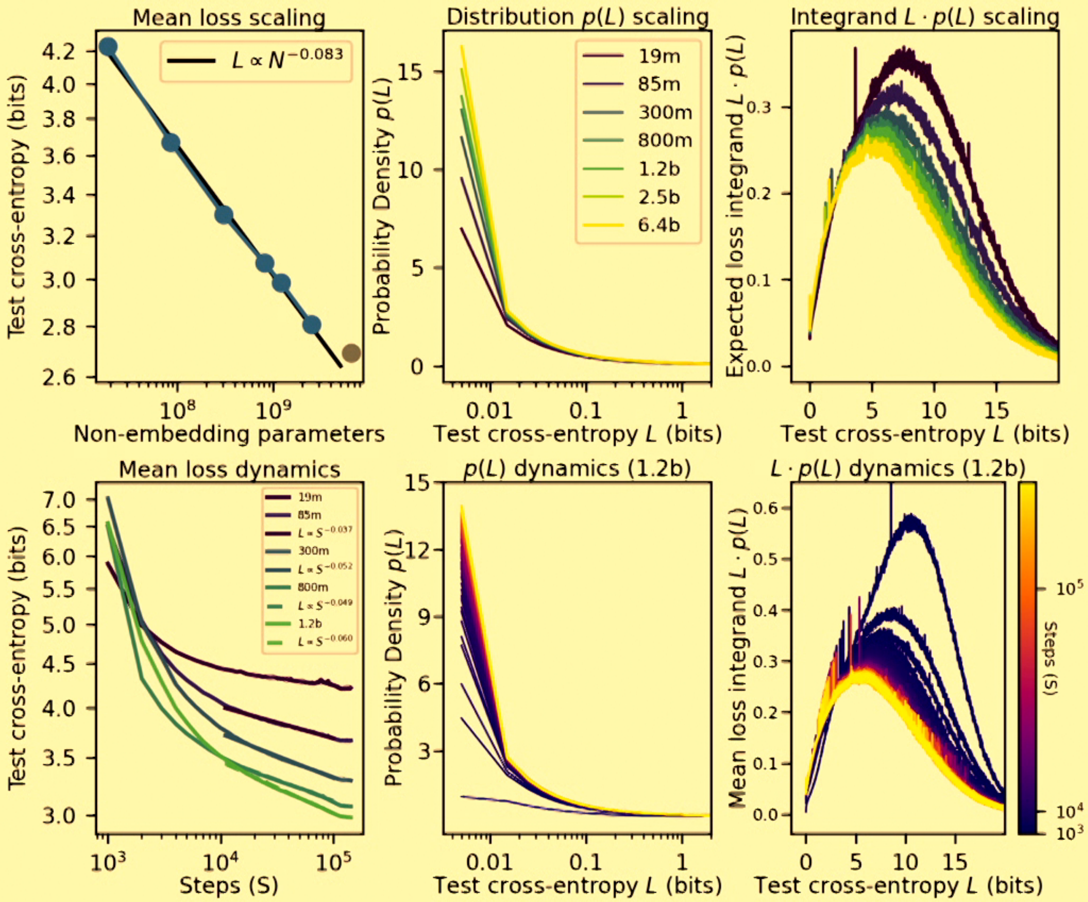
Caveat: relationship between empirical scaling exponents and subtask distribution power law exponent not exactly what theory says



27

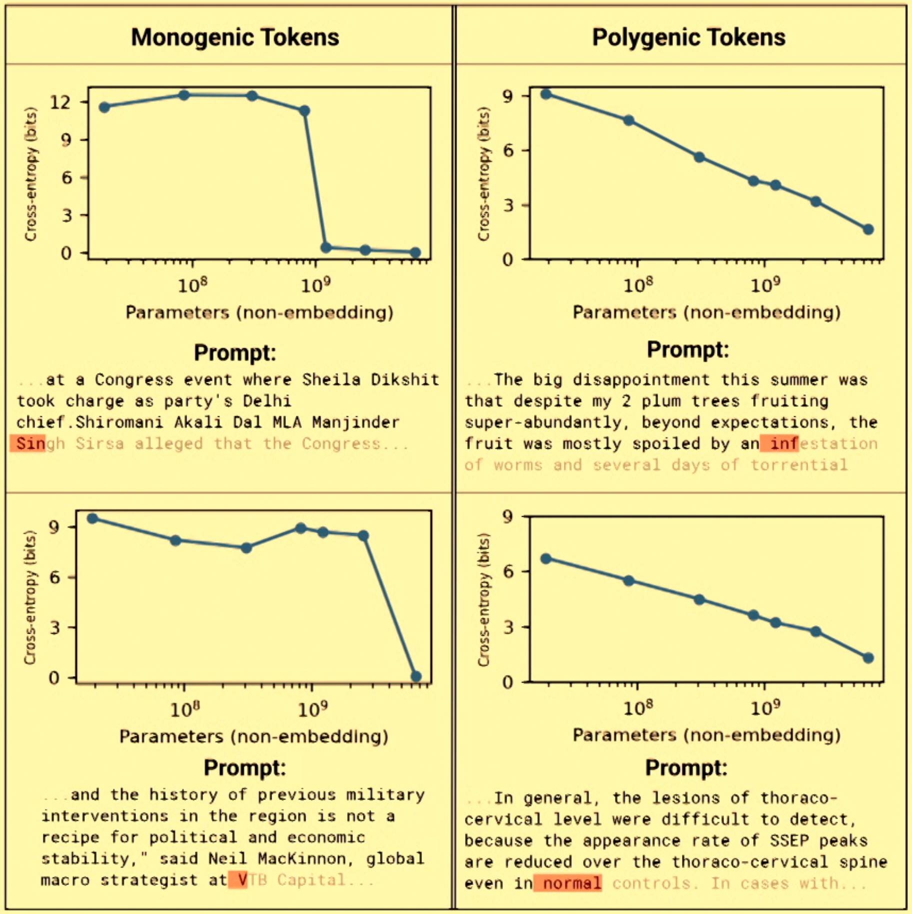
So for data with the right structure, our story of scaling roughly holds.

Statistics of LLM scaling



30

Diverse scaling curves on individual samples

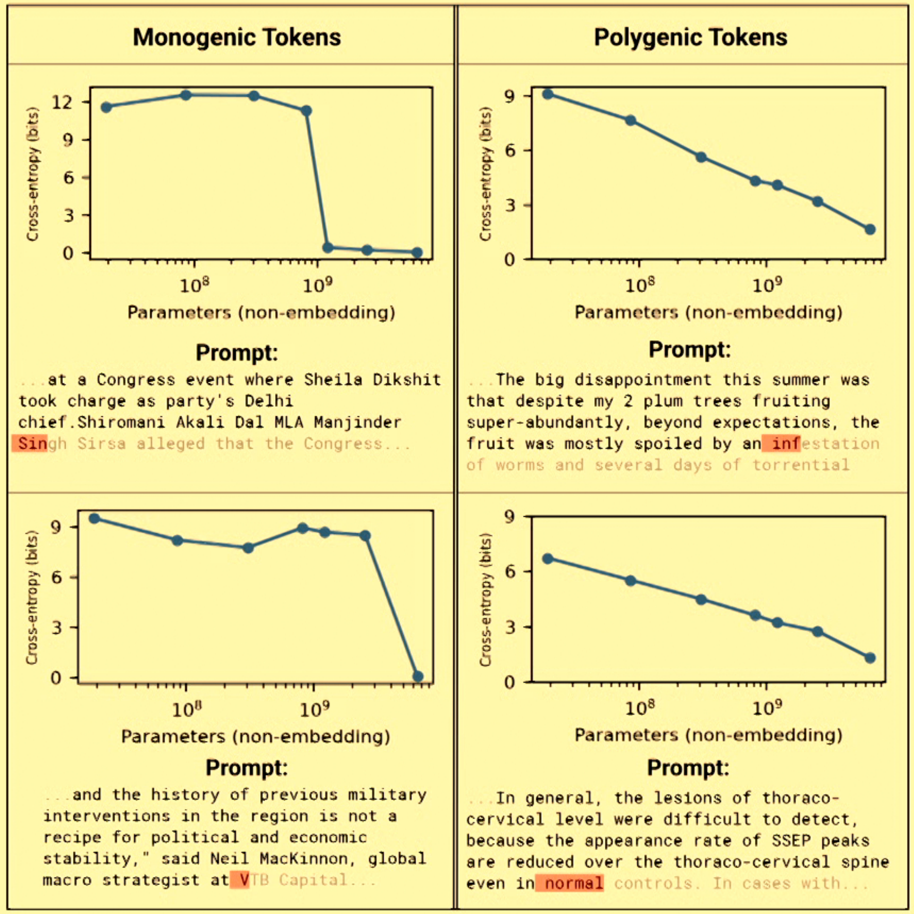


31

Borrowing genetics terminology: Monogenic vs Polygenic

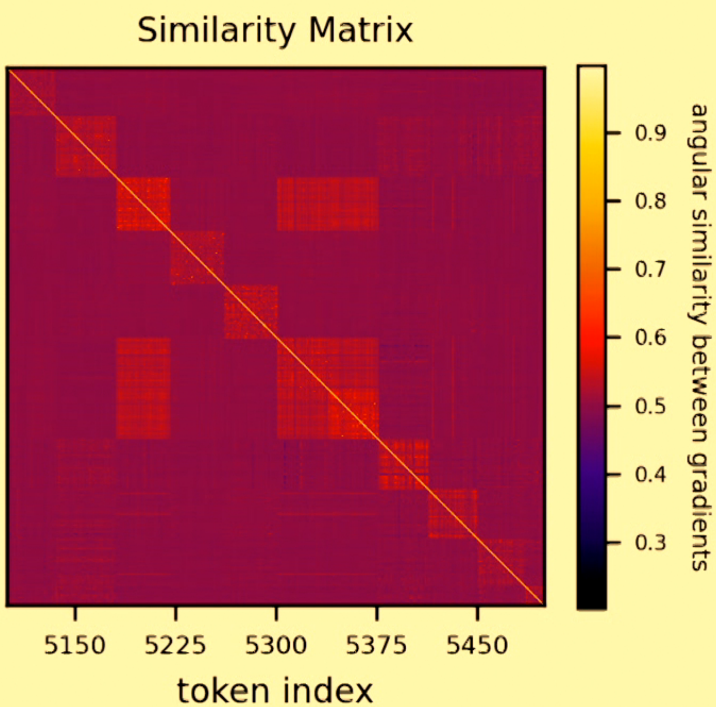
Monogenic: prediction benefits from a single quantum

Polygenic: prediction benefits from multiple quanta



32

Discovering quanta in language modeling



Cluster samples according to their gradients

34

Examples of clusters

"Quanta" of LLM capabilities auto-discovered in natural text

quantum for numerical sequence continuation (examples from cluster 50)

```
...ents his famous tonadas, a genre of the Venezuelan plains folk music.

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
```

```
...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
```

```
...
  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
```

```
...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
```

```
        ...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
```

```
...
    GetPrepareVoteMsg        = 0x07
    PrepareVotesMsg          = 0x08
    GetQCBlockListMsg        = 0x09
    QCBlockListMsg           = 0x0a
    GetLatestStatusMsg       = 0x0b
    LatestStatusMsg          = 0x0c
    PrepareBlockHashMsg      = 0x0d
    GetViewChangeMsg         = 0x0e
    PingMsg                  = 0x0f
```

quantum for predicting newlines to maintain text width (examples from cluster 100)

```
...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
```

```
...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
```

```
...    *
        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
```

```
...
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
```

```
<!--
/**
 * 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
```

```
...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
```

Examples from Cluster 146: comma after day of month

```
After his tweet went viral Aslan apologized on Twitter saying "it's not
like me" to use profanity.

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,
```

```
Sam Willard

Samuel Steven Willard (born September 9,
```

```
215 U.S. 437 (1910)
MECHANICAL APPLIANCE COMPANY
v.
CASTLEMAN.
No. 48.
Supreme Court of United States.
Argued December 3, 1909.
Decided January 3,
```

```
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,
```

Examples from Cluster 278: colon after CSS property

```
.rickshaw_graph.detail {
 pointer-events: none;
 position: absolute;
 top: 0;
 z-index: 2;
 background: rgba(0, 0, 0, 0.1);
 bottom: 0;
 width:
```

```
@import '../../../assets/sass/spin';

.app-header {
  background-color: #282c34;
  min-height: 100vh;
  display:
```

```
...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:
```

Examples from Cluster 269: "s" after start year of decade

```
Romford Ice Arena

Romford Ice Arena was an ice rink located in Romford in the London Borough
of Havering, England. The venue was built in the 1980s
```

```
...ownloadable formats: PDF

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
```

```
...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.

University teaching
During the early 1960s
```

```
...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
```

Examples from Cluster 292: "://" after "http"

```
                                                    ...##################
# TeslaCrypt Ransomware Payment Sites domain blocklist (TC_PS_DOMBL)     #
#                                                                        #
#                                                                        #
# For questions please refer to:                                         #
# https://
```

```
...to that document rather than overwrite it.
If it does not exist, it should insert the new document to the collection.

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://
```

```
Gruber, Martin A. Views of the National Zoological Park in Washington, DC,
showing Exhibit. 1919. Retrieved from the Digital Public Library of
America, http://
```

```
... it be discontinued? I heard Java Swing is discontinued and no more
future enhancements will be made. As a Beginner what should I learn.

A:

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://
```

Sequence continuation cluster

```
...ents his famous tonadas, a genre of the Venezuelan plains folk music.

Track listing
01- Mi Querencia (Simón Díaz)
02- Tonada De Luna Llena (Simón Diaz)
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
```

```
...
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
```

```
...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
```

```
...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
```

```
...
GetPrepareVoteMsg      = 0x07
PrepareVotesMsg        = 0x08
GetQCBlockListMsg      = 0x09
QCBlockListMsg         = 0x0a
GetLatestStatusMsg     = 0x0b
LatestStatusMsg        = 0x0c
PrepareBlockHashMsg    = 0x0d
GetViewChangeMsg       = 0x0e
PingMsg                = 0x0f
```

Newline prediction to maintain text width cluster

```
...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
```

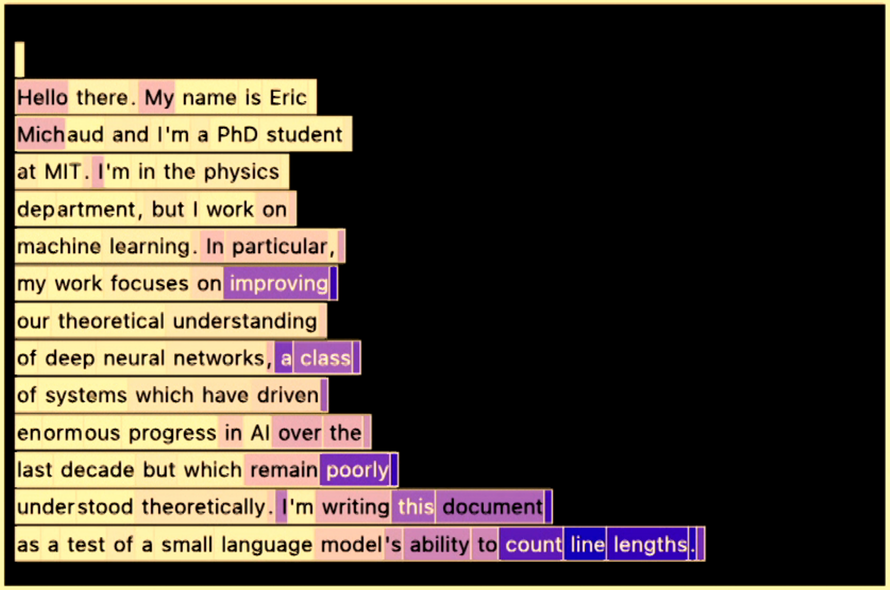
```
...
files (the
// "Software"), to deal in the Software without restriction, including
// without limitation the rights to use, copy, modify, merge, publish,
// distribute, sublicense, and/or sell copies of the Software, and to
permit
// persons to whom the Software is furnished to do so, subject to the\n
```

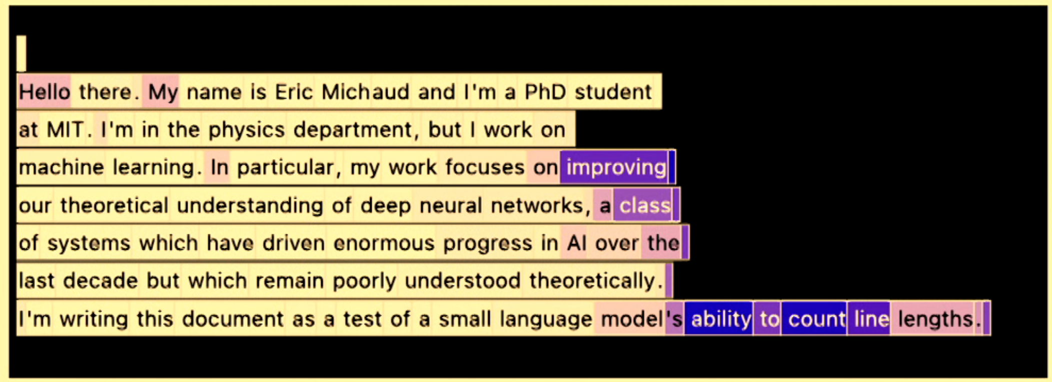
```
...ump is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# creddump is distributed in the hope that it will be useful,\n
```

```
...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[l] over their own envi-\n
```

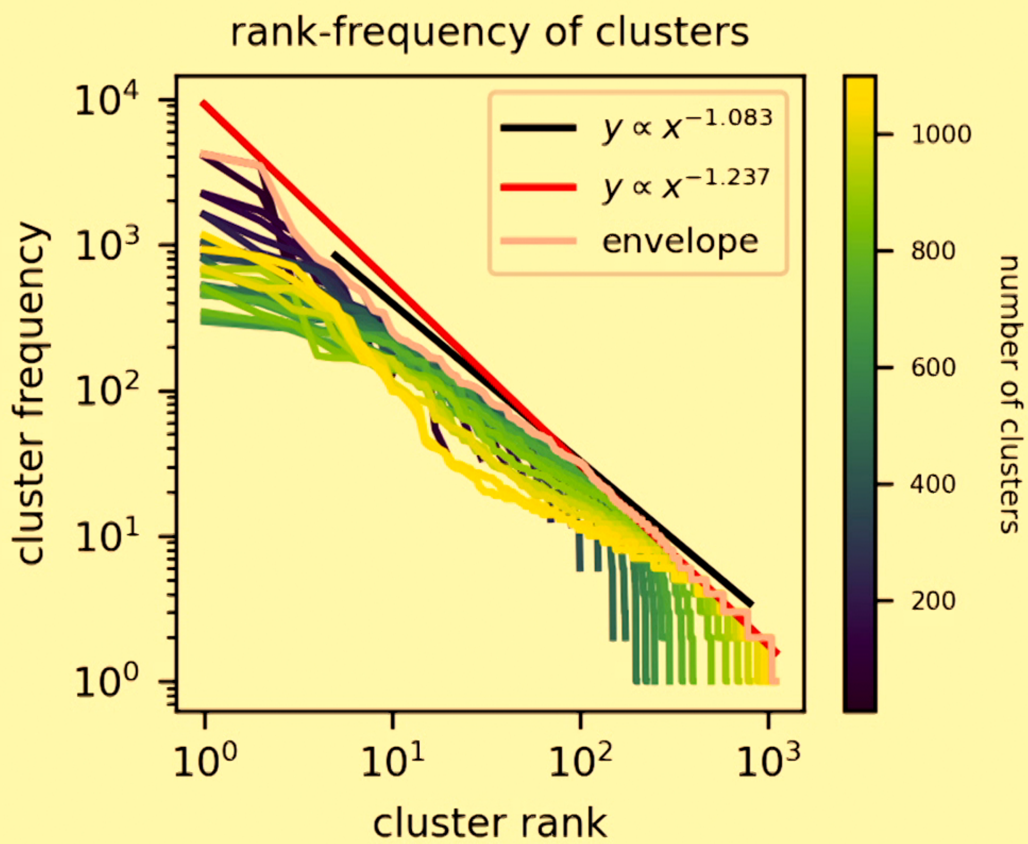
```
...   *
        Pursuant to 5TH CIR. R. 47.5, the court has determined
that this opinion should not be published and is not precedent
except under the limited circumstances set forth in 5TH CIR.\n
```

Newline prediction to maintain text width cluster





Rank-frequency of quanta clusters



Newline prediction to maintain text width cluster

```
...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
```

```
...
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
```

```
...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
```

```
...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[l] over their own envi-\n
```

```
...   *
       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
```

Microscopic

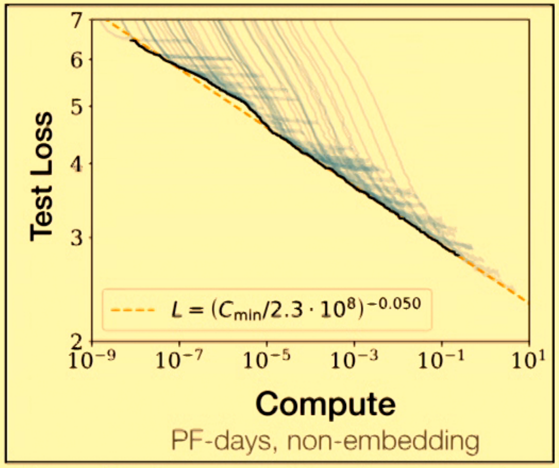
$$\theta_{t+1} = -\eta \nabla_{\theta} L$$

We understand the low-level training dynamics (we implement SGD ourselves) and have access to the full state of the network at all times.

Mesoscale

?

Macroscopic



from Kaplan et al. “Scaling Laws for Neural Language Models”

48