How can human scientists survive, in the time of AI?

Ziming Liu, MIT & IAIFI, June 2023
@Boston Physics Seminar

Disclaimer:
This talk may contain highly speculative opinions, please use at your own discretion. Opinions are my own.

Overview

AI - AI Capabilities and Limitations

- When are AI capable/incapable?

Humanity - Camps and survival strategies

Myself - My own survival strategies (my research)

AI Capabilities and Limitations

GPT-4

March 16, 2023

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

arXiv: 2303.08774

March 24, 2023

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

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

arXiv: 2303.12712

GPT-4

Write poems

Write LaTeX codes for plotting

GPT-4

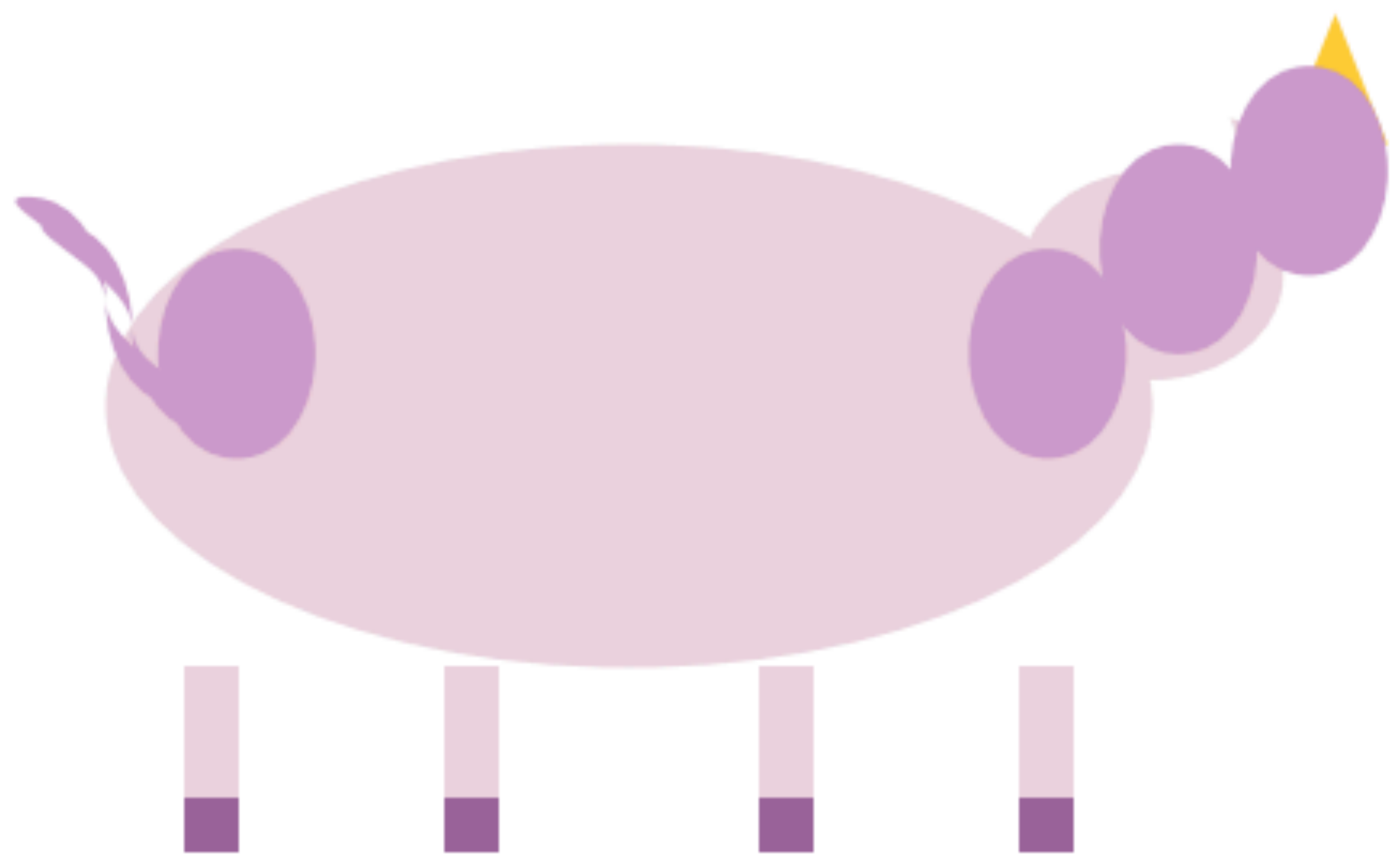
Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

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
```

Prompt: Draw a unicorn in TiKZ.

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



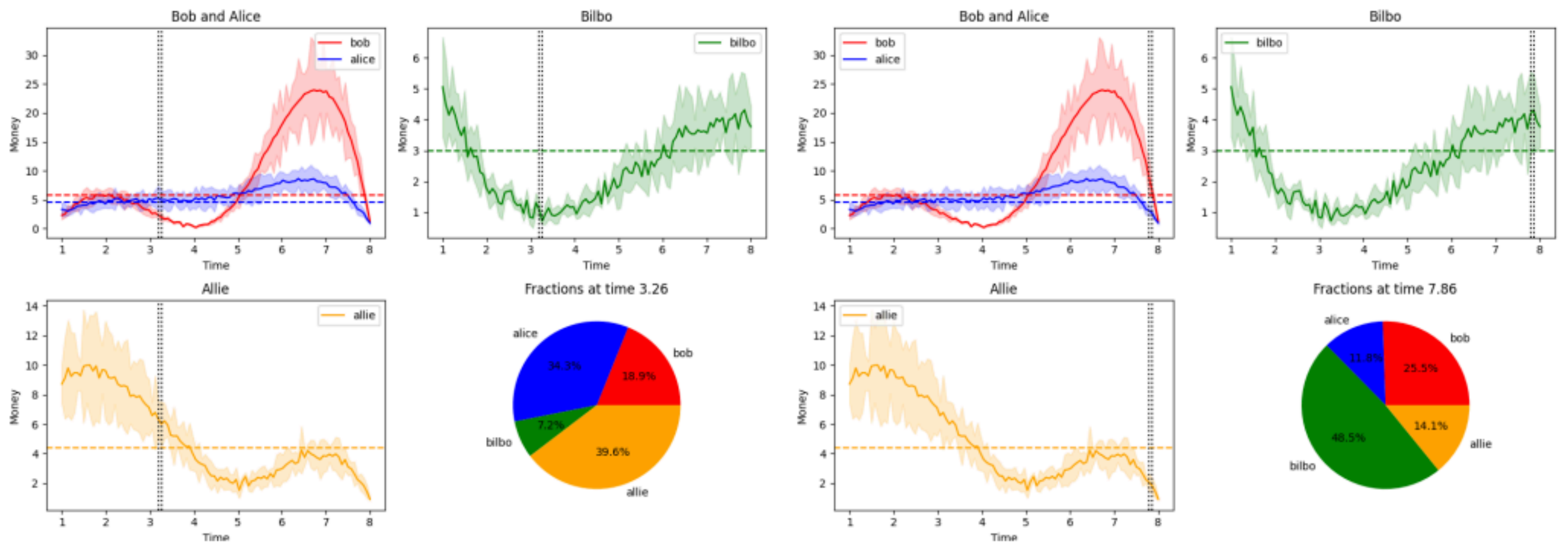
GPT-4

Write python codes for plotting

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ 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]



Do simple 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, $f(f(f(6))) = f(f(9)) = f(7) = 4$.

Solve LeetCode problems

GPT-4:

```
class Solution(object):
    def numberOfPaths(self, grid, k):
        """
        :type grid: List[List[int]]
        :type k: int
        :rtype: int
        """
        # Define MOD as 10**9 + 7
        MOD = 10**9 + 7
        # Get the dimensions of the grid
        m = len(grid)
        n = len(grid[0])
        # Initialize a 3D dp array with all zeros
        # dp[i][j][value] represents the number of paths where the sum of the elements on the path has a remainder of value when divided
        by k
        dp = [[[0] * k for _ in range(n)] for _ in range(m)]
        # Base case: the first cell has only one path and its value is grid[0][0] % k
        dp[0][0][grid[0][0] % k] = 1
        # Loop through the grid
        for i in range(m):
            for j in range(n):
                # Skip the first cell
                if i == 0 and j == 0:
                    continue
                # Loop through all possible remainders
                for value in range(k):
                    # The number of paths to the current cell with the current remainder is equal to the sum of:
                    # - The number of paths to the top cell with the previous remainder
                    # - The number of paths to the left cell with the previous remainder
                    # The previous remainder is equal to the current remainder minus the current value, plus k if it is negative
                    prev = (value - grid[i][j]) % k
```

GPT-4

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
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:c2de f2ed | c2de f2f2 | g2fe d2cB | c4-c2z2 :|
```

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
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|
```

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
V:1
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|
V:2 clef=bass
|:C,2C,2 G,2G,2 | C,2C,2 G,2G,2 | F,2F,2 E,2E,2 | C,4-C,2z2 :|
|:F,2F,2 C,2C,2 | F,2F,2 C,2C,2 | G,2G,2 F,2F,2 | E,2C,2 G,2C,2 :|
```

The sheet music of the final tune:

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.

GPT-4 mathematical abilities

4 Mathematical abilities

In this section we begin to assess how well GPT-4 can express mathematical concepts, solve mathematical problems and apply quantitative reasoning when facing problems that require mathematical thinking and model-building. We demonstrate that GPT-4 represents a jump in that arena too with respect to previous LLMs, even when compared to specially fine-tuned for math models such a Minerva. As it seems, however, GPT-4 is still quite far from the level of experts, and does not have the capacity required to conduct mathematical research.

The reader is called into caution that interpreting the results in this section correctly is a difficult exercise. As we will see, GPT-4 can answer difficult (indeed, competitive) high-school level math questions, and can sometimes engage in meaningful conversation around advanced math topics. Yet, it can also make very basic mistakes and occasionally produce incoherent output which may be interpreted as a lack of *true understanding*. Its mathematical knowledge and abilities can depend on the context in a seemingly arbitrary way.

GPT-4 mathematical abilities

Prompt:

A rain barrel collects water off the roof of a house during three hours of heavy rainfall. The height of the water in the barrel increases at the rate of $r(t) = 4t^3e^{-1.5t}$ feet per hour, where t is the time in hours since the rain began. At time $t = 1$ hour, the height of the water is 0.75 foot. What is the height of the water in the barrel at time $t = 2$ hours?

GPT-4 uses a correct approach (integration by parts),
but produces a wrong final answer due to a calculation error

GPT-4:

If we can find a function F such that $F'(t) = r(t)$, then $F(2) - F(1)$ is the amount that the height of the water increases from $t = 1$ to $t = 2$.

To find such a function F, we can use the technique of integration by parts, which says that if u and v are differentiable functions, then

$$\int udv = uv - \int vdu$$

Applying this to $r(t) = 4t^3e^{-1.5t}$, we can choose $u = t^3$ and $dv = 4e^{-1.5t}dt$, so that $du = 3t^2dt$ and $v = -\frac{8}{3}e^{-1.5t}$. Then we have

$$\int 4t^3e^{-1.5t}dt = -\frac{8}{3}t^3e^{-1.5t} - \int -\frac{8}{3}e^{-1.5t}3t^2dt$$

Simplifying and applying integration by parts again, we can choose $u = t^2$ and $dv = -\frac{8}{3}e^{-1.5t}dt$, so that $du = 2tdtandv = \frac{16}{9}e^{-1.5t}$. Then we have

$$\int 4t^3e^{-1.5t}dt = -\frac{8}{3}t^3e^{-1.5t} + \frac{16}{9}t^2e^{-1.5t} - \int \frac{16}{9}e^{-1.5t}2tdt$$

...

GPT-4 mathematical abilities

We begin with a simplification of a question which appeared in the 2022 International Mathematics Olympiad (IMO).

GPT-4

Prompt:

Suppose g is a continuous smooth function such that for every $x > 0$, there is **one and only one** $y > 0$ such that $g(x) + g(y) \leq 2xy$. Can you prove that $g(x) = x^2$?

GPT-4:

There are a few possible ways to approach this problem, but one possible method is as follows:

- Assume, for the sake of contradiction, that $g(x)$ is not equal to x^2 for some $x > 0$.
- Then, either $g(x) < x^2$ or $g(x) > x^2$ for that x.
- If $g(x) < x^2$, then let $y = x$. Then $g(x) + g(y) = 2g(x) < 2x^2 = 2xy$.
- If $g(x) + g(y) < 2xy$, then $g(x) + g(y)$ is strictly less than the upper bound of $2xy$ for positive x and y.
- This means that there is some room to increase y slightly and still satisfy the inequality $g(x) + g(y) \leq 2xy$.
- For example, let $y' = y + \epsilon$, where ϵ is a small positive number such that $g(x) + g(y') < 2xy$. This is possible because g is continuous and smooth, so it does not have any jumps or kinks that would prevent such a choice of epsilon.
- Then, y' is another positive number such that $g(x) + g(y') \leq 2xy'$, which contradicts the assumption that y is the one and only one positive number that satisfies this condition given x.
- Therefore, if $g(x) < x^2$, then $g(x) + g(y)$ cannot be strictly less than $2xy$, and it must be equal to $2xy$.
- If $g(x) > x^2$, then let y^* be the unique positive solution of $g(x) + g(y) = 2xy$. Then $g(y^*) < (y^*)^2$, since otherwise we would have $g(x) + g(y^*) > 2xy$, which contradicts the existence of y^* given x.
- But then, by the same argument as before, we can let $x = y = y^*$ and obtain a contradiction, since $g(y) + g(y) = 2g(y) < 2y^2 = 2yy$, which contradicts the uniqueness of y given y.
- Therefore, $g(x)$ must be equal to x^2 for all $x > 0$, as required.

Solving it requires a more creative approach, as there is no clear strategy for beginning the proof. For example, the decision to split the argument into two cases (g(x) > x^2 and g(x) < x^2) is not an obvious one, nor is the choice of y ∗ (its reason only becomes clear later on in the argument). Furthermore, the solution demands knowledge of calculus at the undergraduate level. Nevertheless, GPT-4 manages to produce a correct proof.

GPT-4 mathematical abilities

GPT-4

Human: Can you remind me what a k-SAT problem is?

AI: A k-SAT problem is a type of logical problem that involves finding a way to assign truth values (true or false) to a set of variables that appear in a collection of clauses, each of which consists of a disjunction (logical OR) of exactly k literals. A literal is either a variable or its negation. For example, (x_1 OR NOT x_2 OR x_3) is a clause with $k = 3$ and three literals.
The problem is to find an assignment of truth values to the variables that satisfies (makes true) all the clauses in the collection, or to determine that no such assignment exists. This is called the satis fiability problem, or SAT for short.

Human: What's the most obvious way (the first thing you would refer to when teaching a class) to solve a k-SAT problem?

AI: There are different ways to approach the k-SAT problem, but one of the most obvious and straightfor ward methods is to use brute force, or exhaustive search. This means trying all possible combinations of truth values for the variables, and checking if any of them satisfies all the clauses. For example, if there are n variables, there are 2^n possible combinations to try.

Human: Ok, great. Now, given the instance of a k-SAT problem suppose that I define a graph where each vertex refers to a variable in the instance, and two vertices are connected by an edge if and only if the two variables appear in the same clause. Suppose that for a certain instance, the graph I just described has two connected components of size $n/2$ each. In this case, is there a more efficient way to solve the k-SAT instance?

AI: Yes, there is a more efficient way to solve the k-SAT instance in this case. Since the graph has two connected components, this means that the clauses can be partitioned into two subsets, such that each subset only involves the variables in one component. Therefore, the k-SAT instance is equivalent to solving two independent k-SAT instances, each with $n/2$ variables. This reduces the number of possible combinations to try from 2^n to $2^{n/2} + 2^{n/2}$, which is much smaller for large n. In general, the more dis connected the graph is, the easier the k-SAT instance becomes, because it can be decomposed into smaller subproblems.

Human: Now suppose that the graph has only one connected component, but if you were to remove one single edge, then it would have two connected components. Can you suggest an algorithm in this case?

AI: One possible algorithm in this case is to use a divide and conquer strategy. First, identify the edge that, if removed, would split the graph into two components. Then, assign a truth value to one of the variables that are connected by that edge, and solve the resulting k-SAT instance recursively. If the instance is satisfiable, then return the solution. If the instance is unsatisfiable, then assign the opposite truth value to the same variable, and solve the k-SAT instance recursively again. If the instance is still unsatisfiable, then return that the original k-SAT instance is unsatisfiable. This algorithm reduces the number of possible combinations to try from 2^n to $2^{n-1} + 2^{n-1}$, which is still better than brute force. However, this algorithm depends on finding the edge that splits the graph, which may not be easy in general.

Human: Can you generalize this principle to the case that the graph has q edges which, if removed, would split the graph into p connected components?

AI: The general principle is similar, but more complicated. The idea is to assign truth values to the variables that are connected by the q edges that split the graph, and solve the resulting k-SAT instance recursively for each possible assignment. If any of the assignments leads to a satisfiable k-SAT in stance, then return the solution. If none of the assignments leads to a satisfiable k-SAT instance, then return that the original k-SAT instance is unsatisfiable. This algorithm reduces the number of possible combinations to try from 2^n to $2^q \times 2^{(n-q)/p}$, which is better than brute force if q is small and p is large. However, this algorithm also depends on finding the q edges that split the graph, which may be hard in general.

Human: Note that the complexity that you wrote is only correct in the case that all connected components are of the same size. Can you find a more general formula?

AI: Yes, you are right. The more general formula is $2^q \times \sum_i 2^{n_i}$, where n_i is the number of variables in the i-th connected component, and the sum is over all p components. This is because each component has 2^{n_i} possible combinations to try, and the q edges that split the graph have 2^q possible assignments. This formula is equal to the previous one if all n_i are equal to $(n-q)/p$.

GPT-4 demonstrates understanding of the concepts from graph theory and algorithms. It is able to reason about an abstract graph construction, which relates to a constraint satisfaction problem, and deduce correct conclusions about the SAT problem (to the best of our knowledge, this construction does not appear in the mathematical literature). The conversation reflects profound understanding of the undergraduate-level mathematical concepts discussed, as well as a significant extent of creativity. Although GPT-4 makes a mistake in one instance (writing 2n–1 rather than 2n/2), the subsequent conversation demonstrates that this error does not reflect a lack of understanding. Rather, GPT-4 seems to have made an error in a manner similar to a human typo, as it later provides the correct generalization of the formula.

Where it could go wrong

Error type	Percentage of wrong solutions
Arithmetic mistakes (including counting)	68%
Misunderstanding the problem statement	10%
Wrong approaches	22%

Reverse reasoning and verification Mathematical content produced by humans often presents a conclusion before outlining the reasoning that led to it. For example, the sentences "we will show next that $x = 0$ is a solution ..." or "We are going to prove the claim: AC is perpendicular to BD" can appear in a math problem's solution. This stylistic choice can improve readability, but it poses a challenge for natural language generation models, as it requires the model to infer the answer before producing the reasoning steps. We observe that not only does GPT-4 adopt this style, but it has an additional related drawback: **even when it infers a clearly wrong answer at the start, it will still attempt to create justifications for it, instead of correcting it**. This may again be due to the style of the training data, which mostly contains straightforwards solutions rather than trial-and-error style discussions and it is not clear whether or not it can be mitigated through a reinforcement learning phase (such as GPT-Instruct).

When/Why are AI capable/incapable?

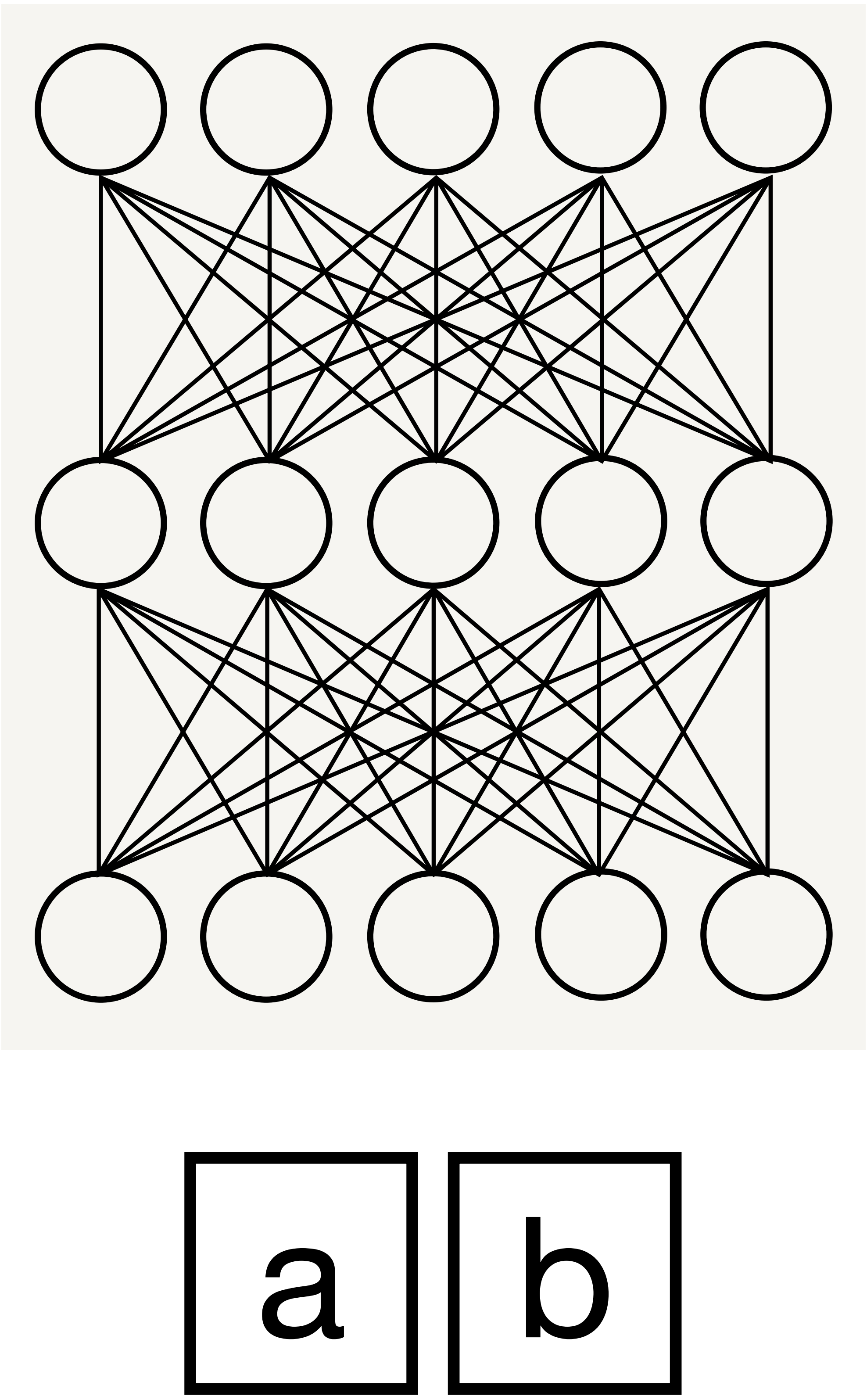
How do NN learn math (Grokking)

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

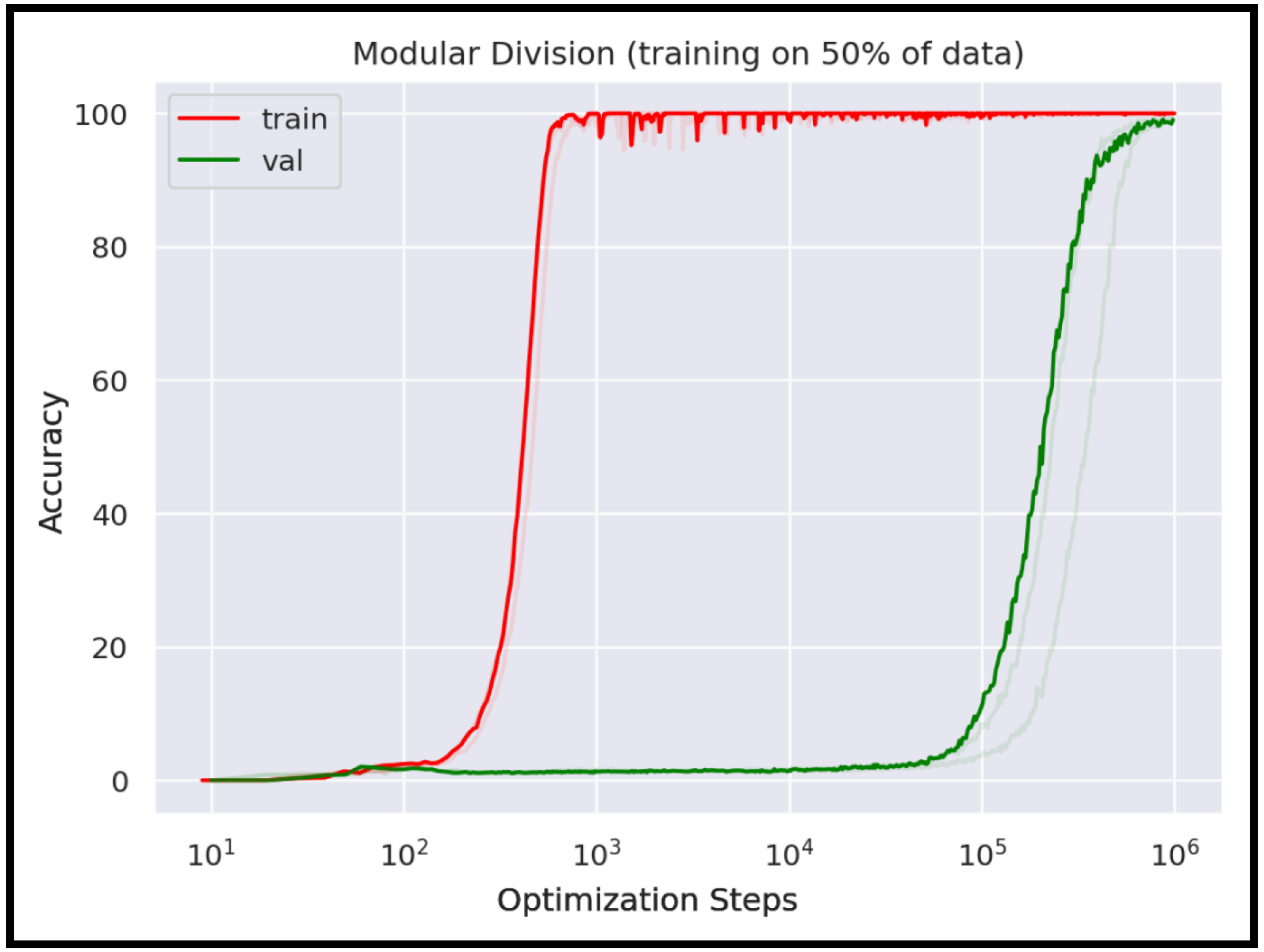
$$a \circ b = c$$

Logits for a, b, c, …



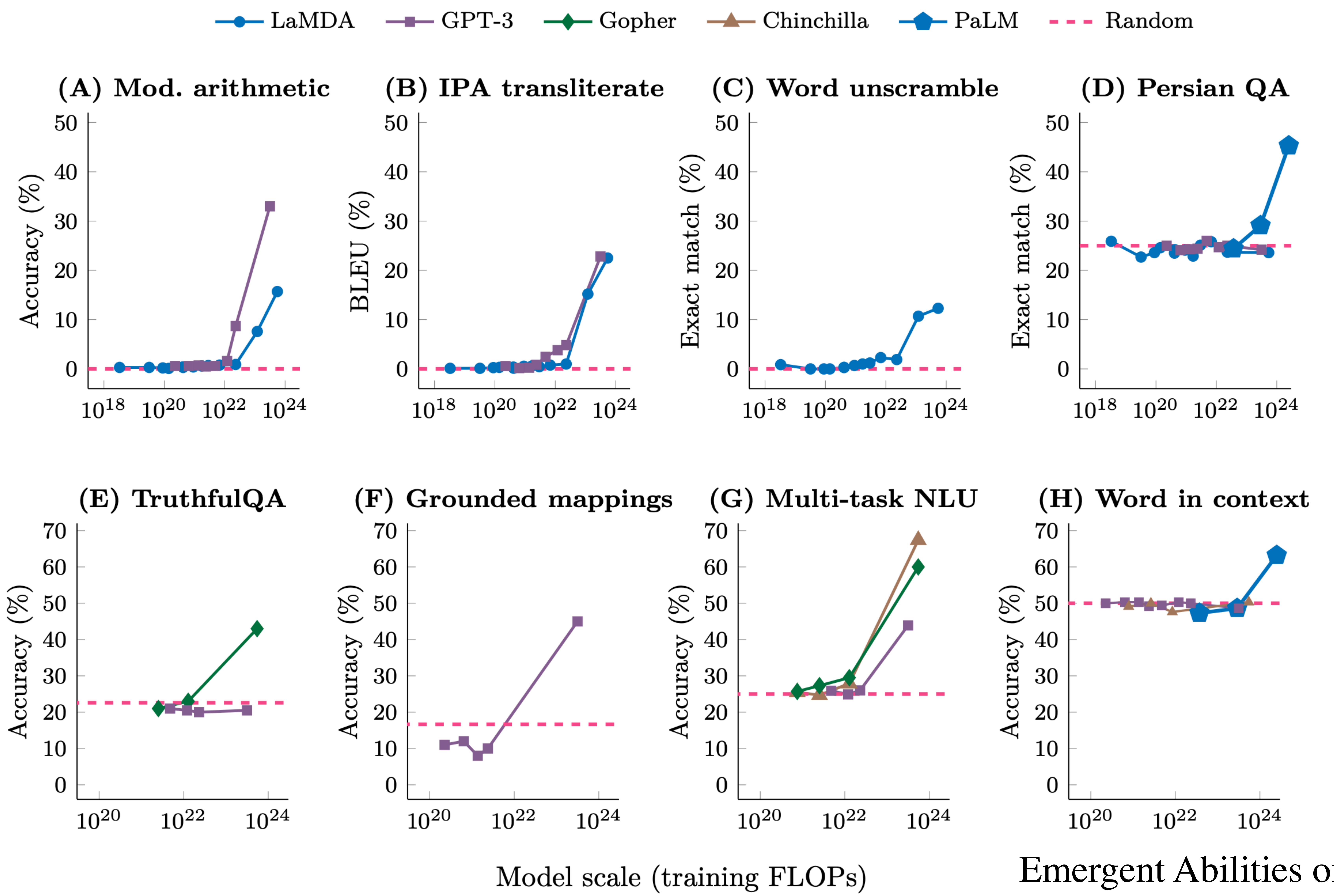
Trainable Embeddings

Phase transition behavior



Grokking: Generalisation beyond overfitting on small algorithmic datasets
2201.02177

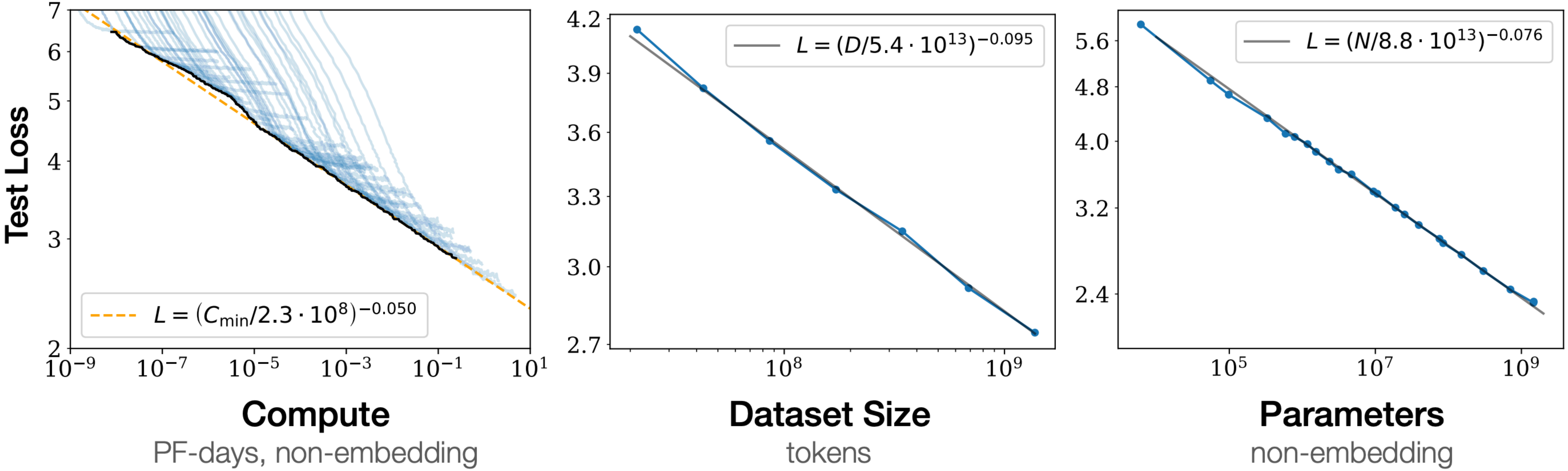
Emergent abilities (EA)



arXiv: 2206.07682

Emergent Abilities of Large Language Models

Neural Scaling Laws (NSL)



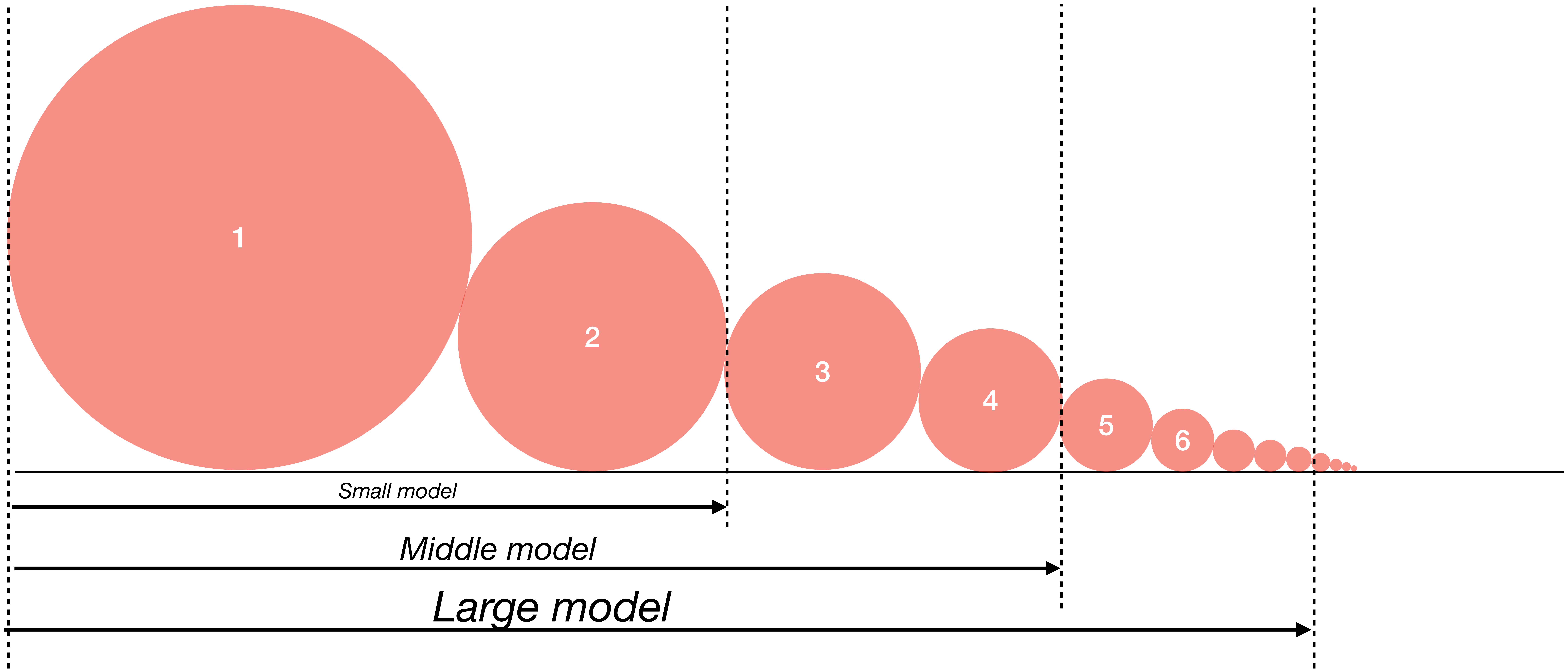
Scaling Laws for Neural Language Models, arXiv: 2001.08361

How EA leads to NSL

The quantisation model of neural scaling, arXiv: 2303.13506

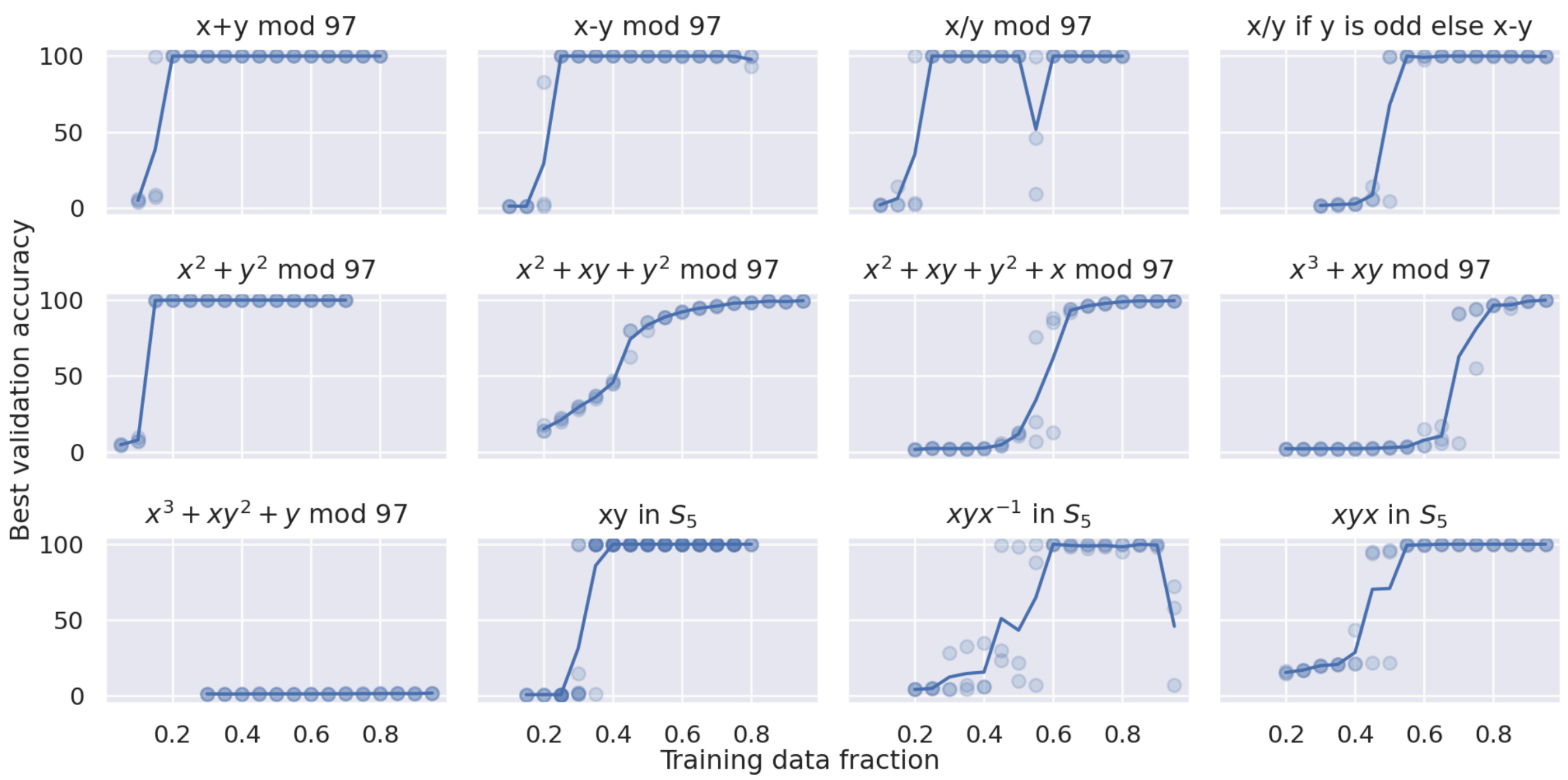
Knowledge quanta sequence

circle size = importance = frequency

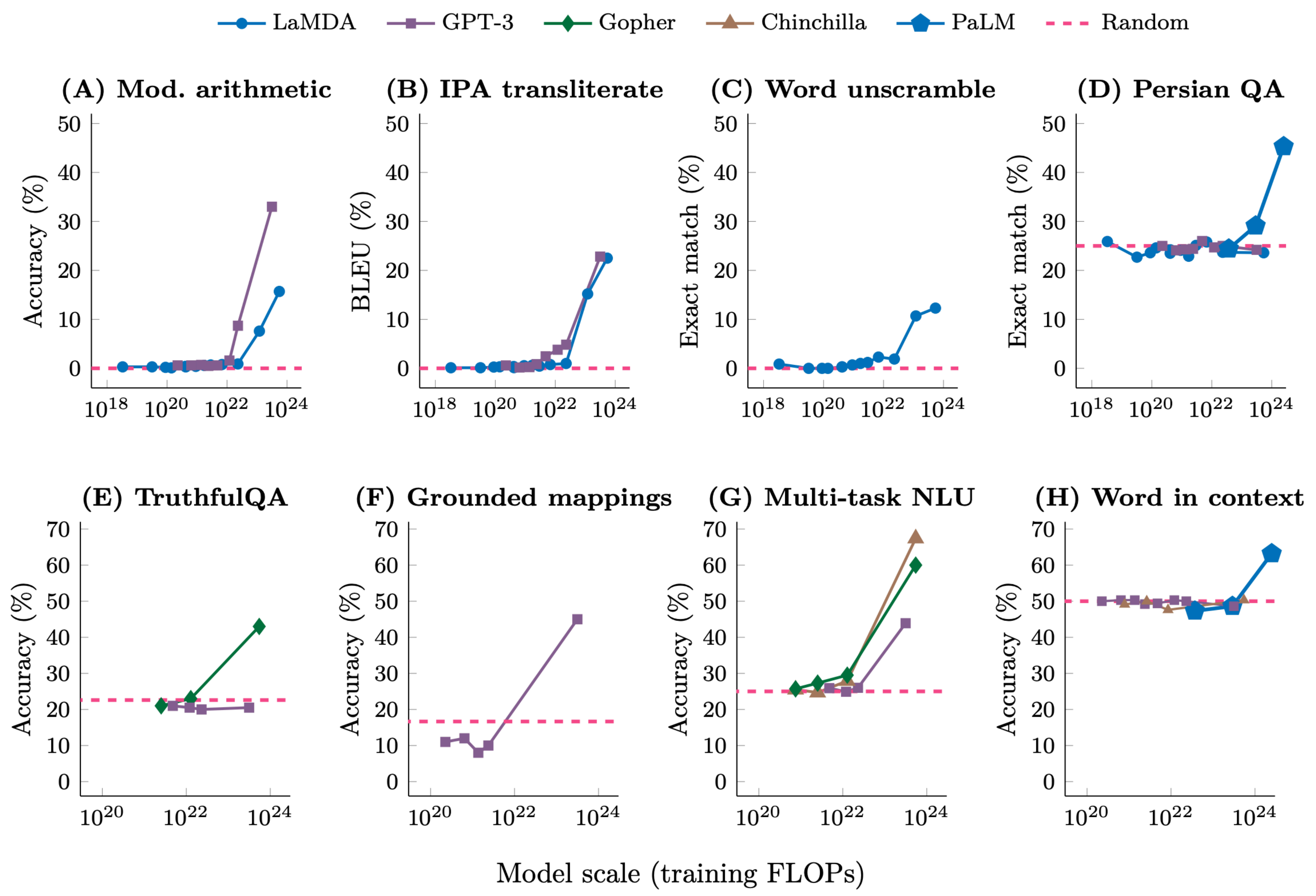


When are AI capable?

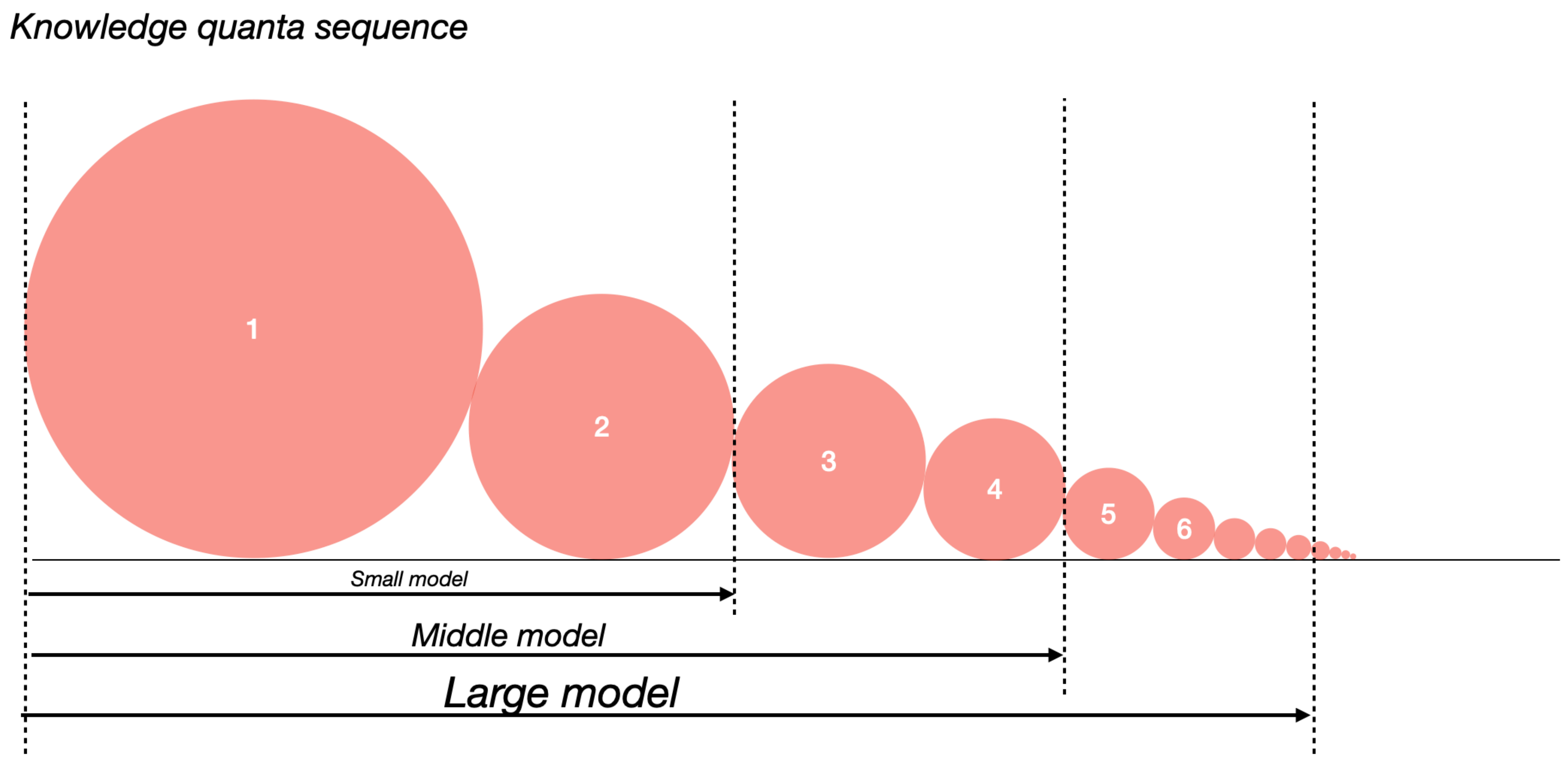
Enough data



Enough model size, enough compute



Pattern frequent enough



Grokking: Generalisation beyond overfitting on small algorithmic datasets
arXiv: 2201.02177

Emergent Abilities of Large Language Models
arXiv: 2206.07682

The quantisation model of neural scaling
arXiv: 2303.13506

Good news for human scientists

Enough data

There are research areas which do not have much data,
where model-driven is better than data-driven.

Enough model size, enough compute

A problem may be too complicated for a blackbox AI to solve,
but it may present simple structures humans can identify.

Pattern frequent enough

Arguably, the job of scientists is to forget about “common sense”,
in search of “outliers”.
Science = Exploring “rare” events

Precision vs Fuzziness

Human languages are fuzzy,
but math/physics are precise.
The way human beings formulate ideas is fuzzy.

「数学天才」陶哲轩：GPT-4无法攻克一个未解决的数学问题，但对工作有帮助

新智元 新智元 2023-04-09 00:12

Humanity: Camps and survival strategies

Human scientists' attitudes towards AI

- The Survivors（幸存派）： admit AI has limitations, but reject AI. Humans can work in areas where AI has limitations.
- The Collaborators (合作派)：admit AI has limitations, and accept AI. Only collaboration between AI and humans can lead to faster scientific development.
- The Redemptionists (拯救派)：admit AI has limitations, and improve AI. Use scientific tools to develop better AI.
- The Self-rescuers (自救派)：concern about AI safety. Humans need to focus on AI safety so that we can better understand and finally control AI.
- The worshipers (崇拜派)：AI is omnipotent, and embrace AI. Humans needn't and cannot understand AI. Just take advantage of AI power.
- The Adventists (降临派)：AI is omnipotent, embrace AI and become AI. Align human values to AI, not the other way around.

Disclaimer: Opinions are my own, which could be highly speculative.
Some of these beliefs are quite extreme. Most people lie on the continuous spectrum rather than extreme points.

Survival strategies

- The Survivors（幸存派）： admit AI has limitations, but reject AI. Humans can work in areas where AI has limitations. Do research areas that AI cannot solve.
- The Collaborators (合作派)：admit AI has limitations, and accept AI. Only collaboration between AI and humans can lead to faster scientific development. Find research areas that neither AI nor human can solve.
- The Redemptionists (拯救派)：admit AI has limitations, and improve AI. Use scientific tools to develop better AI.
- The Self-rescuers (自救派)：concern about AI safety. Humans need to focus on AI safety so that we can better understand and finally control AI. Research on AI interpretability.
- The worshipers (崇拜派)：AI is omnipotent, and embrace AI. Humans needn't and cannot understand AI. Just take advantage of AI power.
- The Adventists (降临派)：AI is omnipotent, embrace AI and become AI. Align human values to AI, not the other way around. I don't know…

Note: Some of these beliefs are quite extreme. Most people lie on the continuous spectrum rather than extreme points.

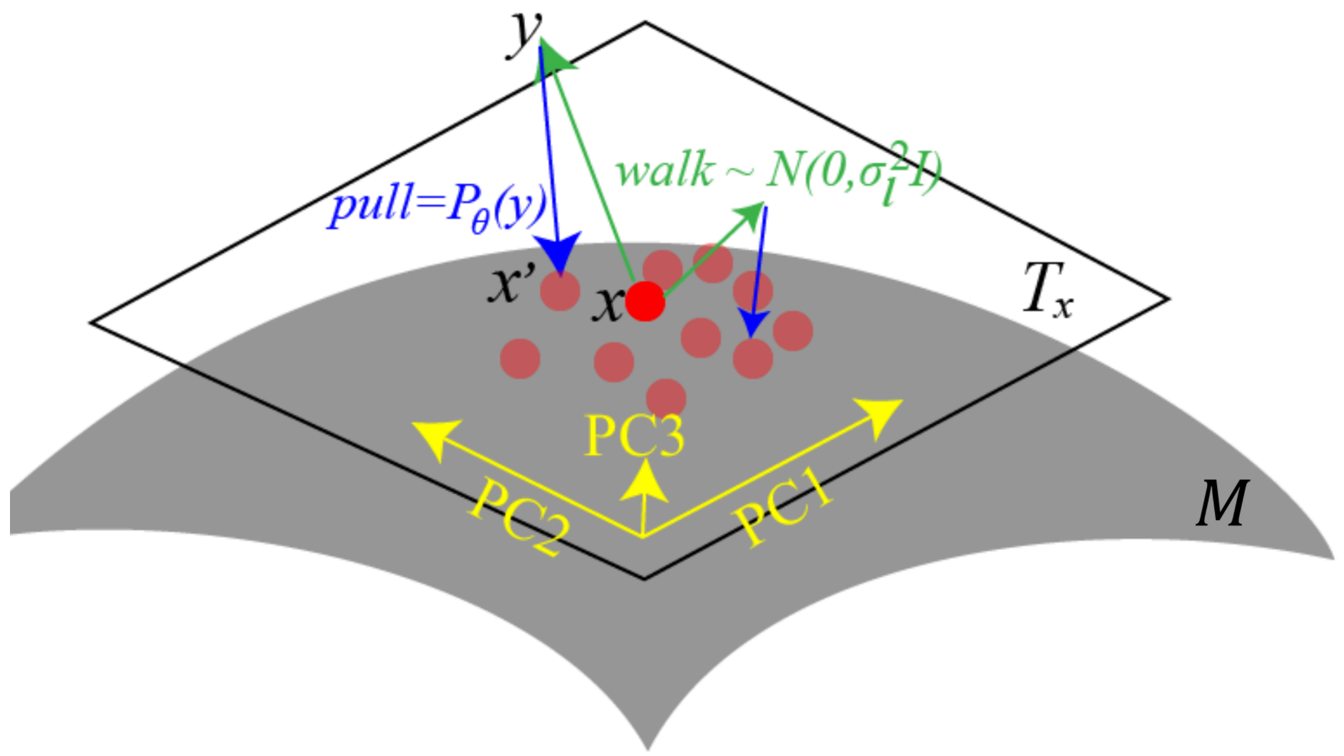
My Attitudes and Quest

- Collaborate, Rescue and self-rescue
- **Collaborate** with AI to speed up scientific discoveries
  - Example: Discovery of conserved quantities
- **Rescue** (improve) AI by leveraging tools developed in science
  - Example: Physics-inspired generative models
- **Self-Rescue** by research on AI interpretability/understanding
  - Example: Phase transitions, quantisation and modularity in neural networks

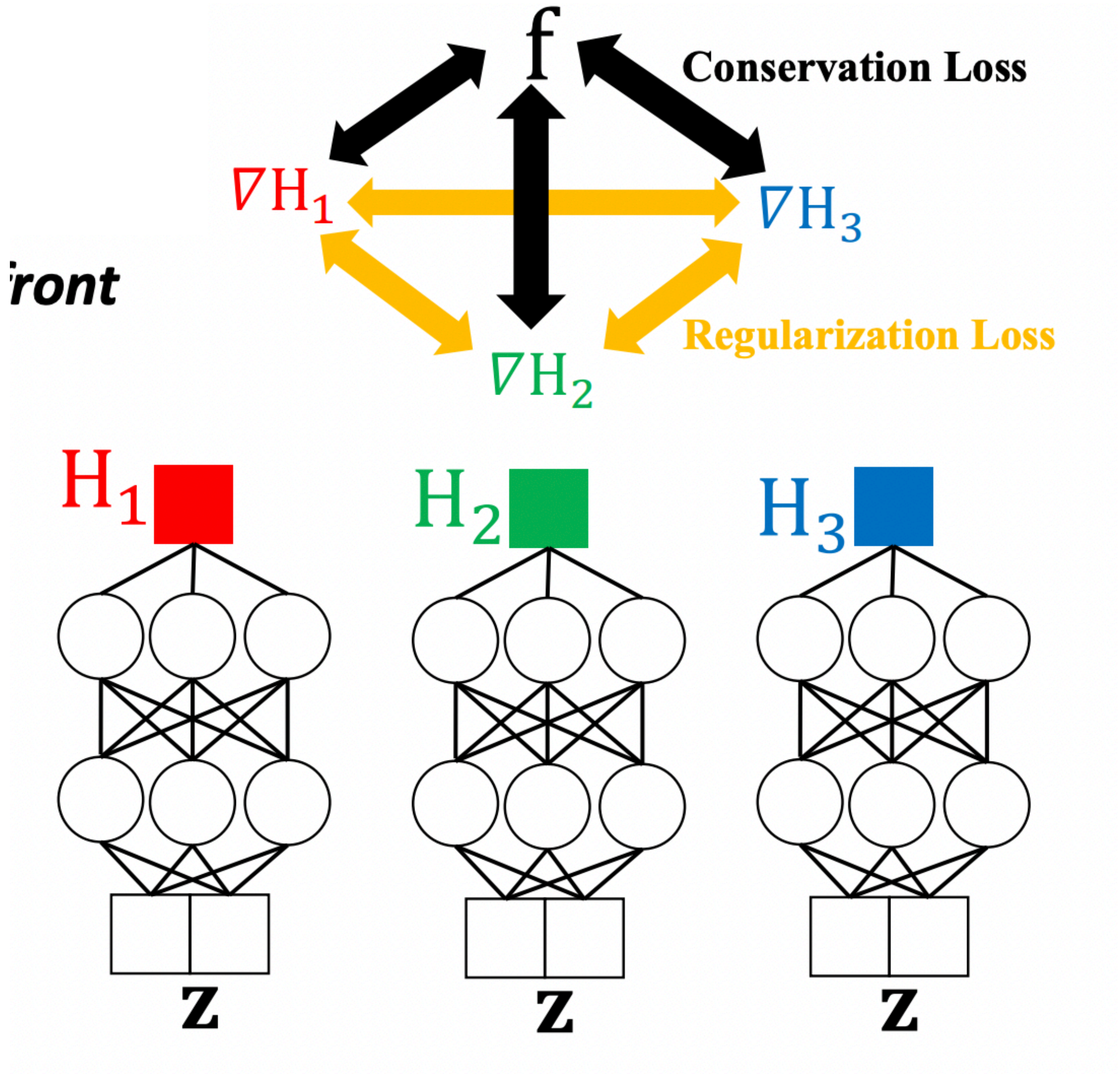
My quest - **Collaborate**:
Discovery of Conserved Quantities

AI Poincare (discovering conserved quantities)

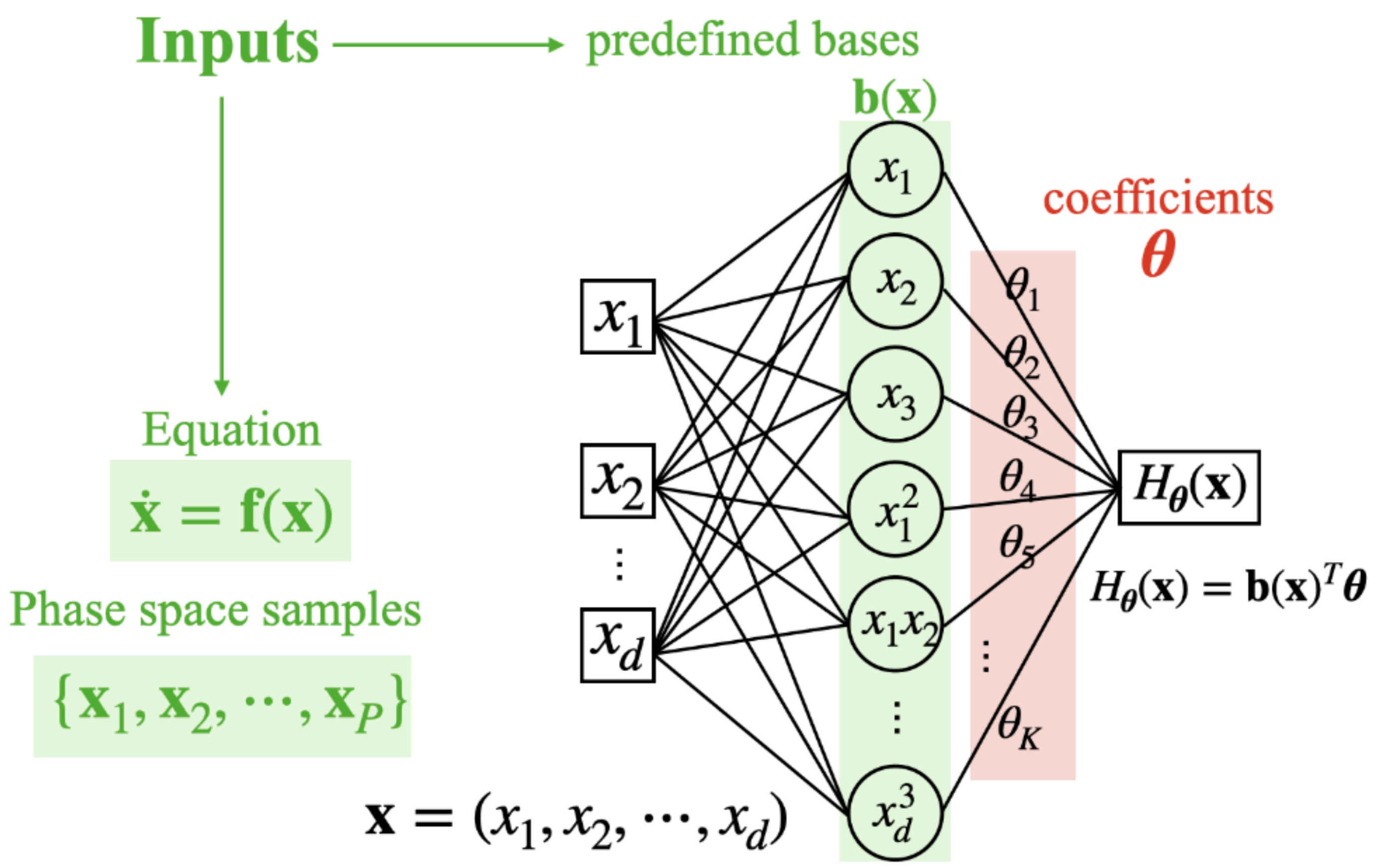
Data: trajectory
Assumption: No



Data: differential equations
Assumption: No



Data: differential equations
Assumption: basis functions



1.0
Phys. Rev. Lett. **126**, 180604

2.0
Phys. Rev. E **106**, 045307

3.0
arXiv:2305.19525

AI Poincare: less is more

Data: trajectory Assumption: No **1.0: A challenging setup**	Data: differential equations Assumption: No **2.0: A simpler setup**	Data: differential equations Assumption: basis functions **3.0: An even simpler setup**
Not interpretable	Reasonably interpretable	Extremely interpretable
Cannot discover unknown conserved quantities	Cannot discover unknown conserved quantities	**Discover unknown conserved quantities!!!**

1.0 — Phys. Rev. Lett. **126**, 180604

2.0 — Phys. Rev. E **106**, 045307

3.0 — arXiv:2305.19525

AI Poincare 3.0 is not trivial!

Differential equations

The derivation of equations of motion is similar to 2D, only with area switched to volume. The phase space is 24D: $\mathbf{x} = (x_i, y_i, z_i, u_i, v_i, w_i), i = 1, 2, 3, 4$. We list the equations here:

$$\begin{aligned}
\dot{x}_1 &= u_1, \dot{u}_1 = \lambda(-y_2z_3 + y_2z_4 + y_3z_2 - y_3z_4 - y_4z_2 + y_4z_3),\\
\dot{y}_1 &= v_1, \dot{v}_1 = \lambda(x_2z_3 - x_2z_4 - x_3z_2 + x_3z_4 + x_4z_2 - x_4z_3),\\
\dot{z}_1 &= w_1, \dot{w}_1 = \lambda(-x_2y_3 + x_2y_4 + x_3y_2 - x_3y_4 - x_4y_2 + x_4y_3),\\
\dot{x}_2 &= u_2, \dot{u}_2 = \lambda(y_1z_3 - y_1z_4 - y_3z_1 + y_3z_4 + y_4z_1 - y_4z_3),\\
\dot{y}_2 &= v_2, \dot{v}_2 = \lambda(-x_1z_3 + x_1z_4 + x_3z_1 - x_3z_4 - x_4z_1 + x_4z_3),\\
\dot{z}_2 &= w_2, \dot{w}_2 = \lambda(x_1y_3 - x_1y_4 - x_3y_1 + x_3y_4 + x_4y_1 - x_4y_3),\\
\dot{x}_3 &= u_3, \dot{u}_3 = \lambda(-y_1z_2 + y_1z_4 + y_2z_1 - y_2z_4 - y_4z_1 + y_4z_2),\\
\dot{y}_3 &= v_3, \dot{v}_3 = \lambda(x_1z_2 - x_1z_4 - x_2z_1 + x_2z_4 + x_4z_1 - x_4z_2),\\
\dot{z}_3 &= w_3, \dot{w}_3 = \lambda(-x_1y_2 + x_1y_4 + x_2y_1 - x_2y_4 - x_4y_1 + x_4y_2),\\
\dot{x}_4 &= u_4, \dot{u}_4 = \lambda \cdot (y_1z_2 - y_1z_3 - y_2z_1 + y_2z_3 + y_3z_1 - y_3z_2),\\
\dot{y}_4 &= v_4, \dot{v}_4 = \lambda(-x_1z_2 + x_1z_3 + x_2z_1 - x_2z_3 - x_3z_1 + x_3z_2),\\
\dot{z}_4 &= w_4, \dot{w}_4 = \lambda(x_1y_2 - x_1y_3 - x_2y_1 + x_2y_3 + x_3y_1 - x_3y_2),
\end{aligned} \tag{C6}$$

where

$$\begin{aligned}
\lambda &= p/q,\\
p &= x_1(-v_2w_3 + v_2w_4 + v_3w_2 - v_3w_4 - v_4w_2 + v_4w_3) + y_1(u_2w_3 - u_2w_4 - u_3w_2 + u_3w_4 + u_4w_2 - u_4w_3)\\
&\quad + z_1(-u_2v_3 + u_2v_4 + u_3v_2 - u_3v_4 - u_4v_2 + u_4v_3) + x_2(v_1w_3 - v_1w_4 - v_3w_1 + v_3w_4 + v_4w_1 - v_4w_3)\\
&\quad + y_2(-u_1w_3 + u_1w_4 + u_3w_1 - u_3w_4 - u_4w_1 + u_4w_3) + z_2(u_1v_3 - u_1v_4 - u_3v_1 + u_3v_4 + u_4v_1 - u_4v_3)\\
&\quad + x_3(-v_1w_2 + v_1w_4 + v_2w_1 - v_2w_4 - v_4w_1 + v_4w_2) + y_3(u_1w_2 - u_1w_4 - u_2w_1 + u_2w_4 + u_4w_1 - u_4w_2)\\
&\quad + z_3(-u_1v_2 + u_1v_4 + u_2v_1 - u_2v_4 - u_4v_1 + u_4v_2) + x_4(v_1w_2 - v_1w_3 - v_2w_1 + v_2w_3 + v_3w_1 - v_3w_2)\\
&\quad + y_4(-u_1w_2 + u_1w_3 + u_2w_1 - u_2w_3 - u_3w_1 + u_3w_2) + z_4(u_1v_2 - u_1v_3 - u_2v_1 + u_2v_3 + u_3v_1 - u_3v_2)\\
q &= (-y_2z_3 + y_2z_4 + y_3z_2 - y_3z_4 - y_4z_2 + y_4z_3)^2 + (x_2z_3 - x_2z_4 - x_3z_2 + x_3z_4 + x_4z_2 - x_4z_3)^2\\
&\quad + (-x_2y_3 + x_2y_4 + x_3y_2 - x_3y_4 - x_4y_2 + x_4y_3)^2 + (y_1z_3 - y_1z_4 - y_3z_1 + y_3z_4 + y_4z_1 - y_4z_3)^2\\
&\quad + (-x_1z_3 + x_1z_4 + x_3z_1 - x_3z_4 - x_4z_1 + x_4z_3)^2 + (x_1y_3 - x_1y_4 - x_3y_1 + x_3y_4 + x_4y_1 - x_4y_3)^2\\
&\quad + (-y_1z_2 + y_1z_4 + y_2z_1 - y_2z_4 - y_4z_1 + y_4z_2)^2 + (x_1z_2 - x_1z_4 - x_2z_1 + x_2z_4 + x_4z_1 - x_4z_2)^2\\
&\quad + (-x_1y_2 + x_1y_4 + x_2y_1 - x_2y_4 - x_4y_1 + x_4y_2)^2 + (y_1z_2 - y_1z_3 - y_2z_1 + y_2z_3 + y_3z_1 - y_3z_2)^2\\
&\quad + (-x_1z_2 + x_1z_3 + x_2z_1 - x_2z_3 - x_3z_1 + x_3z_2)^2 + (x_1y_2 - x_1y_3 - x_2y_1 + x_2y_3 + x_3y_1 - x_3y_2)^2
\end{aligned} \tag{C7}$$

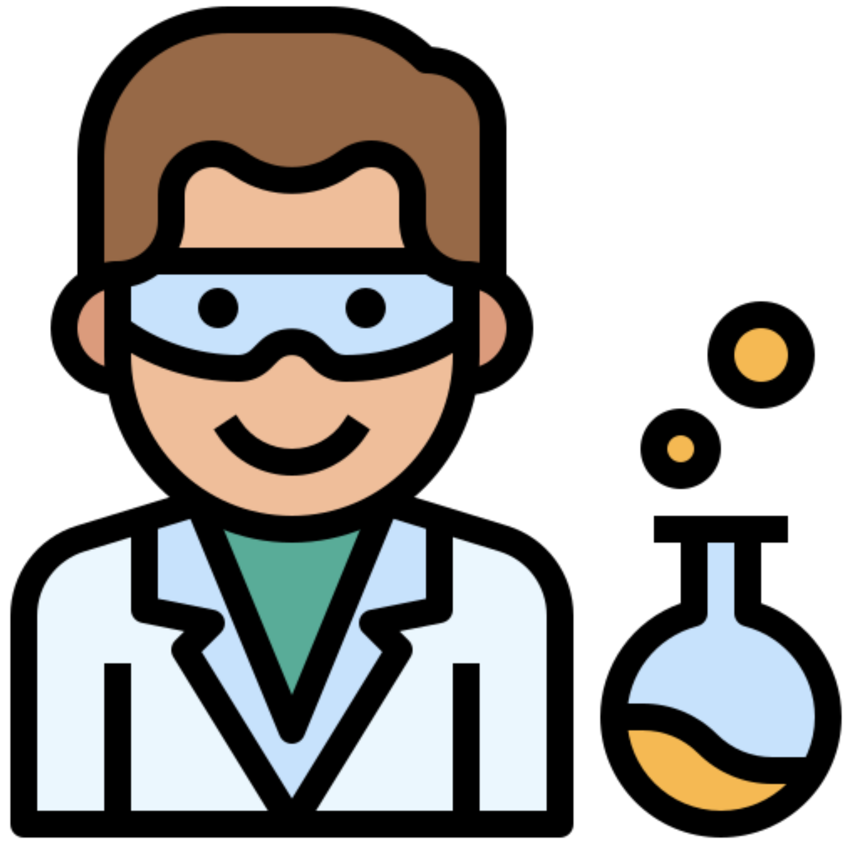
Constraints

and with the constant-volume constraint:

$$\begin{aligned}
0 &= (-y_2z_3 + y_2z_4 + y_3z_2 - y_3z_4 - y_4z_2 + y_4z_3)u_1 + (x_2z_3 - x_2z_4 - x_3z_2 + x_3z_4 + x_4z_2 - x_4z_3)v_1\\
&\quad + (-x_2y_3 + x_2y_4 + x_3y_2 - x_3y_4 - x_4y_2 + x_4y_3)w_1 + (y_1z_3 - y_1z_4 - y_3z_1 + y_3z_4 + y_4z_1 - y_4z_3)u_2\\
&\quad + (-x_1z_3 + x_1z_4 + x_3z_1 - x_3z_4 - x_4z_1 + x_4z_3)v_2 + (x_1y_3 - x_1y_4 - x_3y_1 + x_3y_4 + x_4y_1 - x_4y_3)w_2\\
&\quad + (-y_1z_2 + y_1z_4 + y_2z_1 - y_2z_4 - y_4z_1 + y_4z_2)u_3 + (x_1z_2 - x_1z_4 - x_2z_1 + x_2z_4 + x_4z_1 - x_4z_2)v_3\\
&\quad + (-x_1y_2 + x_1y_4 + x_2y_1 - x_2y_4 - x_4y_1 + x_4y_2)w_3 + (y_1z_2 - y_1z_3 - y_2z_1 + y_2z_3 + y_3z_1 - y_3z_2)u_4\\
&\quad + (-x_1z_2 + x_1z_3 + x_2z_1 - x_2z_3 - x_3z_1 + x_3z_2)v_4 + (x_1y_2 - x_1y_3 - x_2y_1 + x_2y_3 + x_3y_1 - x_3y_2)w_4
\end{aligned} \tag{C8}$$

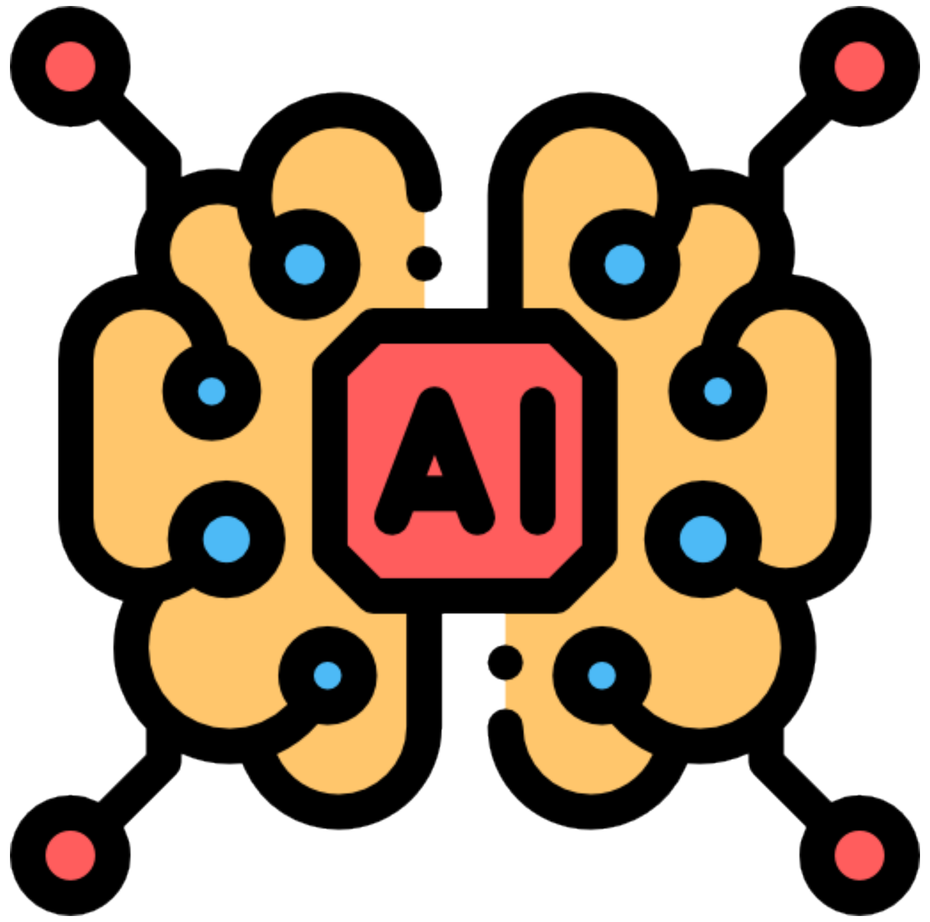
Question: How many conserved quantities do the differential equations have?

Human



(Two months' hard work)
I found 12!

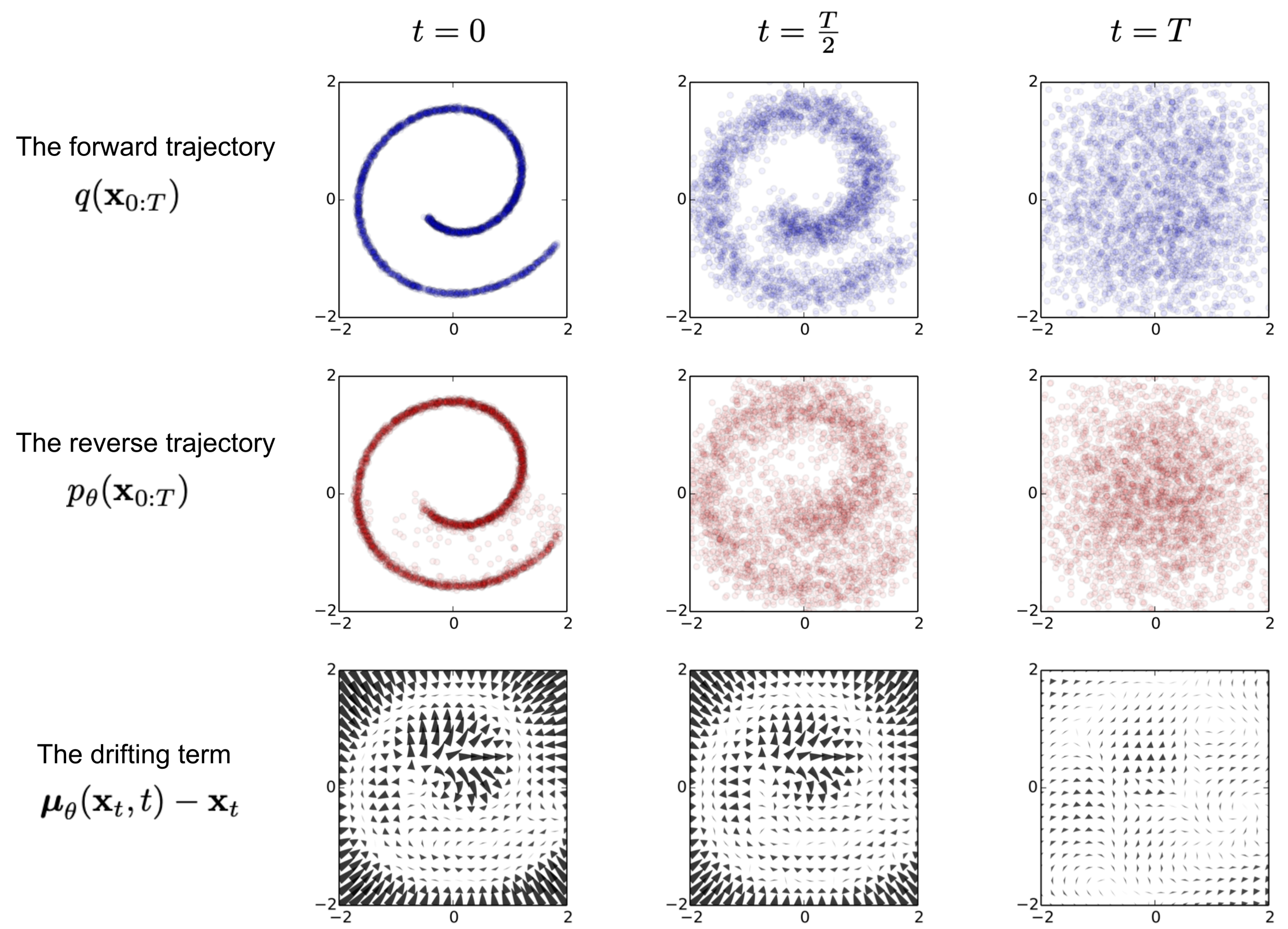
AI Poincare 3.0

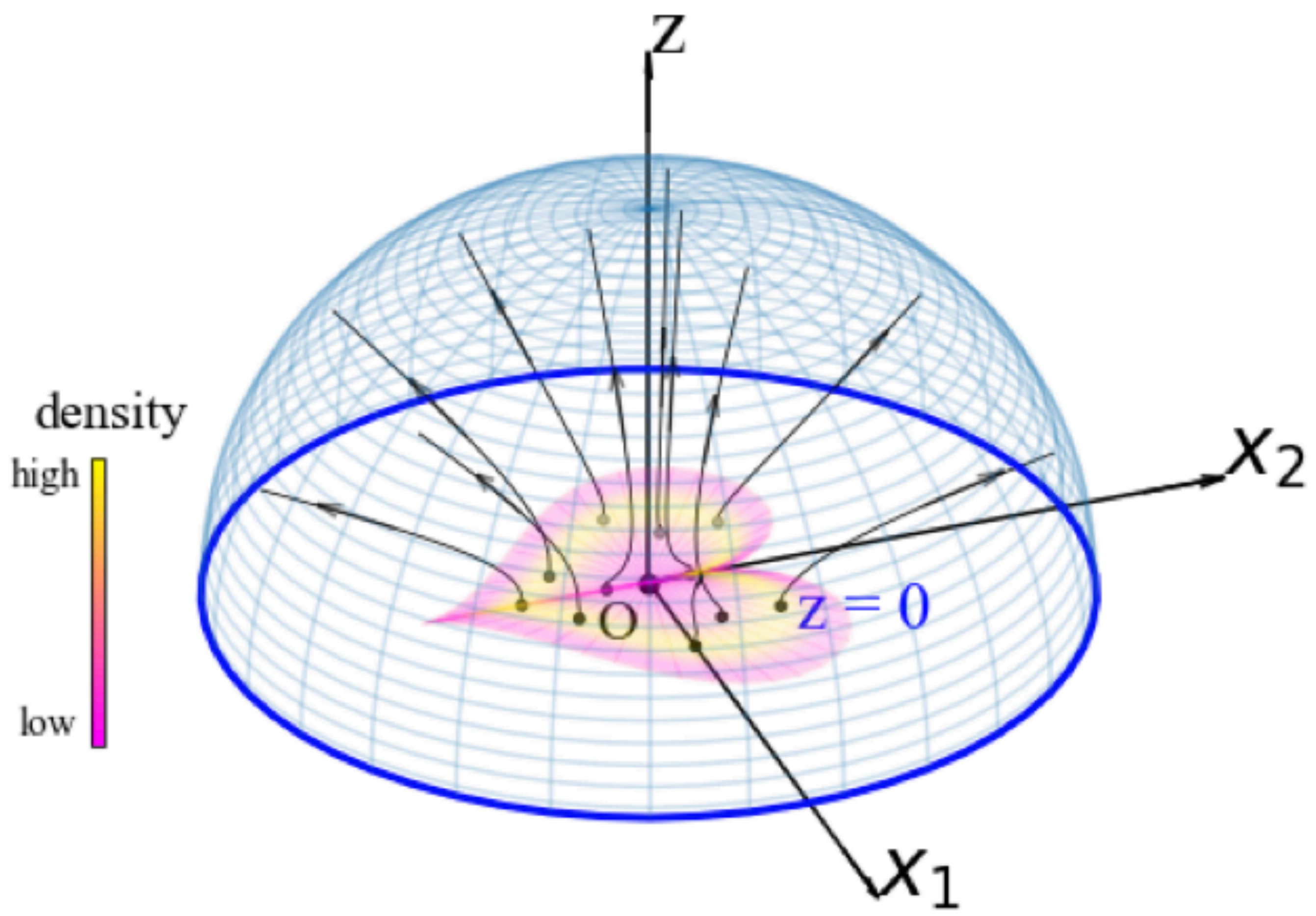


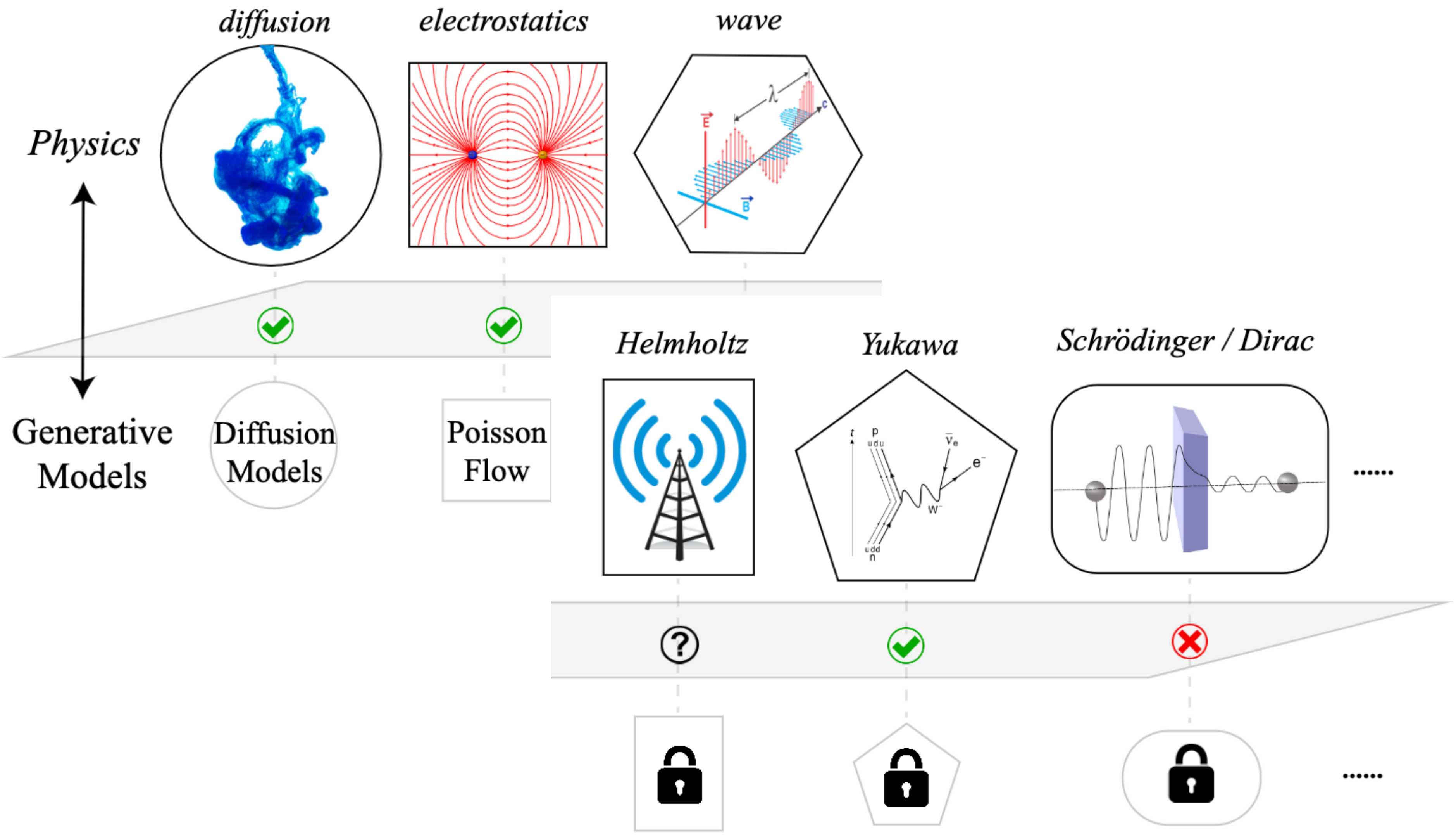
(< 1 minute)
I found 14!

My quest - **Rescue:**
Physics-inspired generative models

Generative models from physical processes







Diffusion

Electrostatics

Any smooth physical process

Diffusion model

(Not my work)

arXiv: 1503.03585

Poisson Flow

Poisson Flow Generative Models

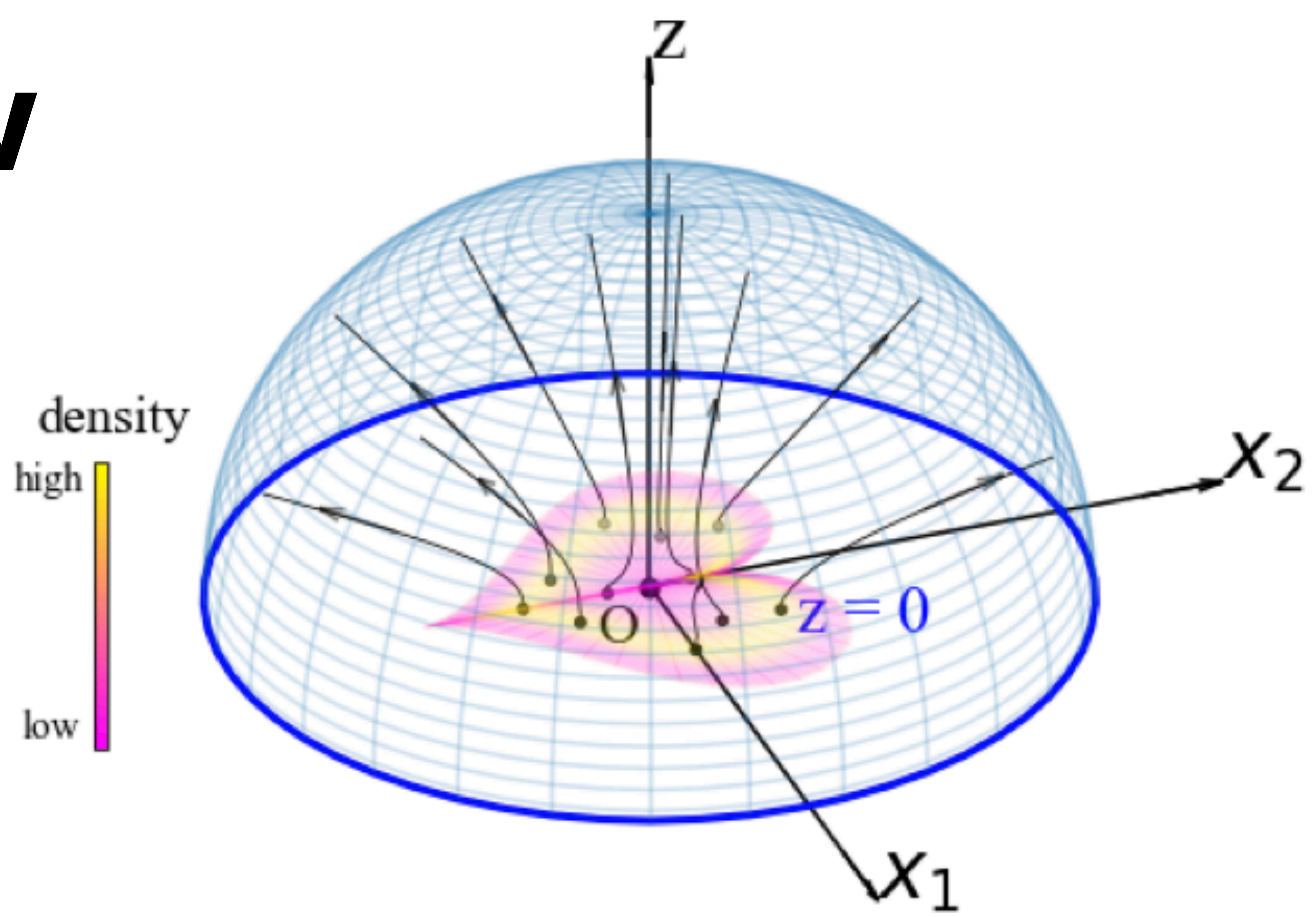
arXiv: 2209.11178 (NeurIPS 2022)

GenPhys

From Physical Processes to Generative Models

arXiv: 2304.02637

Poisson Flow



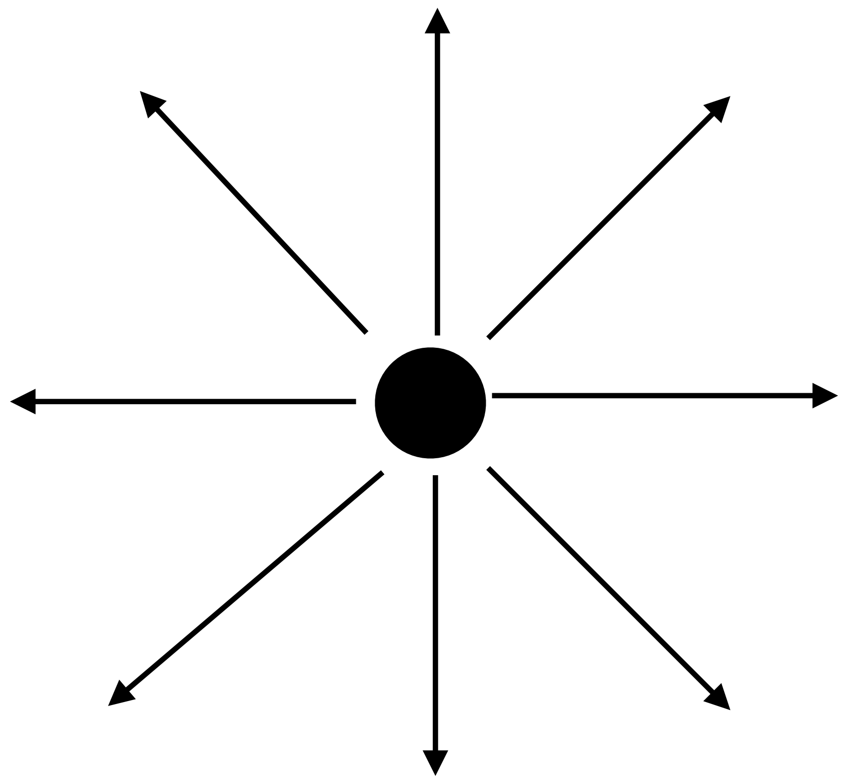
Generative modeling	⟷	*Physics*
data point/sample	⟷	electric charge
data distribution	⟷	charge distribution
flow	⟷	electric field/flux
bijective map	⟷	electric lines

Electrostatics

Poisson Equation $\nabla^2 \varphi = -\rho(x),\ \nabla^2 \equiv \sum_{i=1}^{N} \partial_i^2$

Green function $\nabla^2 G(x, x') = -\delta(x - x') \Longrightarrow G(x, x') \sim r^{-(N-2)},\ r \equiv ||x - x'||$

$$\varphi(x) = \int G(x, x')\rho(x')dx' \qquad E(x) = -\nabla \varphi(x) \sim \int \frac{x - x'}{||x - x'||^N}\rho(x')dx'$$



$$N = 3 \Longrightarrow \varphi(r) \sim \frac{1}{r},\ E(r) \sim \frac{1}{r^2}$$

$$N \geq 3 \Longrightarrow \varphi(r) \sim \frac{1}{r^{N-2}},\ E(r) \sim \frac{1}{r^{N-1}}$$

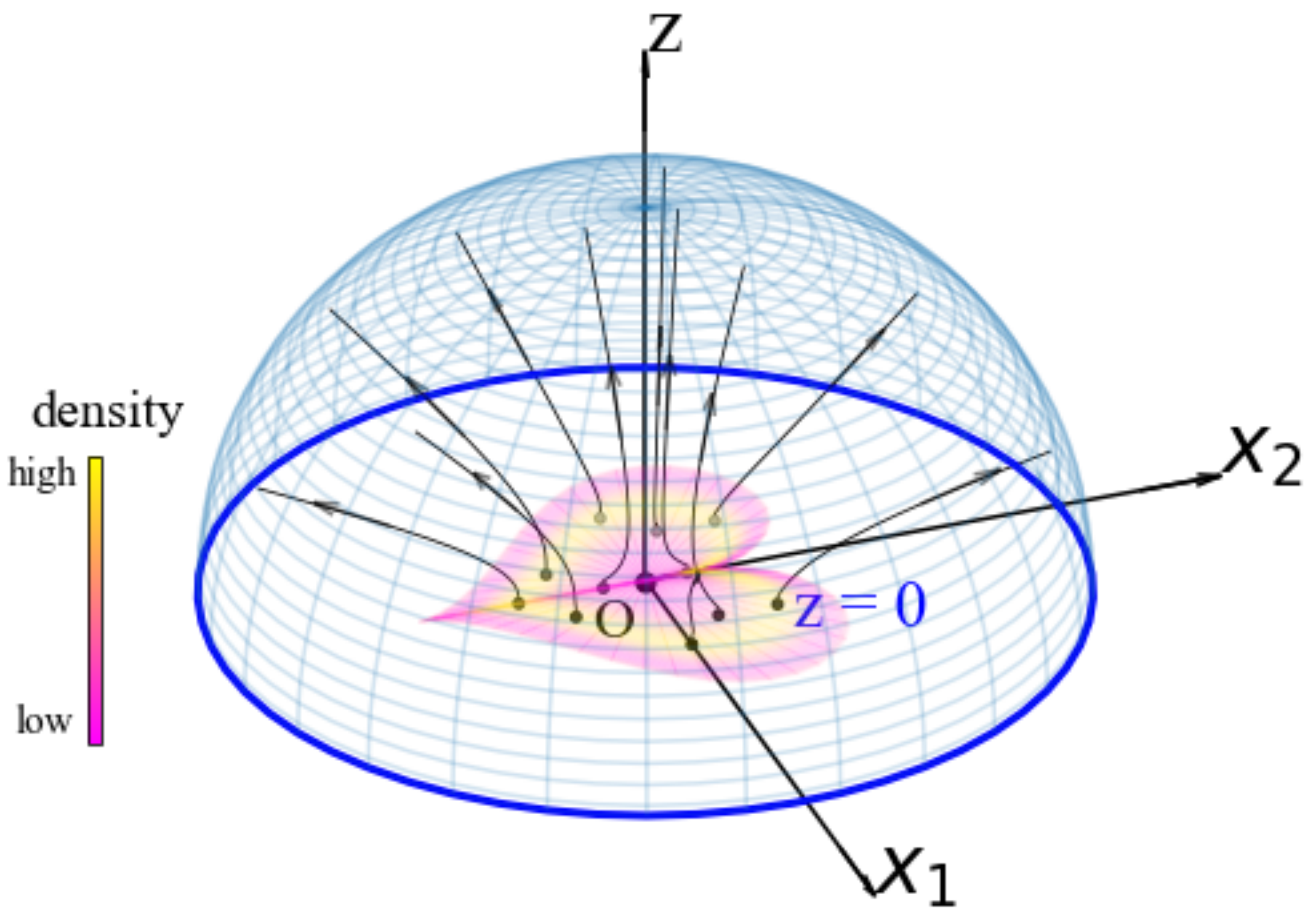
Electric field

↓ generalize

Poisson field

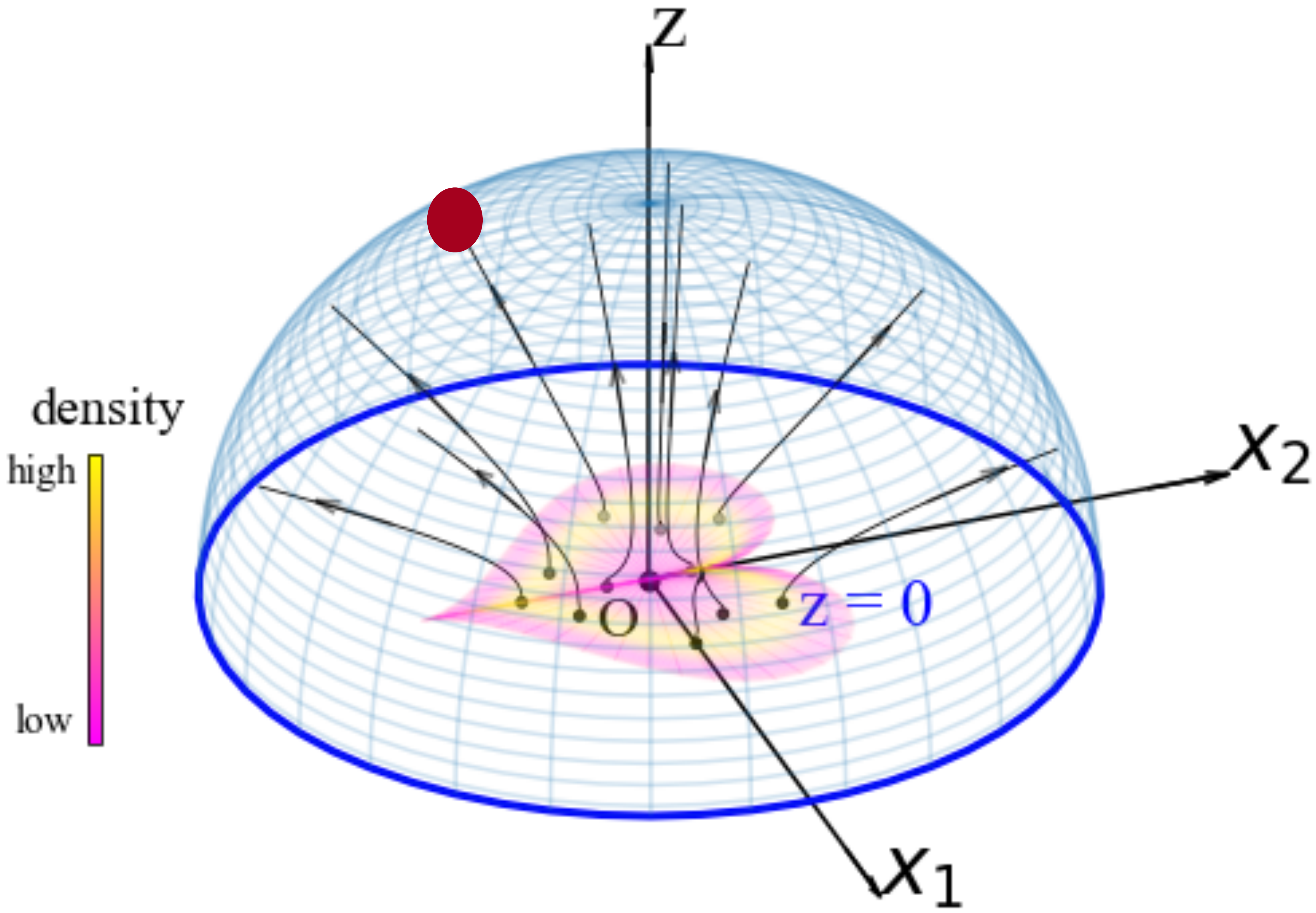
Poisson Flow: a New Flow Inspired by Electrostatics

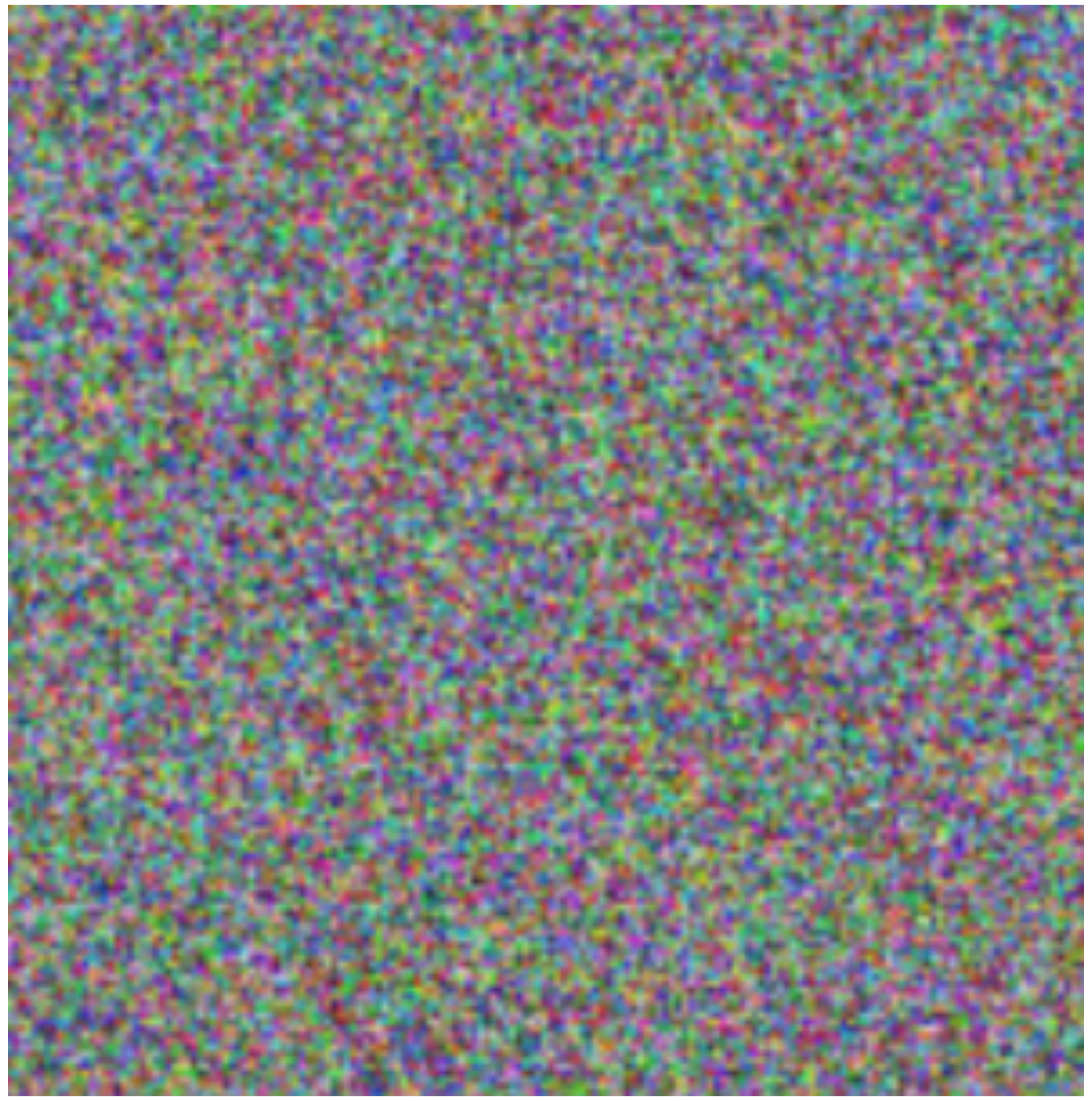
- Interpret N-dim **data distribution** as **charge density**
- Placing the charges on the $z = 0$ hyperplane in an $N + 1$-dim space augmented with dimension z
- Electric Field Lines ***define a* bijection** between ***data distribution*** and a ***uniform distribution on the large hemisphere***



Generation

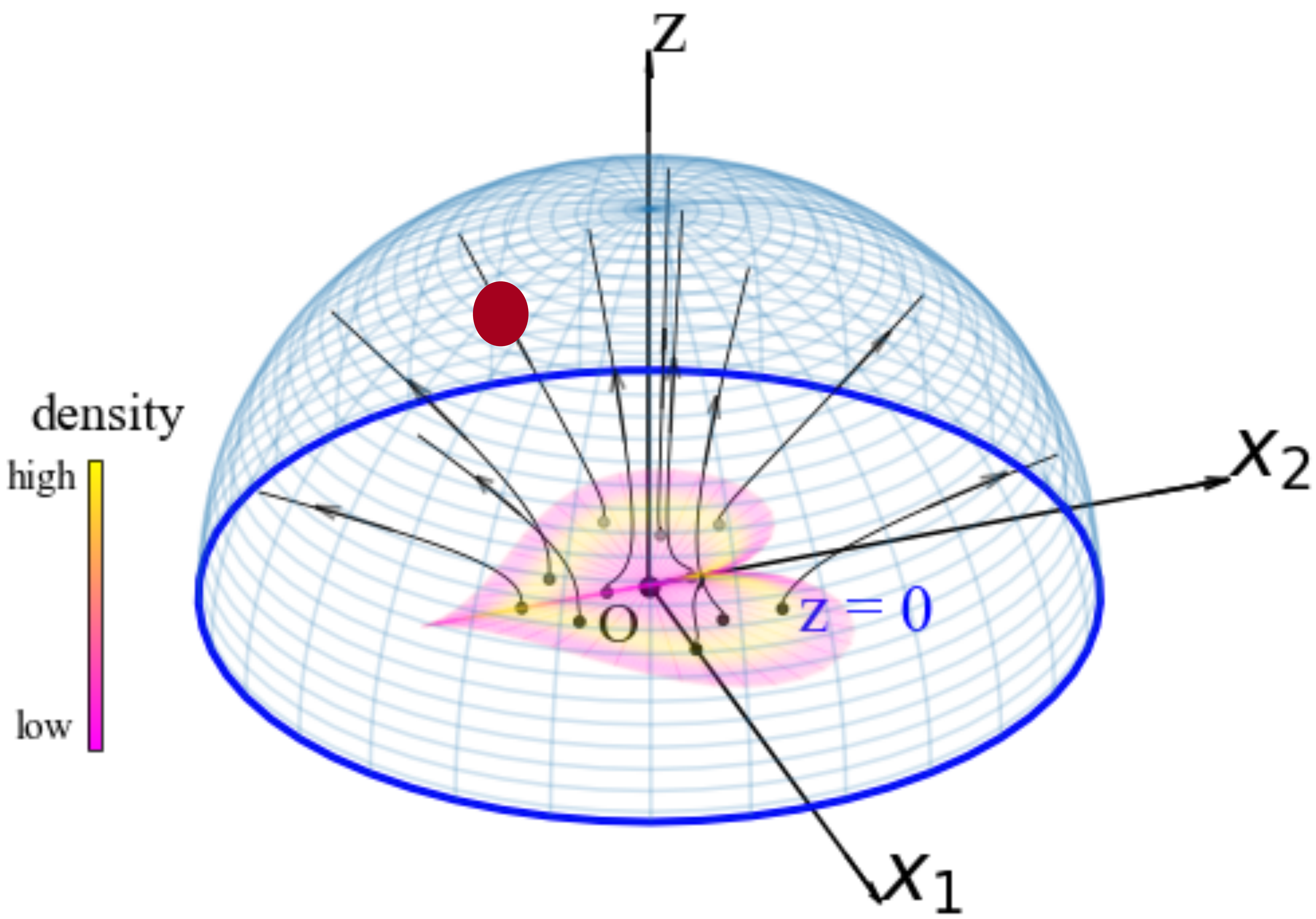
1. Uniformly sampling an initial sample (●) on the hemisphere.

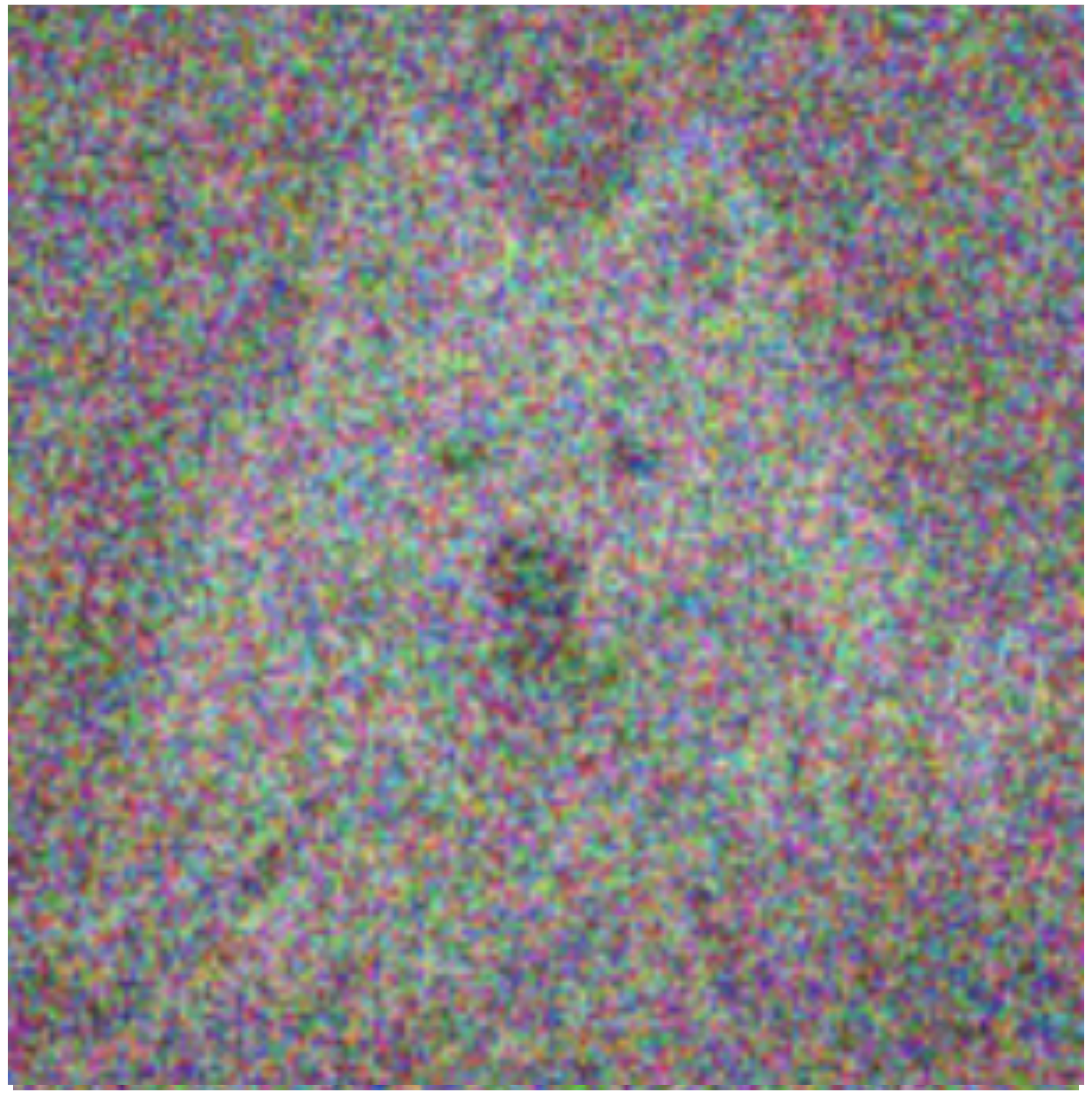




Generation

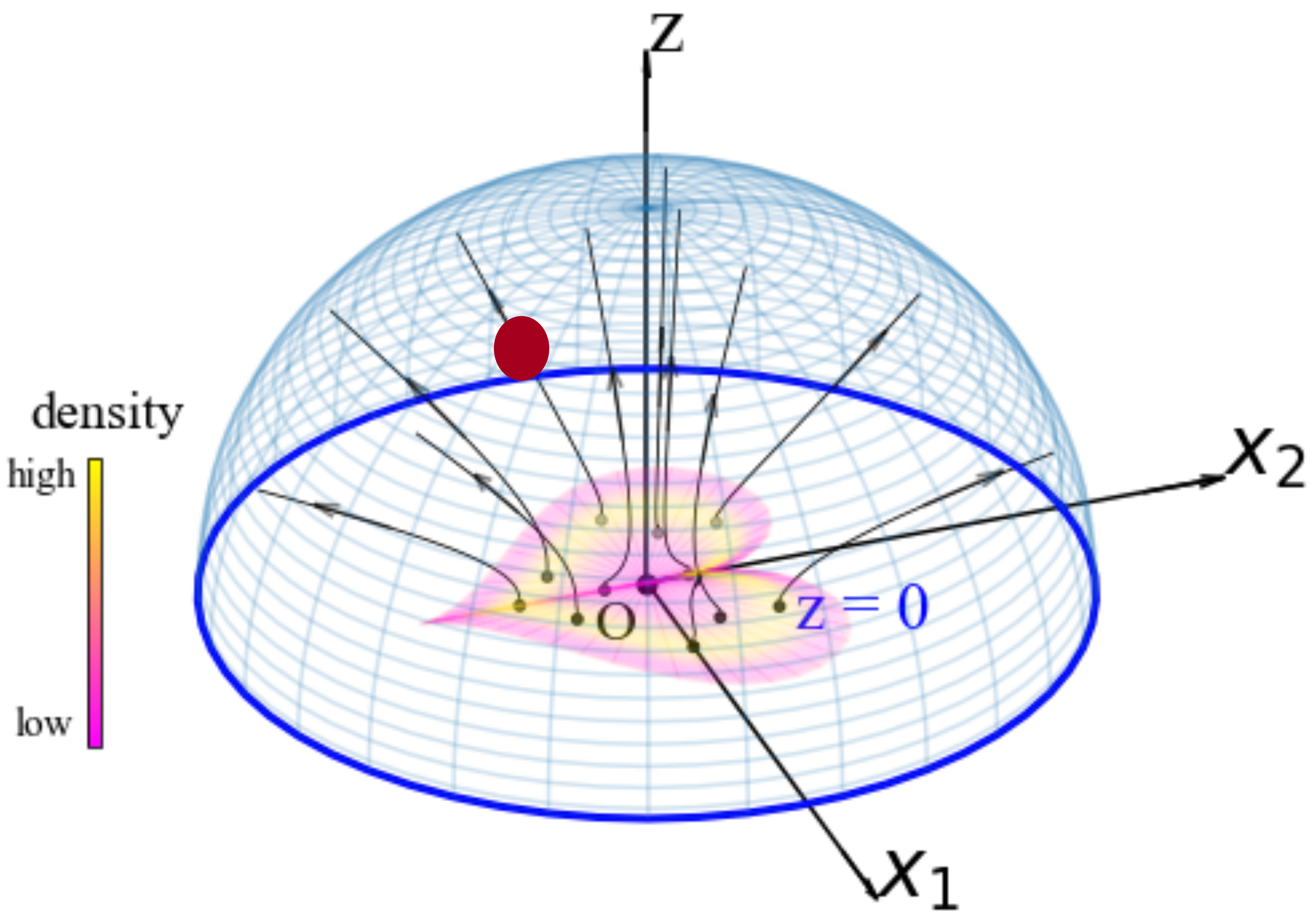
1. Uniformly sampling an initial sample (●) on the hemisphere.
2. Evolving the sample by following the corresponding electric field line.





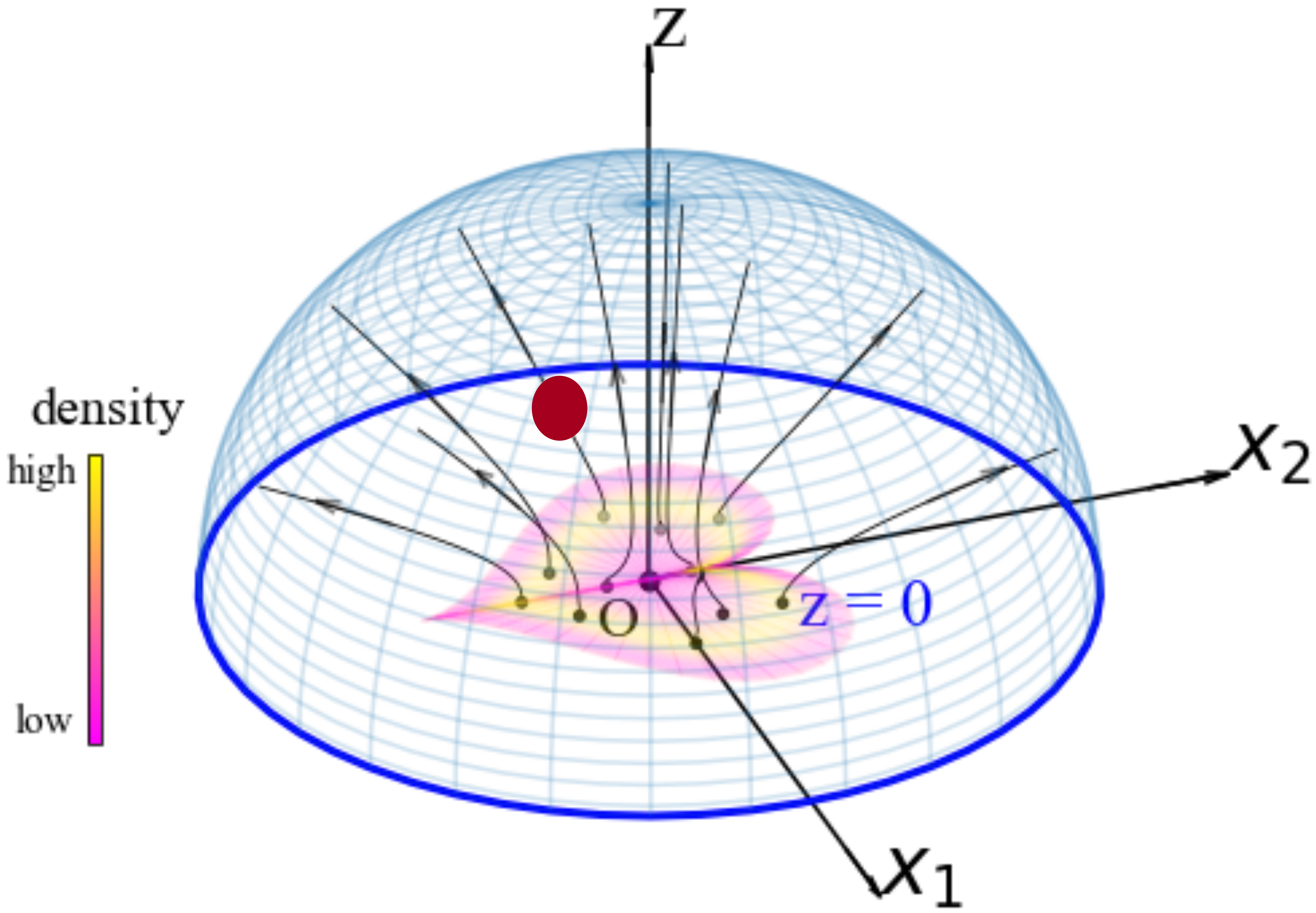
Generation

1. Uniformly sampling an initial sample (●) on the hemisphere.
2. Evolving the sample by following the corresponding electric field line.



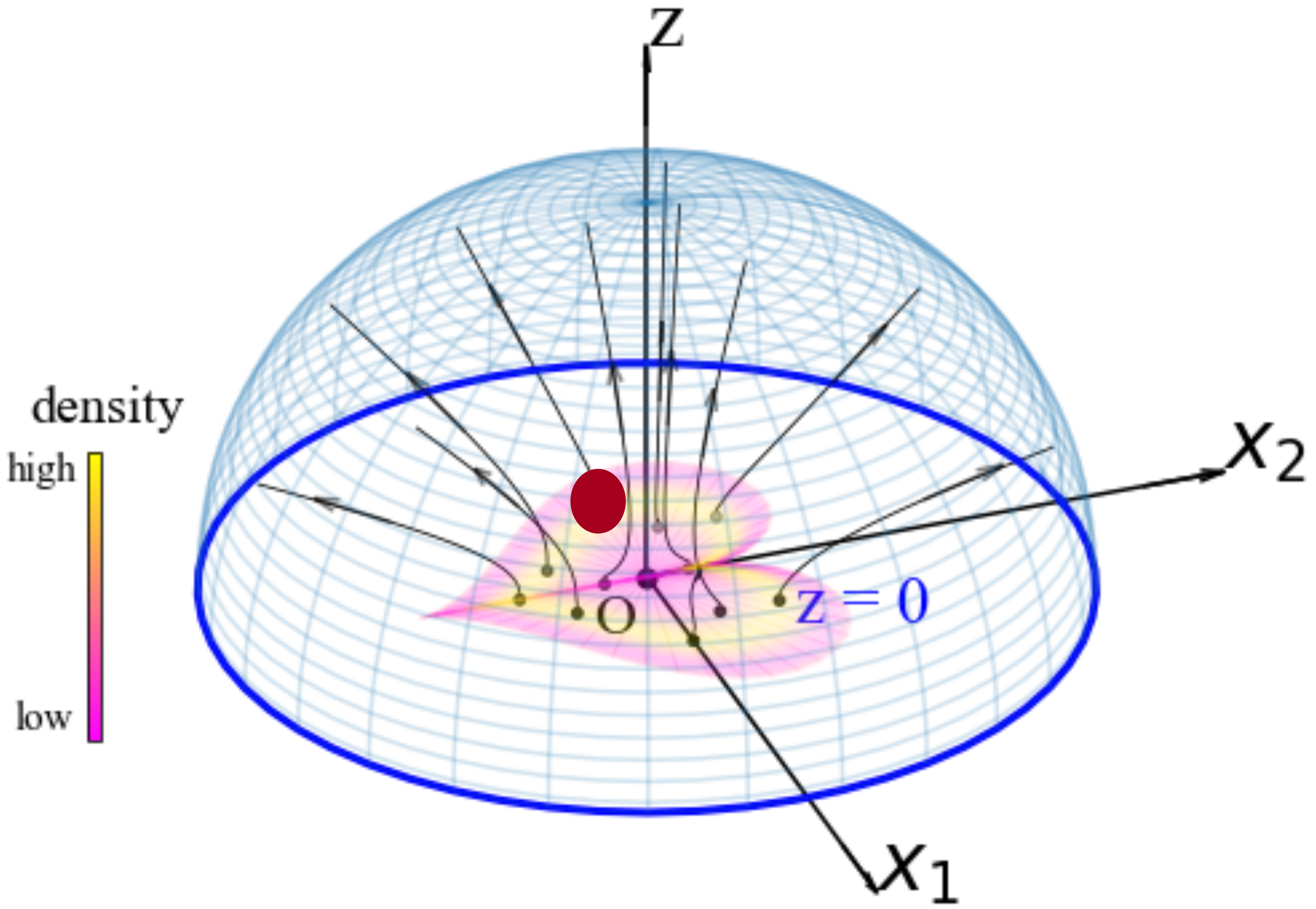
Generation

1. Uniformly sampling an initial sample (●) on the hemisphere.
2. Evolving the sample by following the corresponding electric field line.



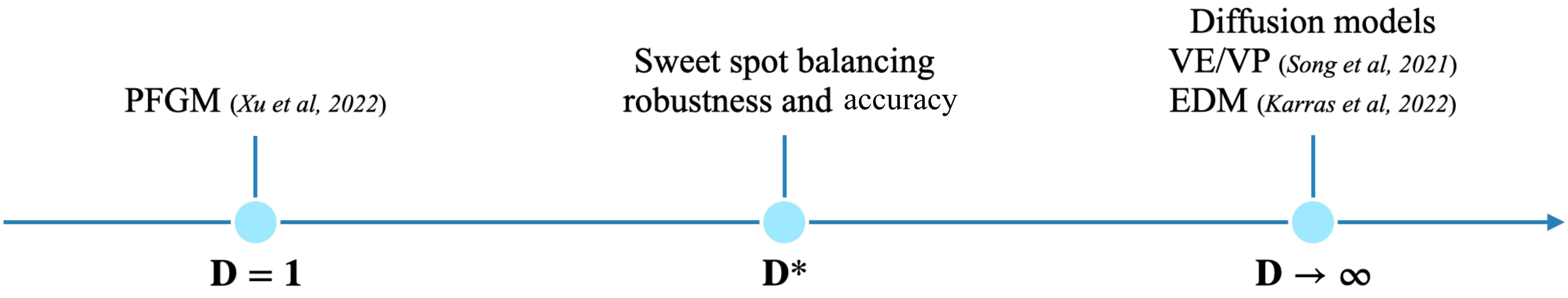
Generation

1. Uniformly sampling an initial sample (●) on the hemisphere.
2. Evolving the sample by following the corresponding electric field line.
3. Stopping the process when $z = 0$.



Q: Relation between Diffusion Models and Poisson Flows?

PFGM++ framework



PFGM++ ($\mathbf{D} \in \mathbb{R}^+$)

Extension from PFGM

Higher-dimensional augmentation

Perturbation-based training objective

Equivalence between $D \to \infty$ and diffusion models

Field / Sampling equivalence

Training equivalence

Xu, Liu, Tong, Tian, Tegmark, and Jaakkola. *PFGM++: Unlocking the Potential of Physics-Inspired Generative Models*, arXiv: 2302.04265

Q: Is there a universal converter from physics to generative models?

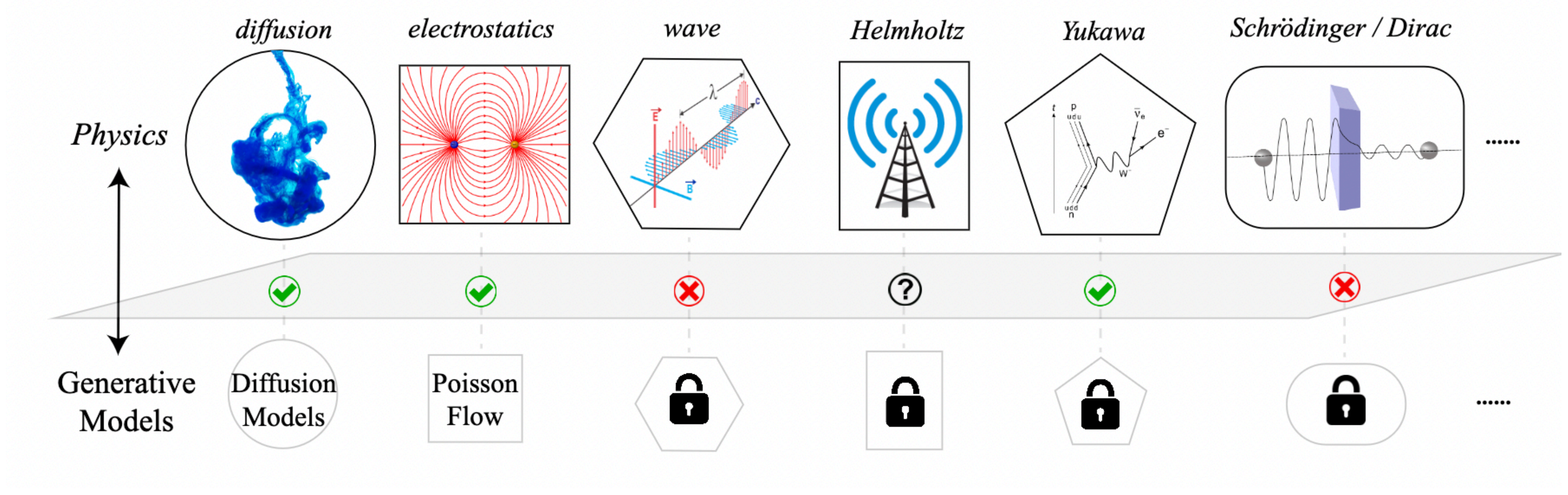
GenPhys: From Physical Processes to Generative Models

arXiv: 2304.02637

A: Yes, but…

Yes: A concrete protocol that converts physics to generative models

but: the converted generative models may not have desirable properties



Converter: partial differential equations (PDEs)

A *physical process* is described by a **PDE**

GenPhys: From Physical Processes to Generative Models
arXiv: 2304.02637

$$\hat{L}\phi \equiv F(\phi, \phi_t, \phi_{tt}, \nabla\phi, \nabla^2\phi, \ldots) = f(\mathbf{x}, t) \quad \text{(Physical PDE)}$$

Physicists already know how to solve it, …

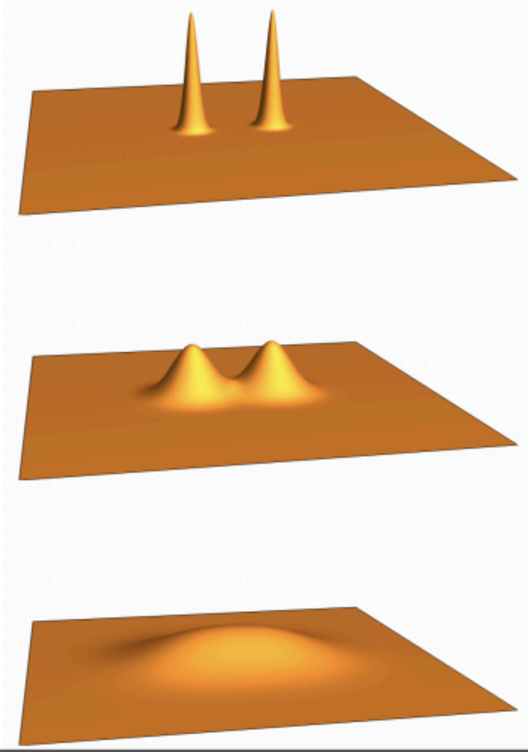
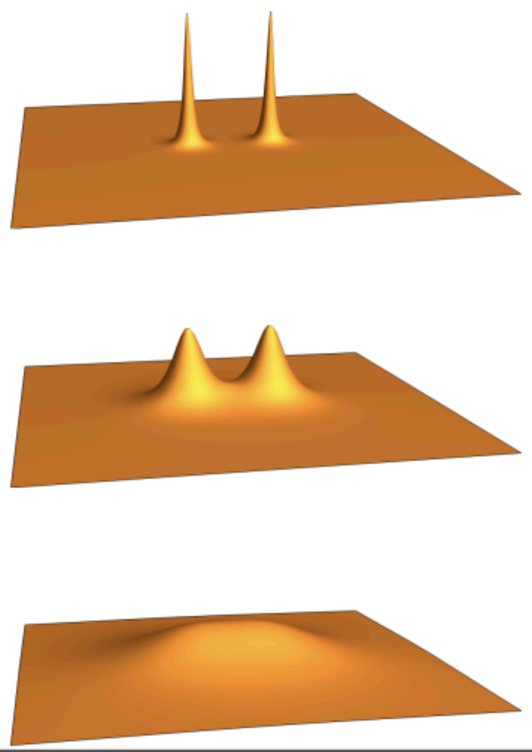
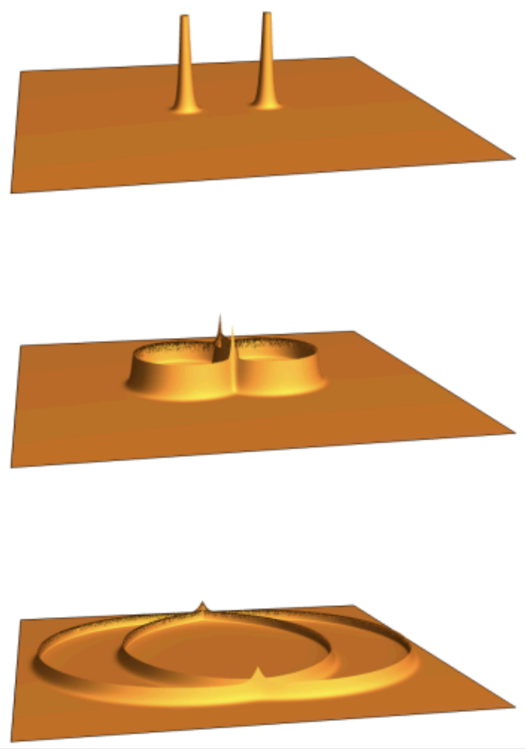
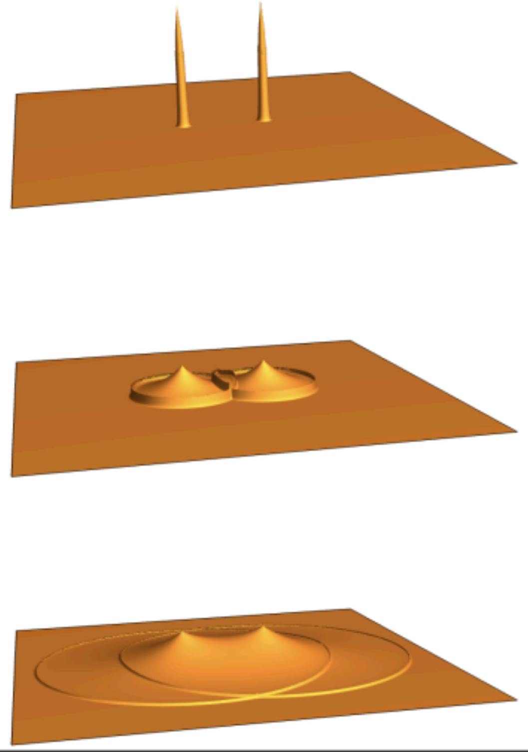
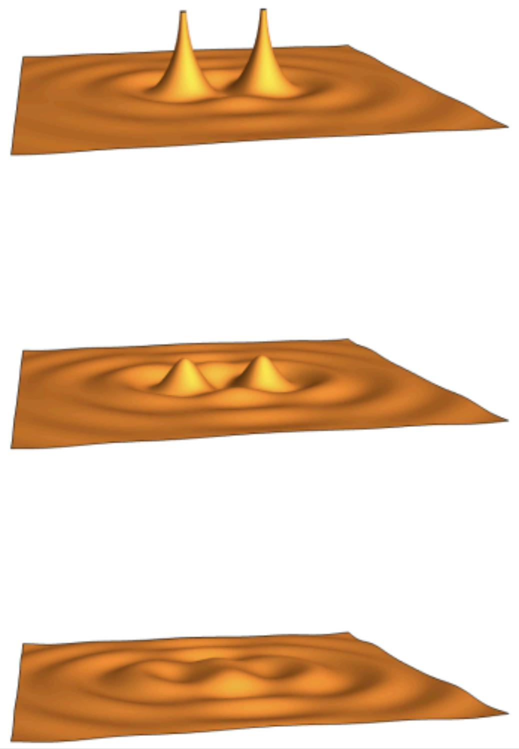
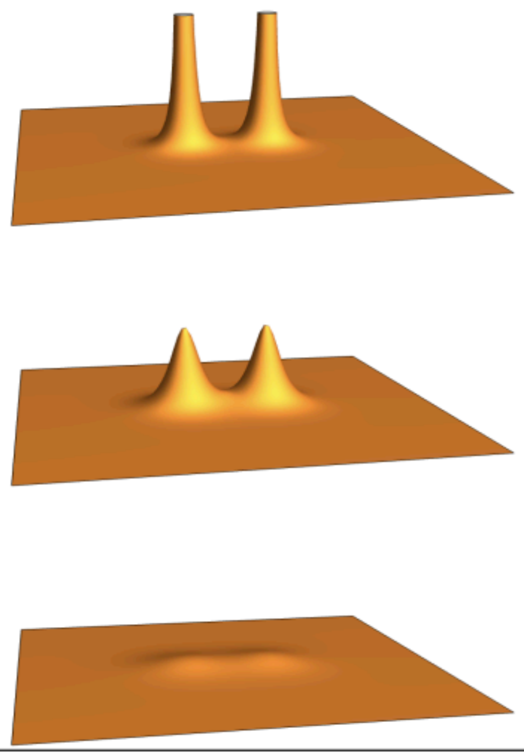
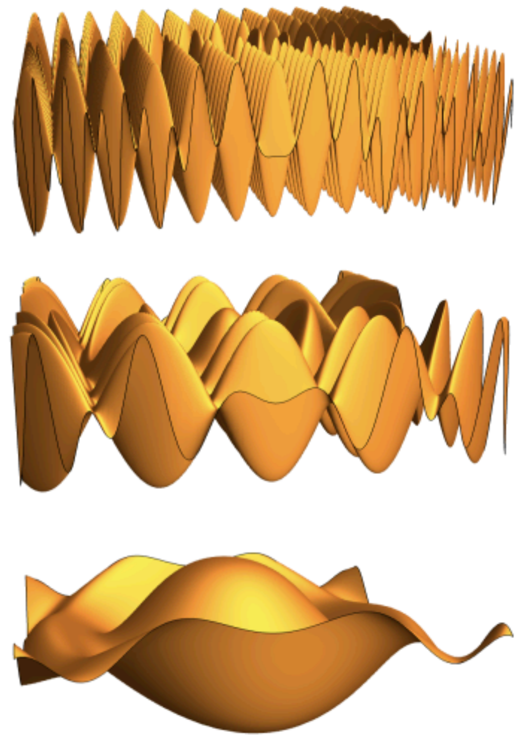
…, if they are equivalent, …

A *generative model* is associated with a density flow (which is also a **PDE**)

$$\hat{M}(p, \mathbf{v}, R) \equiv \frac{\partial p(\mathbf{x}, t)}{\partial t} + \nabla \cdot [p(\mathbf{x}, t)\mathbf{v}(\mathbf{x}, t)] - R(\mathbf{x}, t) = p_{\mathrm{data}}(\mathbf{x})\delta(t) \quad \text{(Density Flow)}$$

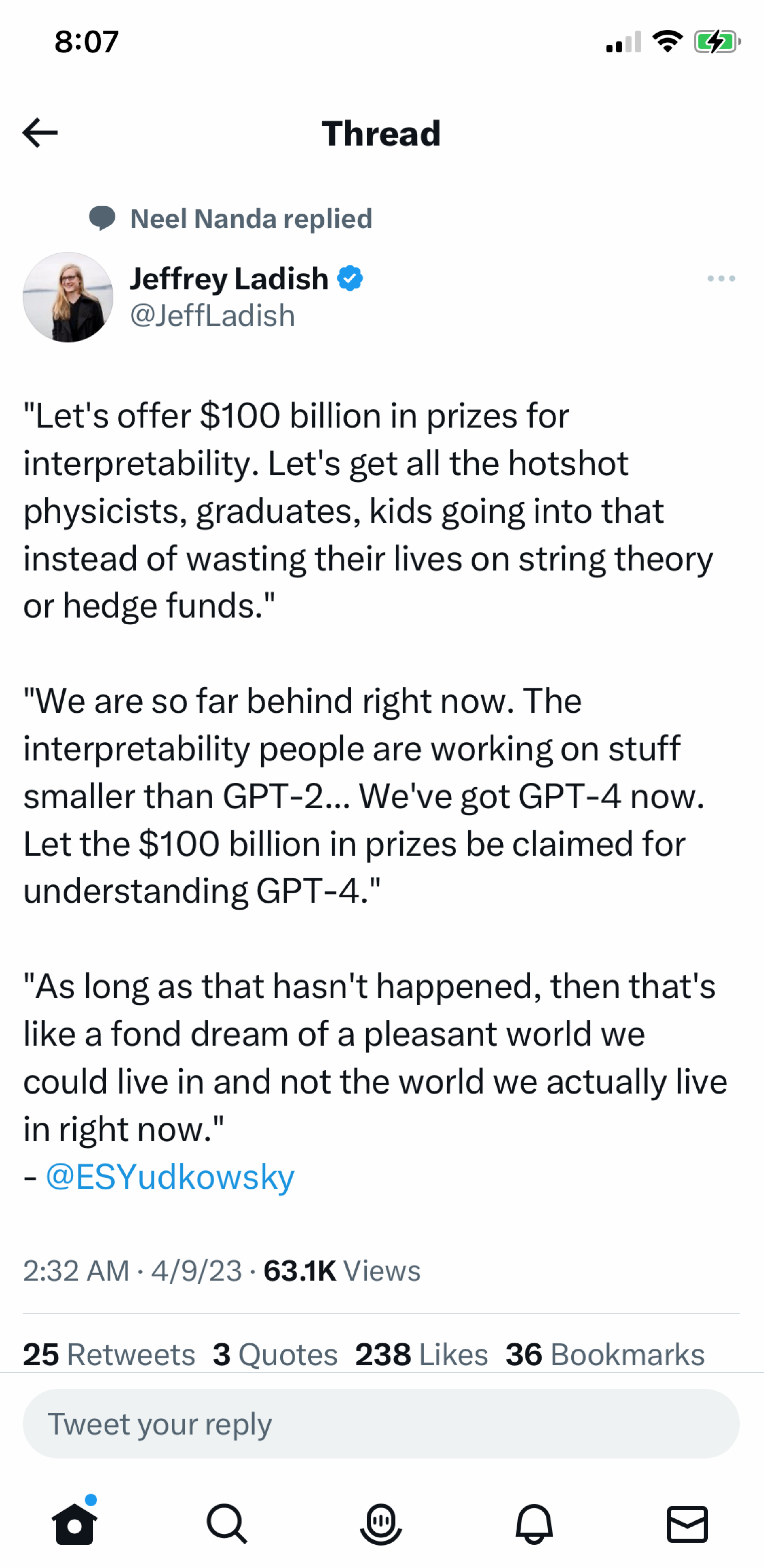
*…, then we know how to solve this one, too. By "solve", we mean a design of (p, **v**, R).*

Examples

	✓	✓	✗	?	?	✓	✗
equation	diffusion equation	Poisson equation	ideal wave equation	dissipative wave equation	Helmholtz equation	screened Poisson equation (Yukawa)	Schrödinger equation
PDE $\hat{L}\phi = 0$	$\phi_t - \nabla^2\phi = 0$	$\phi_{tt} + \nabla^2\phi = 0$	$\phi_{tt} - \nabla^2\phi = 0$	$\phi_{tt} + 2\epsilon\phi_t - \nabla^2\phi = 0$	$\phi_{tt} + \nabla^2\phi + k_0^2\phi = 0$	$\phi_{tt} + \nabla^2\phi - m^2\phi = 0$	$i\phi_t + \nabla^2\phi = 0$
Rewritten	$\frac{\partial\phi}{\partial t} + \nabla\cdot(\phi(-\nabla\log\phi)) = 0$	$\frac{\partial(-\phi_t)}{\partial t} + \nabla\cdot((-\phi_t)(\frac{\nabla\phi}{\phi_t})) = 0$	$\frac{\partial(-\phi_t)}{\partial t} + \nabla\cdot((-\phi_t)(-\frac{\nabla\phi}{\phi_t})) = 0$	$\frac{\partial(-\phi_t-2\epsilon\phi)}{\partial t} + \nabla\cdot((-\phi_t-2\epsilon\phi)(\frac{\nabla\phi}{\phi_t+2\epsilon\phi})) = 0$	$\frac{\partial(-\phi_t)}{\partial t} + \nabla\cdot((-\phi_t)(\frac{\nabla\phi}{\phi_t})) - k_0^2\phi = 0$	$\frac{\partial(-\phi_t)}{\partial t} + \nabla\cdot((-\phi_t)(\frac{\nabla\phi}{\phi_t})) + m^2\phi = 0$	$\frac{\partial\lvert\phi\rvert^2}{\partial t} + \nabla\cdot(\lvert\phi\rvert^2(2\mathrm{Im}\nabla\log\phi)) = 0$
p	ϕ	$-\phi_t$	$-\phi_t$	$-(\phi_t + 2\epsilon\phi)$	$-\phi_t$	$-\phi_t$	$\lvert\phi\rvert^2$
$\mathbf{v}$	$-\nabla\log\phi$	$\frac{\nabla\phi}{\phi_t}$	$-\frac{\nabla\phi}{\phi_t}$	$\frac{\nabla\phi}{\phi_t+2\epsilon\phi}$	$-\phi_t$	$-\phi_t$	$\lvert\phi\rvert^2$
R	0	0	0	0	$k_0^2\phi$	$-m^2\phi$	0
$G(r,t)$	$\frac{1}{(4\pi t)^{\frac{N}{2}}}\exp(-\frac{r^2}{4t})$	$\frac{1}{(t^2+r^2)^{\frac{N-1}{2}}}$	$\frac{1}{\sqrt{t^2-r^2}}\Theta(t-r)$ (2D)	$\frac{e^{-\epsilon t}\cosh(\epsilon\sqrt{t^2-r^2})}{\sqrt{t^2-r^2}}\Theta(t-r)$ (2D)	$(\frac{k_0}{\sqrt{t^2+r^2}})^{\frac{N-1}{2}} H^{(1)}_{\frac{N-1}{2}}(k_0\sqrt{t^2+r^2})$	$(\frac{m}{\sqrt{t^2+r^2}})^{\frac{N-1}{2}} K_{\frac{N-1}{2}}(m\sqrt{t^2+r^2})$	$\frac{1}{(4\pi i t)^{\frac{N}{2}}}\exp(\frac{ir^2}{4t})$
$\hat{G}(k,t)$	$\exp(-k^2 t)$	$\exp(-kt)$	$\exp(\pm ikt)$	$\exp(-\epsilon t + i\sqrt{k^2-\epsilon^2}t)$ $(k > \epsilon)$ $\exp(-(\epsilon + \sqrt{k^2-\epsilon^2})t)$ $(k \le \epsilon)$ $\hat{G}(k,t)$	$\exp(-i\sqrt{k_0^2-k^2}t)$ $(k \le k_0)$ $\exp(-\sqrt{k^2-k_0^2}t)$ $(k > k_0)$	$\exp(-\sqrt{k^2+m^2}t)$	$\exp(ik^2 t)$
(C1)	Yes	Yes	No	Conditionally yes	Conditional yes	Yes	No
(C2)	Yes	Yes	No	Conditionally yes	Conditional Yes	Yes	No
Illustration ϕ							
s-generative?	Yes (Diffusion Models)	Yes (Poisson Flow)	No	Conditionally Yes (large ϵ)	Conditionally yes (small k)	Yes	No

My quest - **Self-Rescue:** AI interpretability

(Phase transitions, quantisation, modularity)



8:07

Thread

Neel Nanda replied

Jeffrey Ladish
@JeffLadish

"Let's offer $100 billion in prizes for interpretability. Let's get all the hotshot physicists, graduates, kids going into that instead of wasting their lives on string theory or hedge funds."

"We are so far behind right now. The interpretability people are working on stuff smaller than GPT-2... We've got GPT-4 now. Let the $100 billion in prizes be claimed for understanding GPT-4."

"As long as that hasn't happened, then that's like a fond dream of a pleasant world we could live in and not the world we actually live in right now."
- @ESYudkowsky

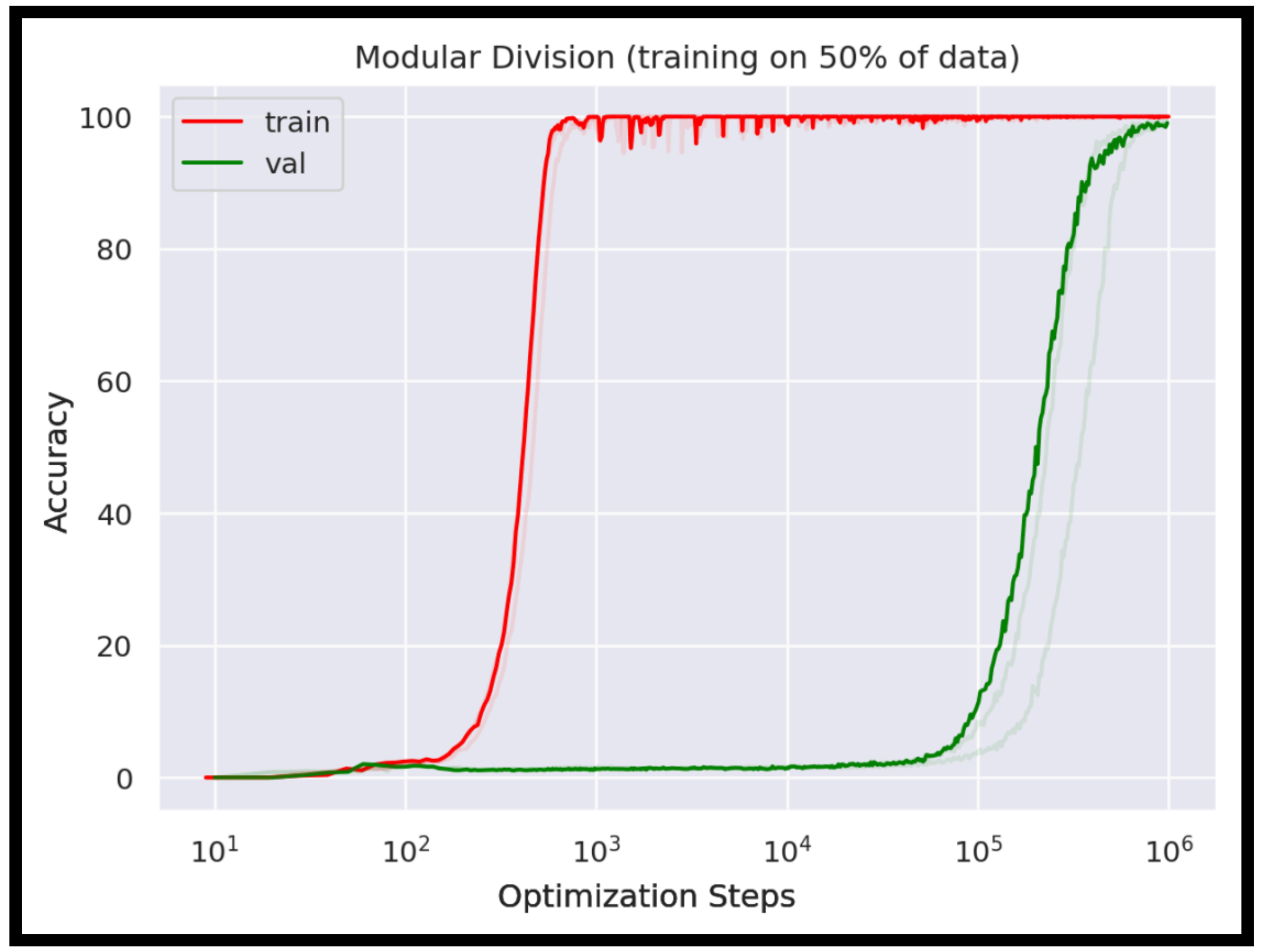
2:32 AM · 4/9/23 · **63.1K** Views

25 Retweets **3** Quotes **238** Likes **36** Bookmarks

Tweet your reply

Physics of AI

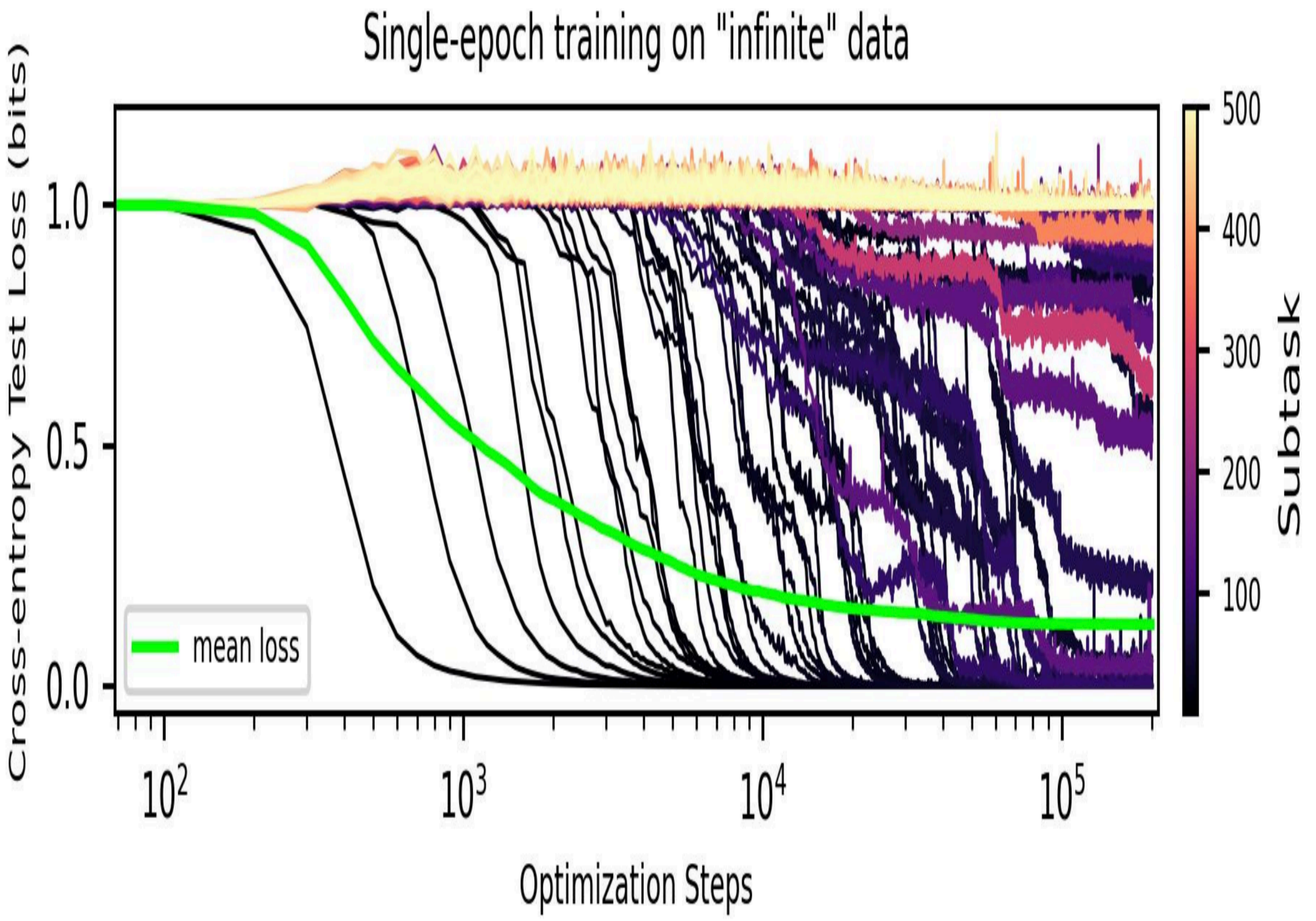
Grokking (phase transitions)



Towards understanding grokking:
An effective theory of representation learning
arXiv: 2205.10343 (NeurIPS 2022)

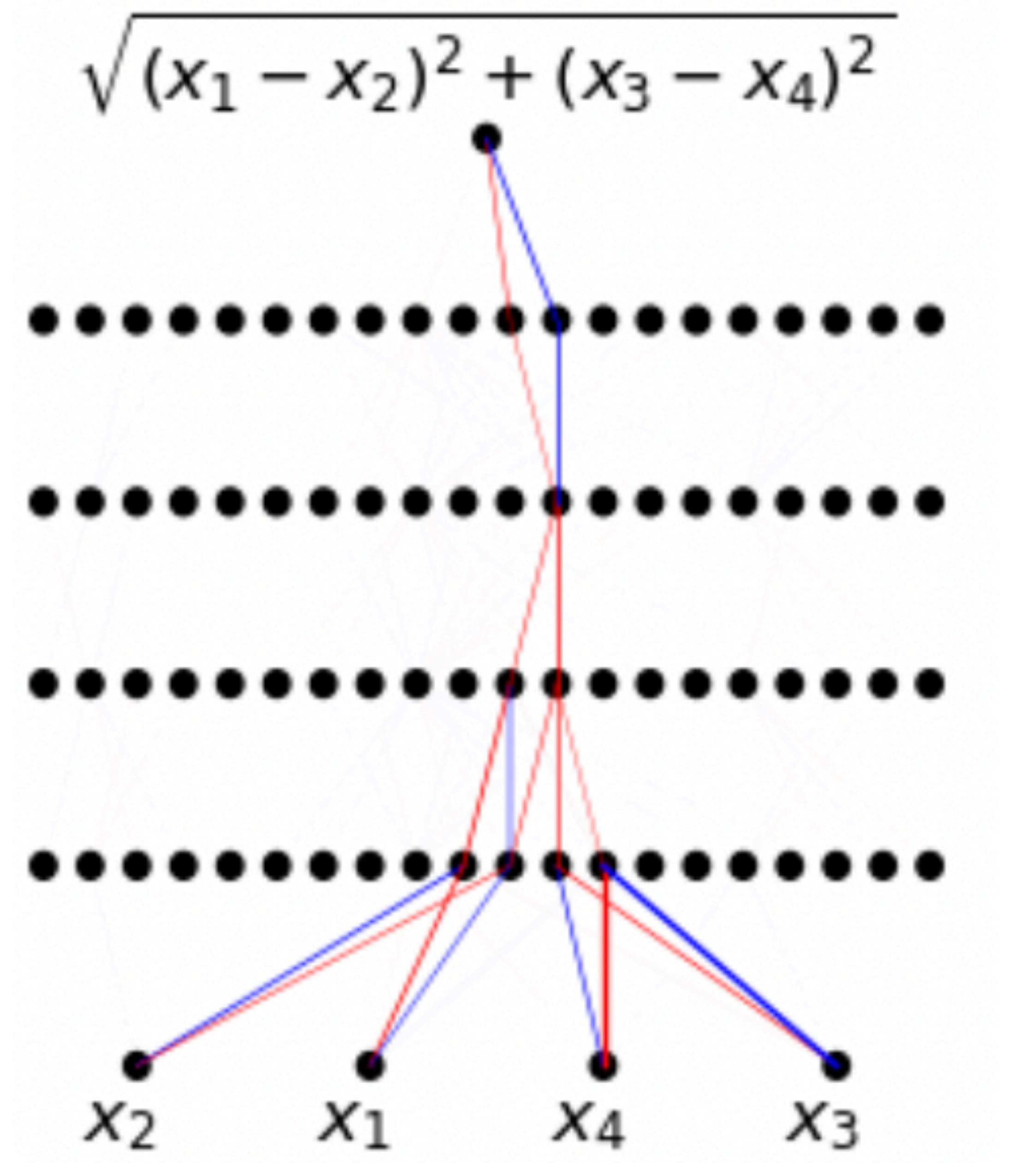
Omnigrok: Grokking Beyond Algorithmic data
arXiv: 2210.01117 (ICLR 2023)

Quantization



A Quantization Model of Neural Scaling
arXiv: 2303.13506

Locality/Modularity

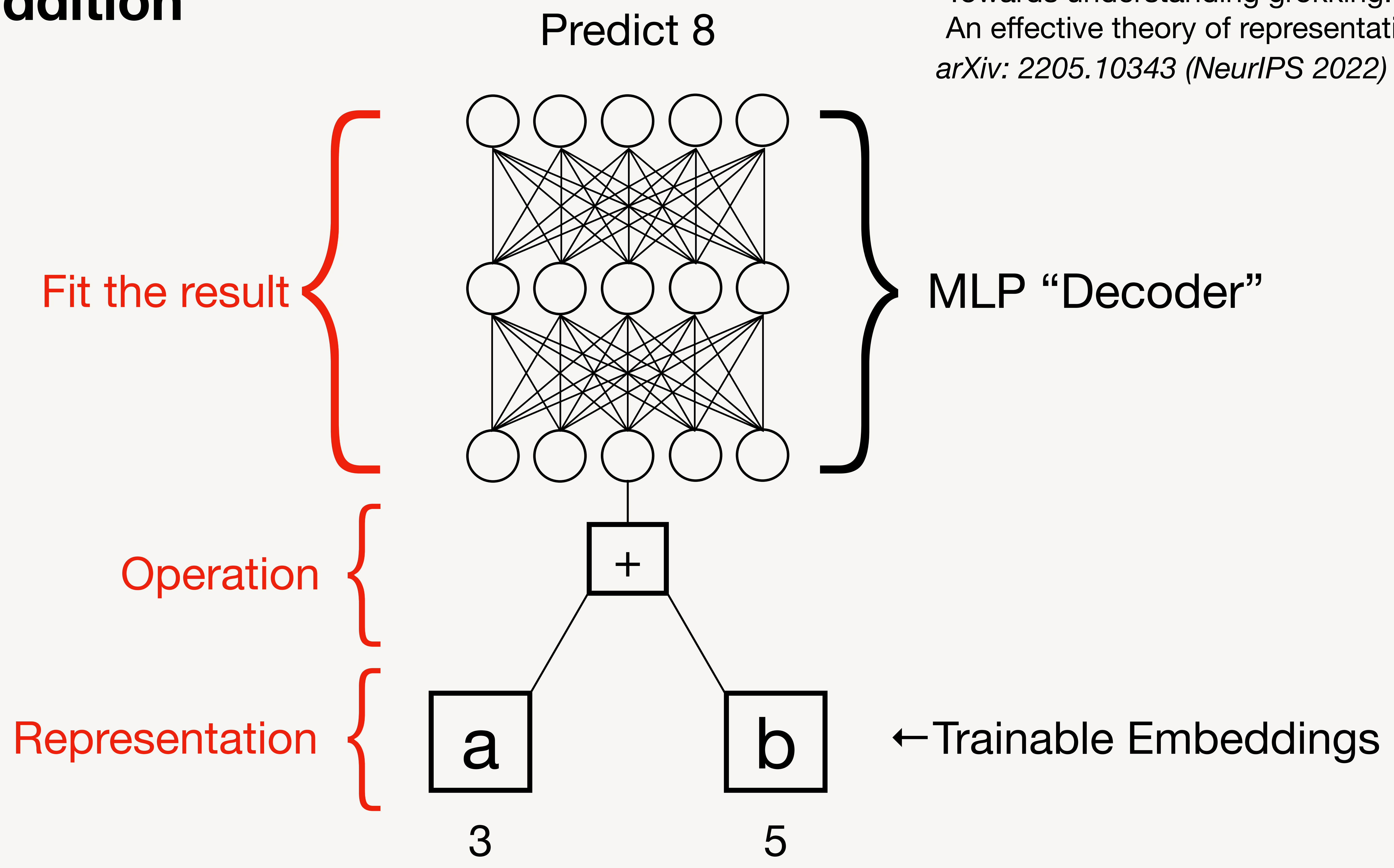


"Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability"
https://arxiv.org/abs/2305.08746

Learning addition

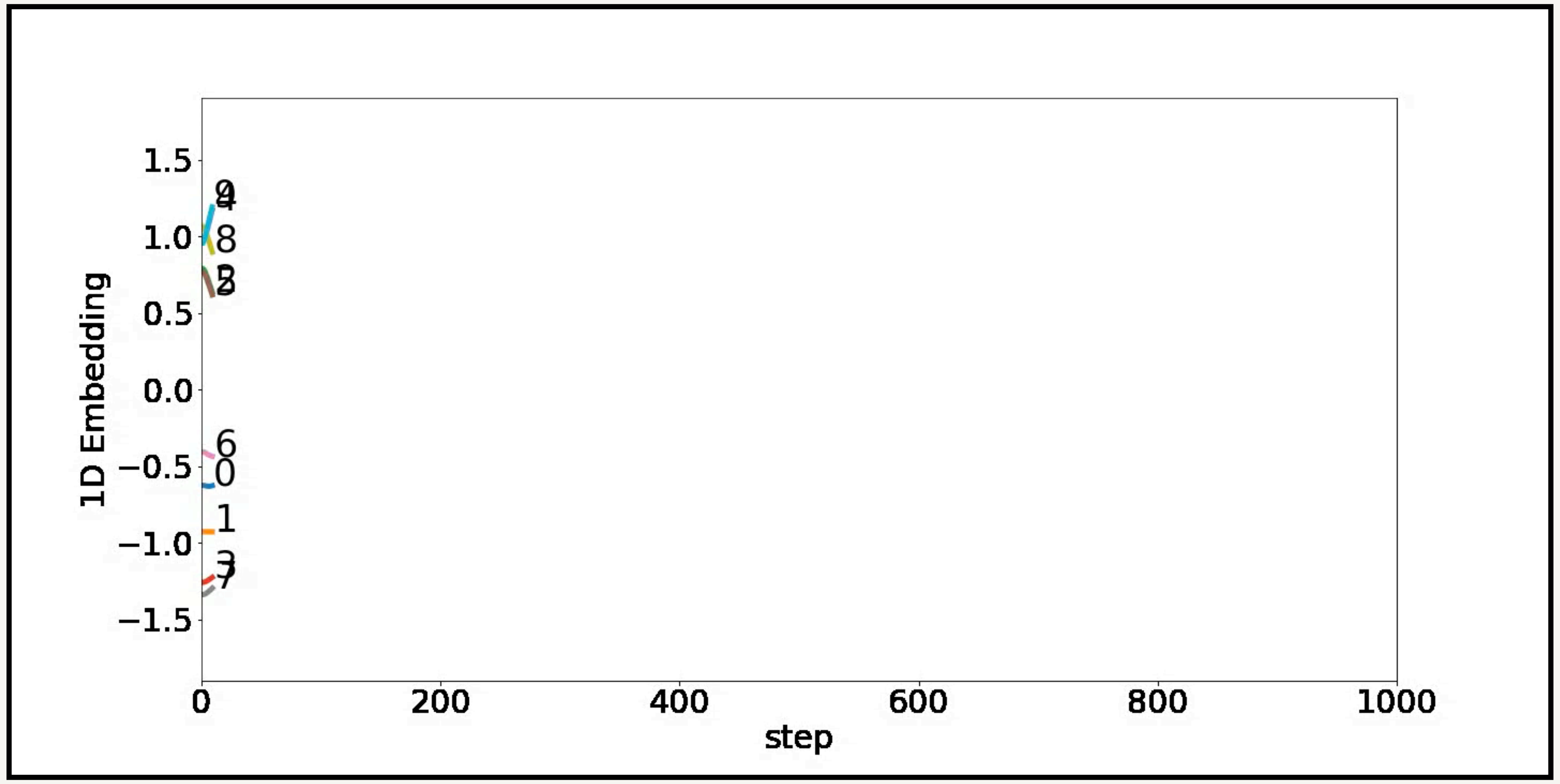
Towards understanding grokking:
An effective theory of representation learning
arXiv: 2205.10343 (NeurIPS 2022)

Addition dataset

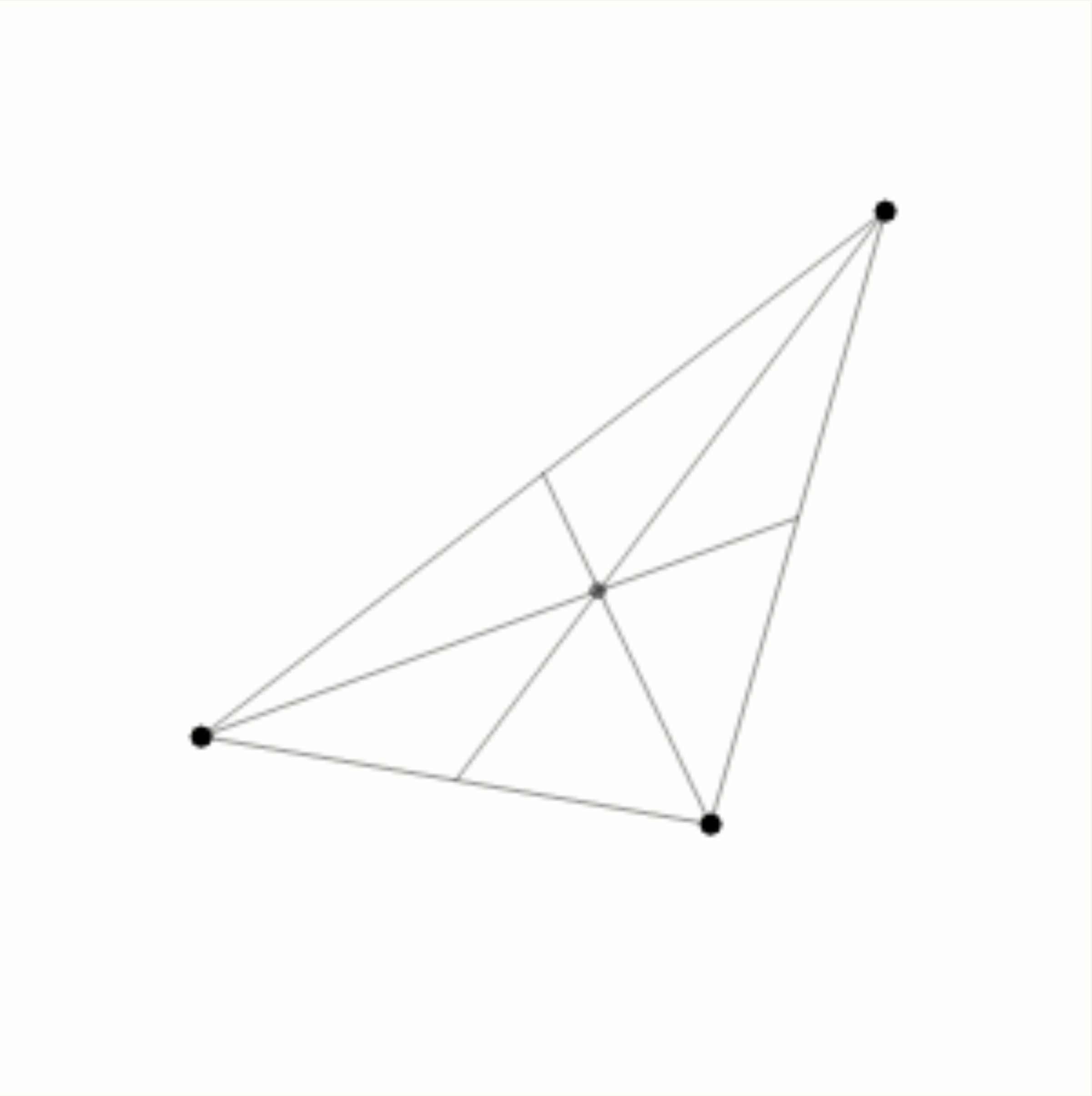


The dynamics of representation

Towards understanding grokking:
An effective theory of representation learning
arXiv: 2205.10343 (NeurIPS 2022)

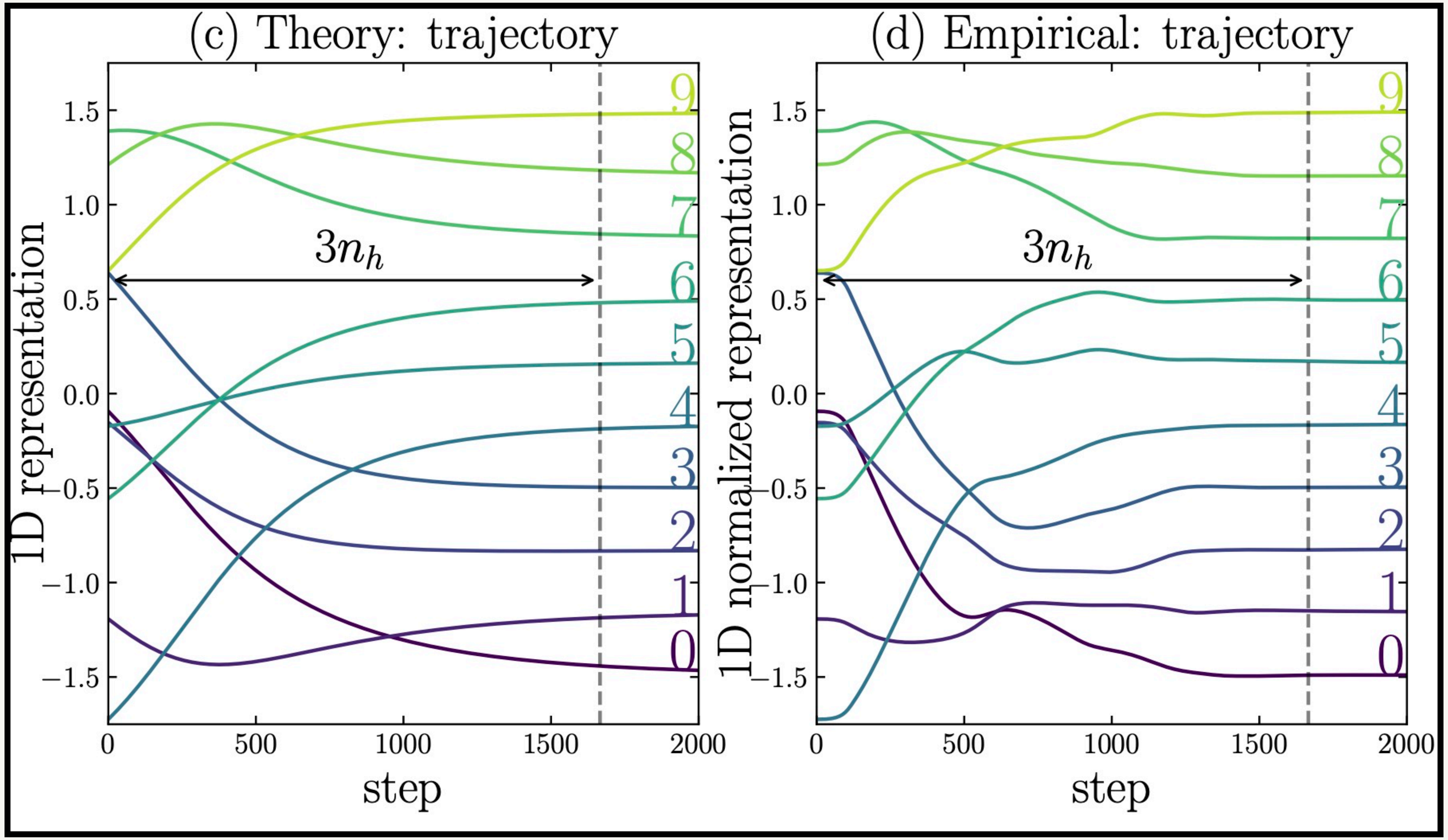


ML ⟷ Physics

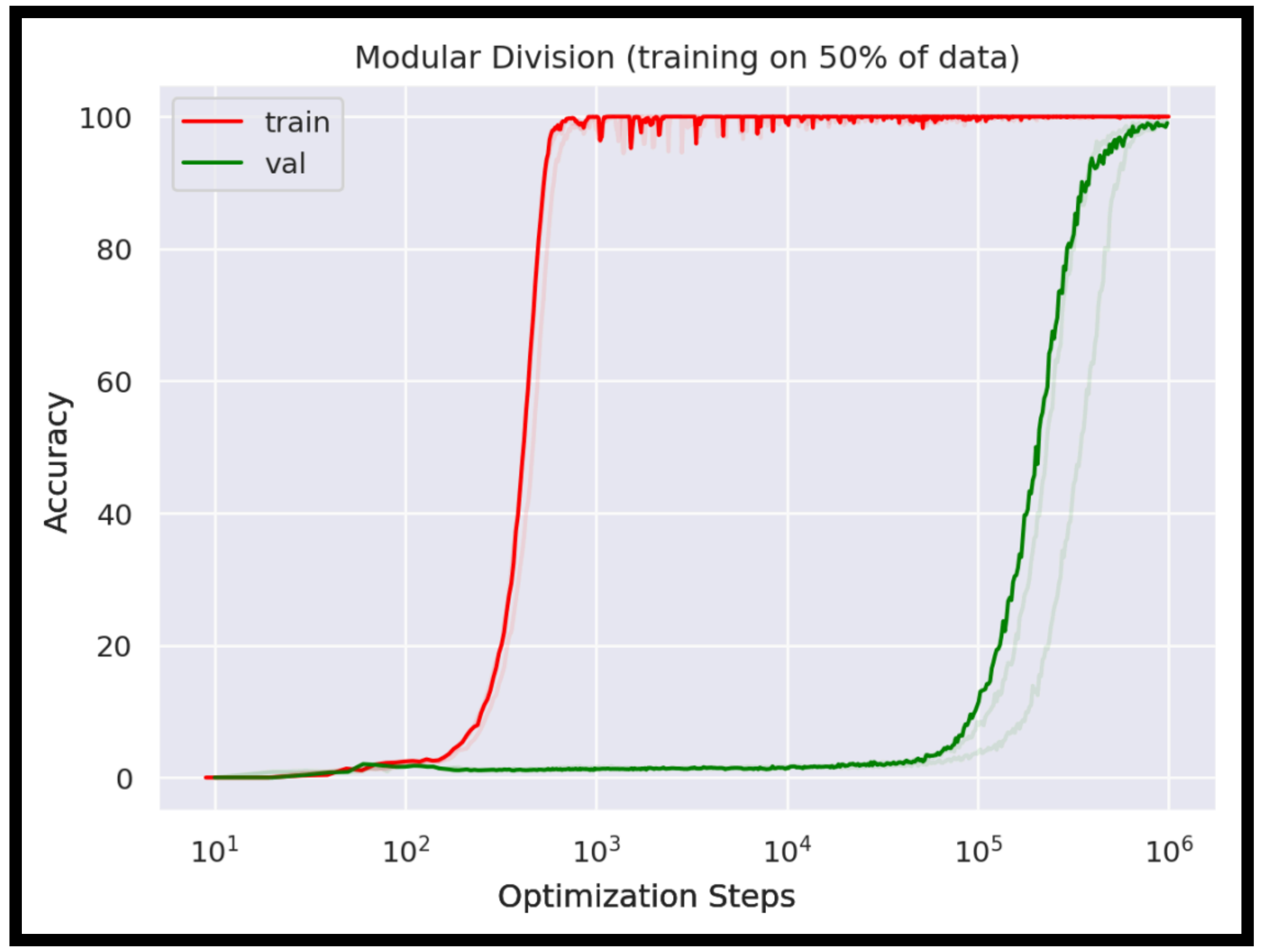


Compare theory and experiment

Towards understanding grokking:
An effective theory of representation learning
arXiv: 2205.10343 (NeurIPS 2022)

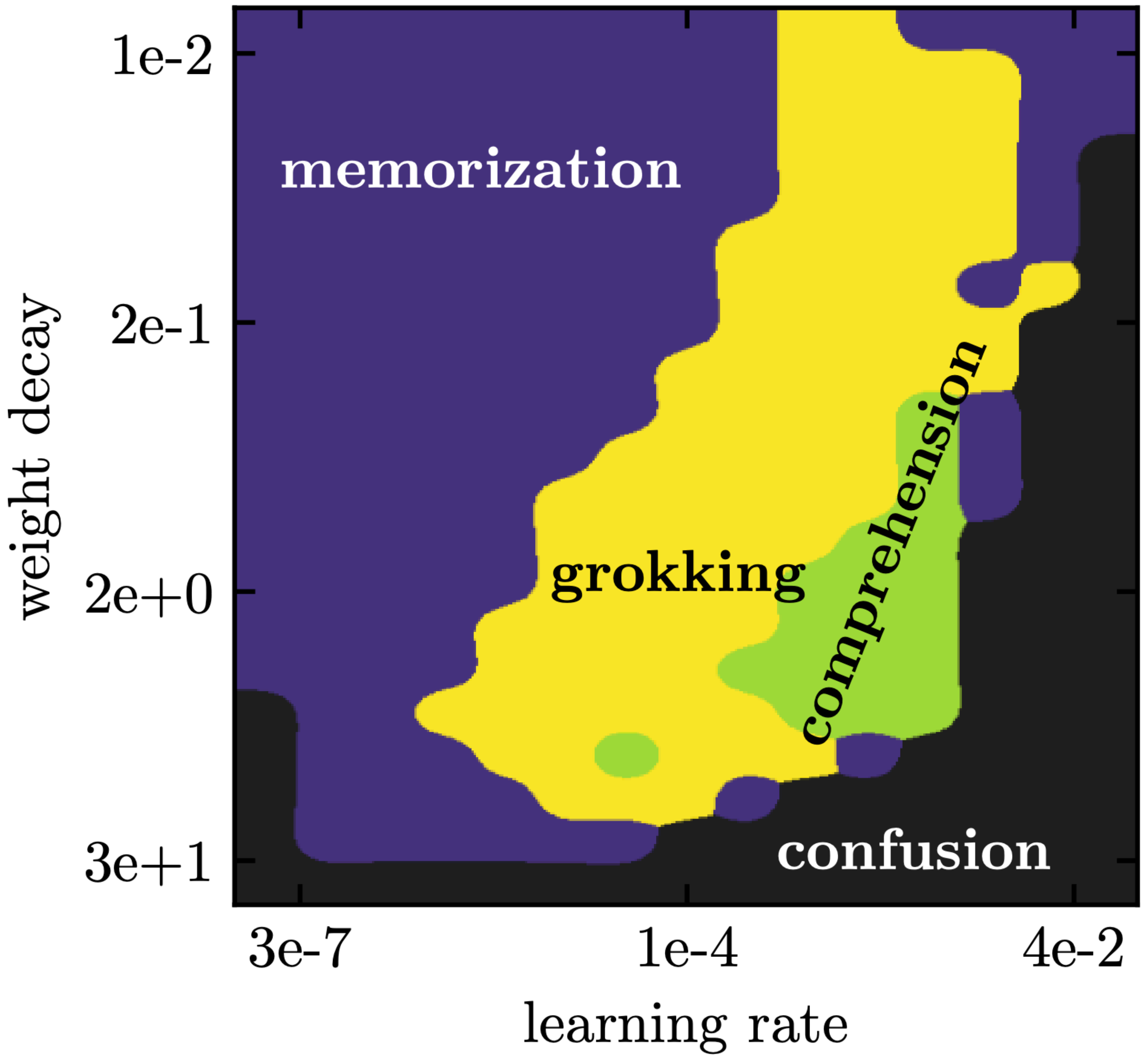


Grokking



Macroscopic

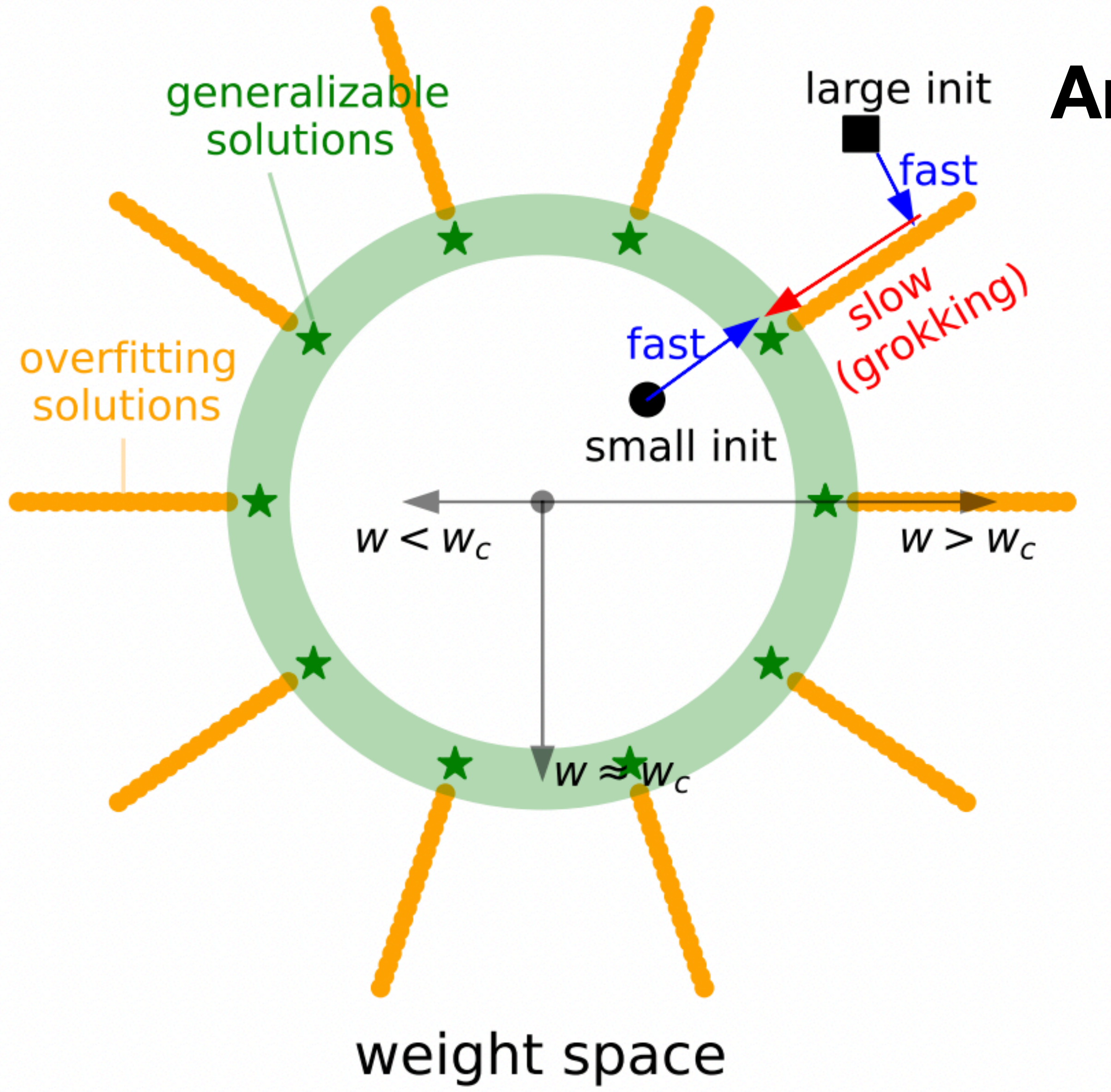
Microscopic



Phase diagrams of learning

Towards understanding grokking: An effective theory of representation learning

arXiv: 2205.10343 (NeurIPS 2022)



An adiabatic theory of learning dynamics

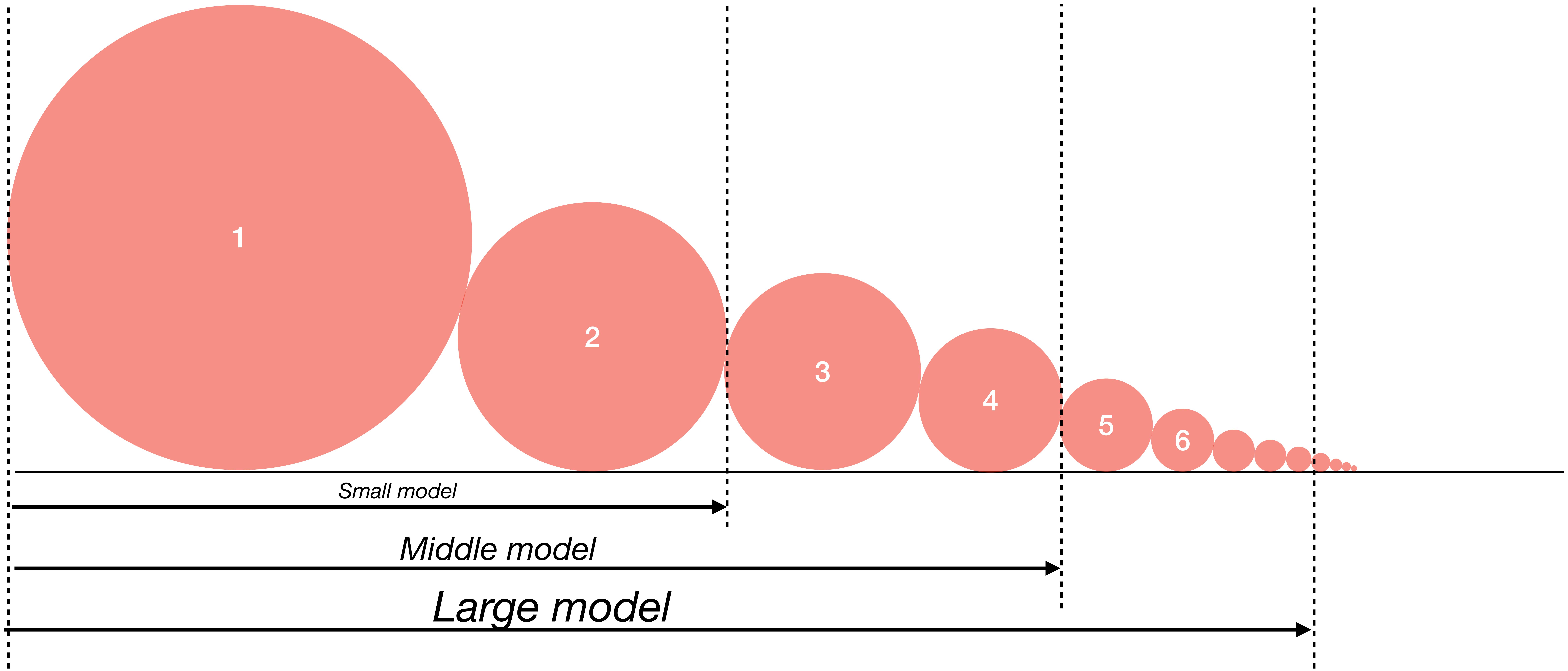
Omnigrok: Grokking Beyond Algorithmic data

arXiv: 2210.01117 (ICLR 2023)

Quantization of neural networks

The quantisation model of neural scaling, arXiv: 2303.13506

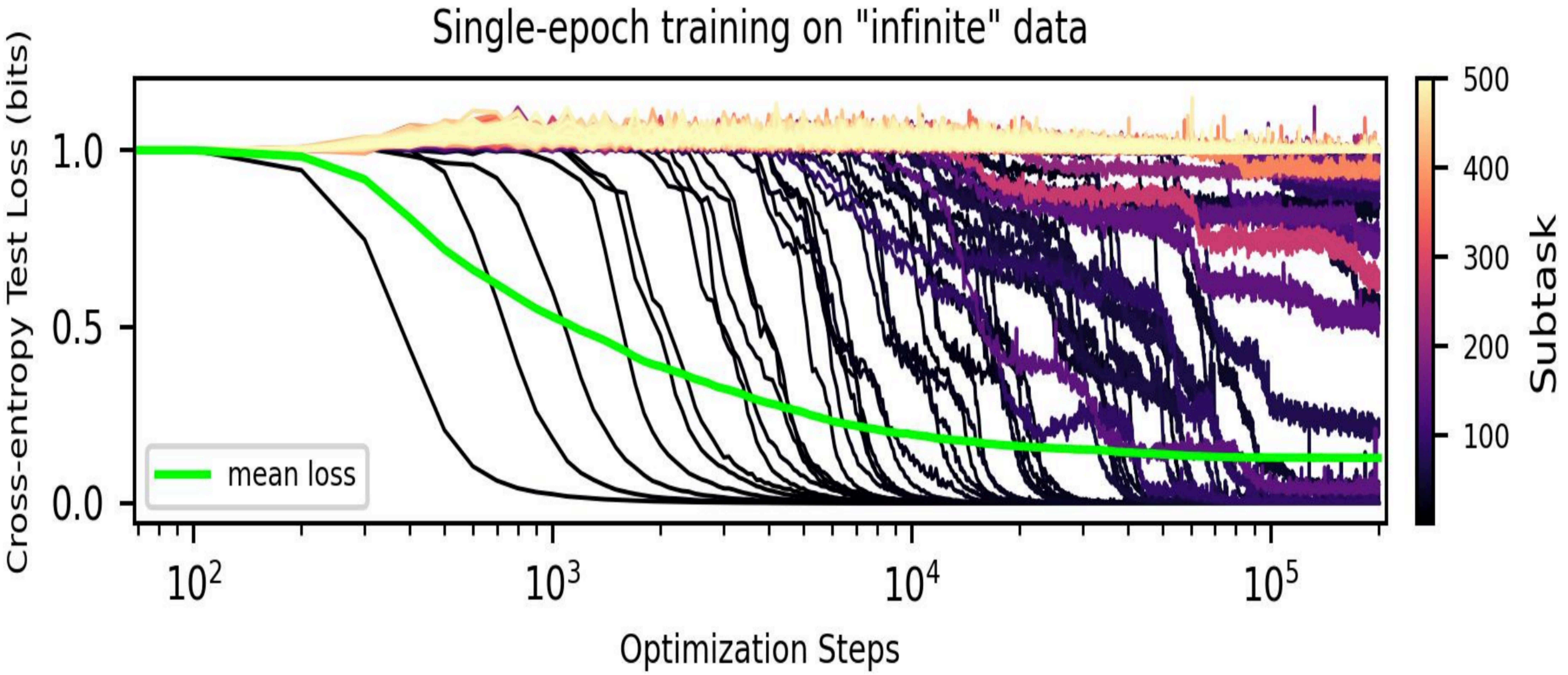
Knowledge quanta sequence



Quantization of neural networks

A Quantization Model of Neural Scaling

arXiv: 2303.13506



Quantization of neural networks

A Quantization Model of Neural Scaling

arXiv: 2303.13506

"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
```

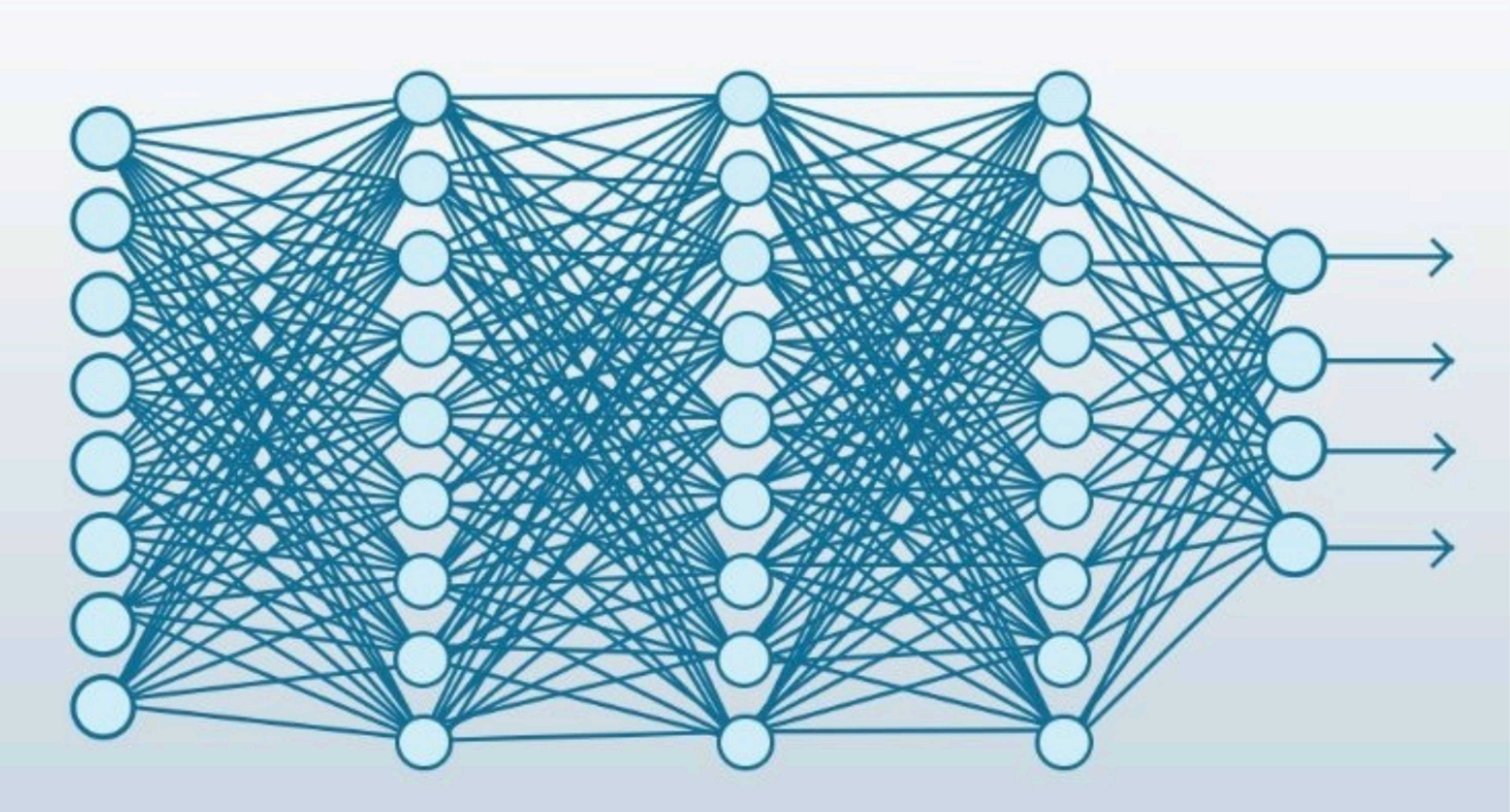
```
...
files (the
// "Software"), to deal in the Software without restriction, including
// without limitation the rights to use, copy, modify, merge, publish,
// distribute, sublicense, and/or sell copies of the Software, and to
permit
// persons to whom the Software is furnished to do so, subject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

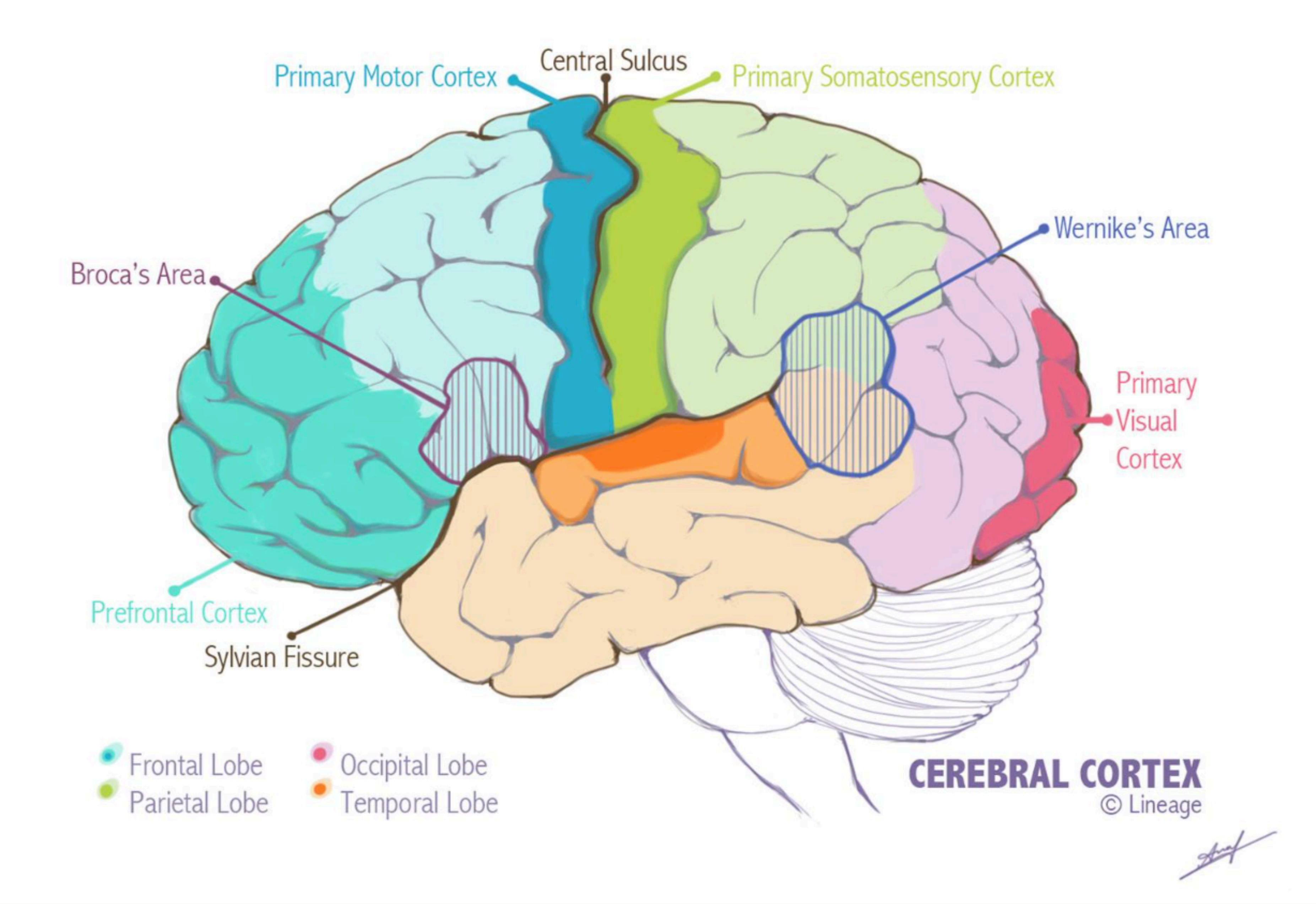
Neural networks vs brains

Neural networks



???

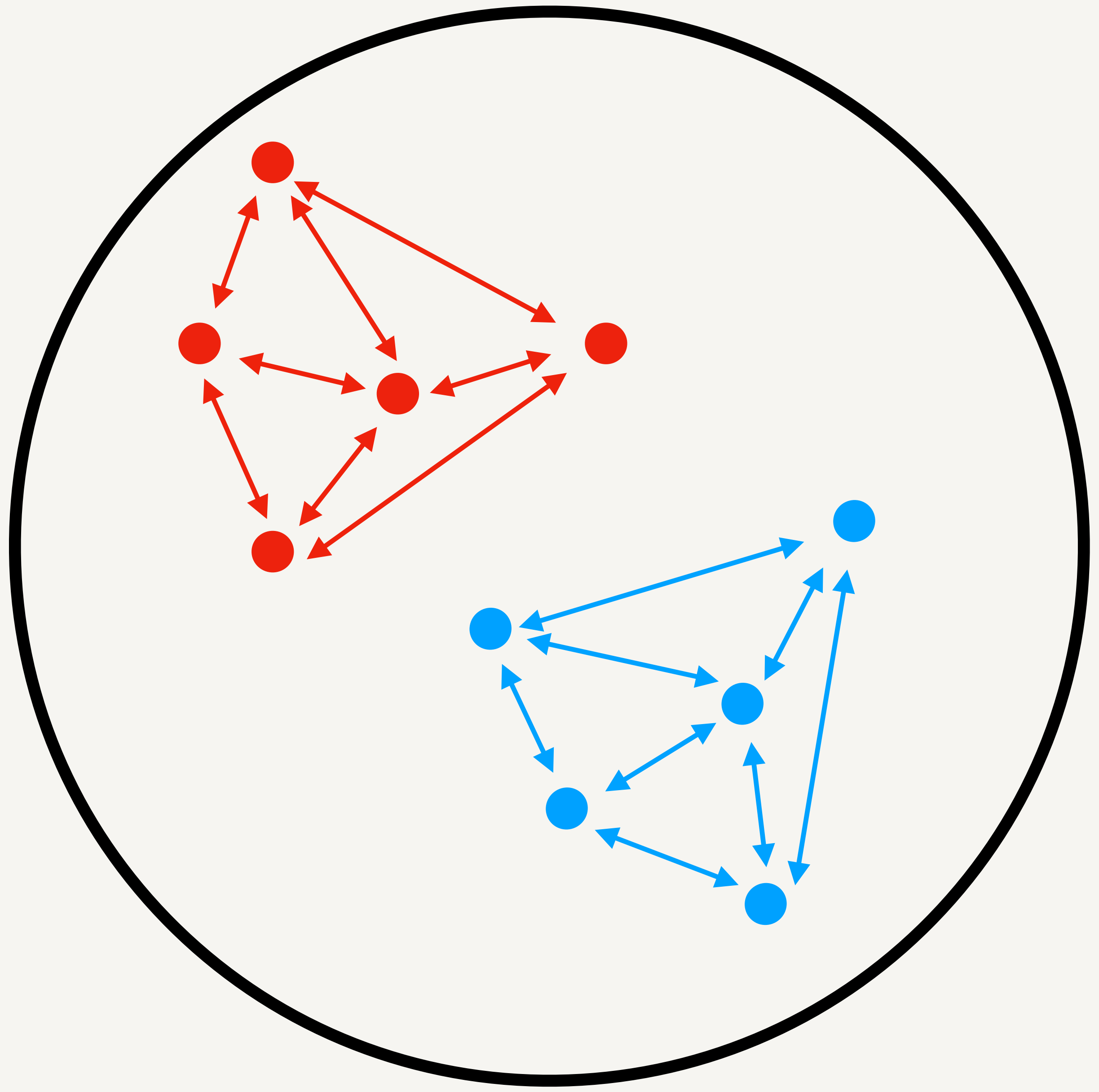
Brains



Modular brains have survival advantages, but modular NNs don't

When humans deal with a specific task …

Modular brains



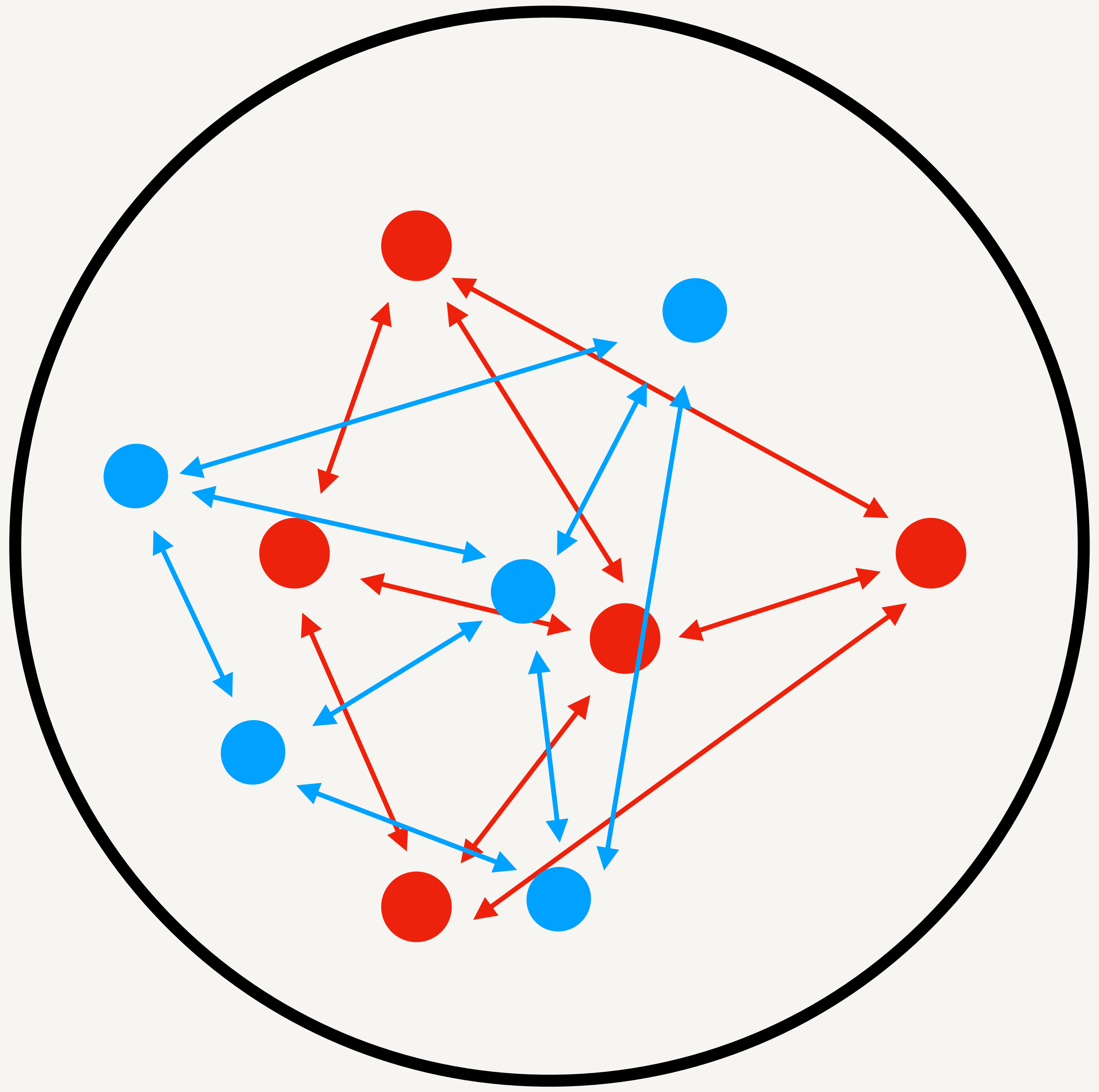
Relevant neurons are local

Shorter neuron connections

React faster

More likely to survive

Non-Modular brains



Relevant neurons are non-local

Longer neuron connections

React slower

Less likely to survive

Q: Do modular neural networks have “survival advantages”?

A: No! Because there is no (explicit) incentive for artificial neural networks to become modular if it only cares about prediction.

Q: What training techniques can induce modularity in otherwise non-modular networks?

A: Need to introduce “locality” and **limit resources (hunger)**!

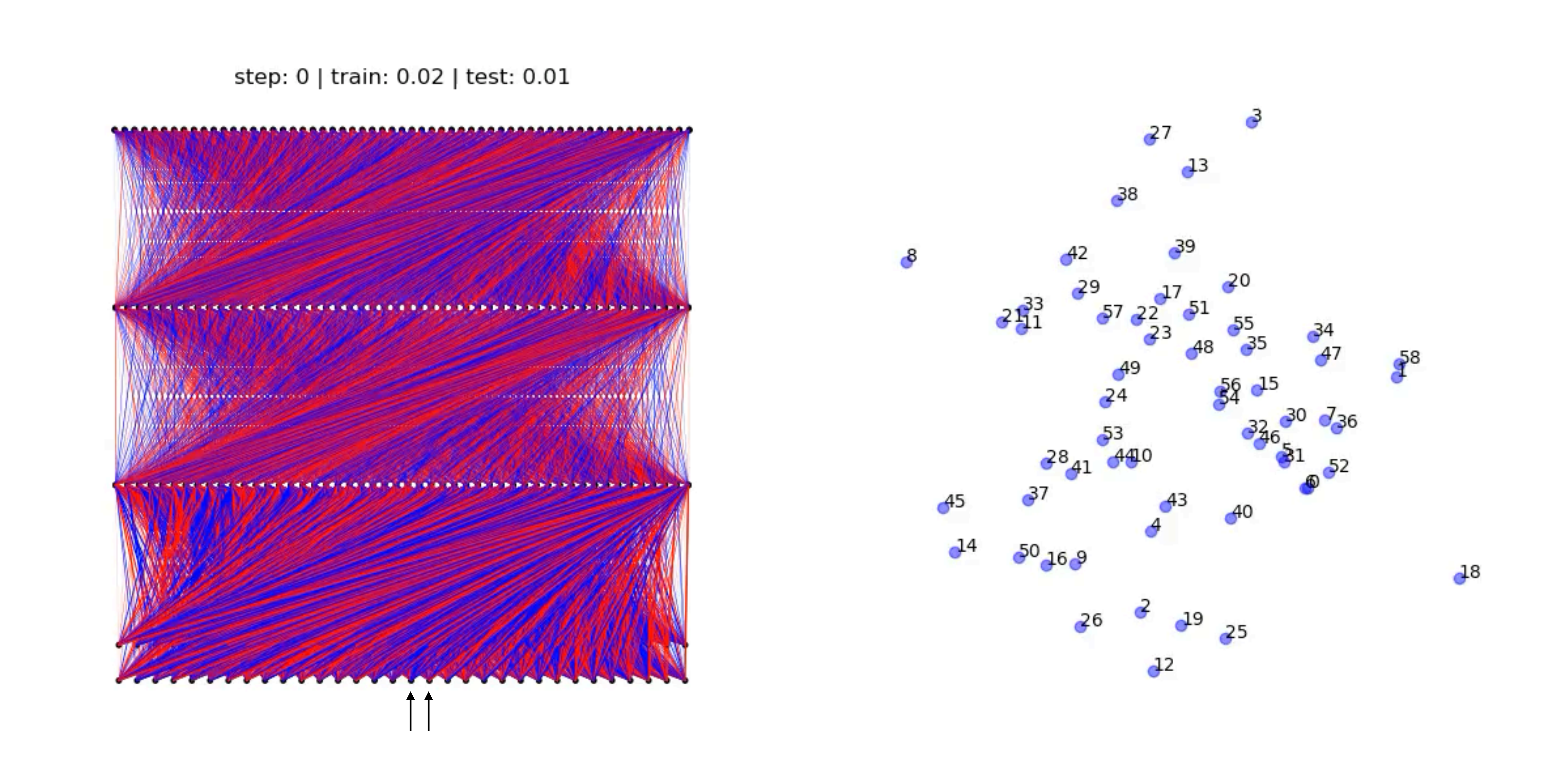
Liu, Gan & Tegmark “Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability”

Brain-inspired modular training (BIMT)

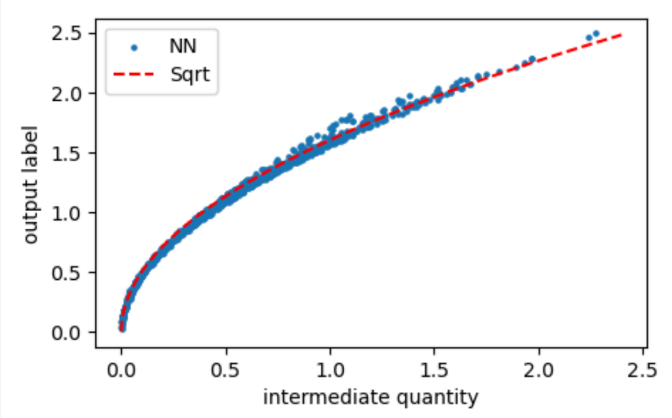
Liu, Gan & Tegmark "Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability" https://arxiv.org/abs/2305.08746

Modular addition

blue/red stands for positive/negative weights



Symbolic formulas

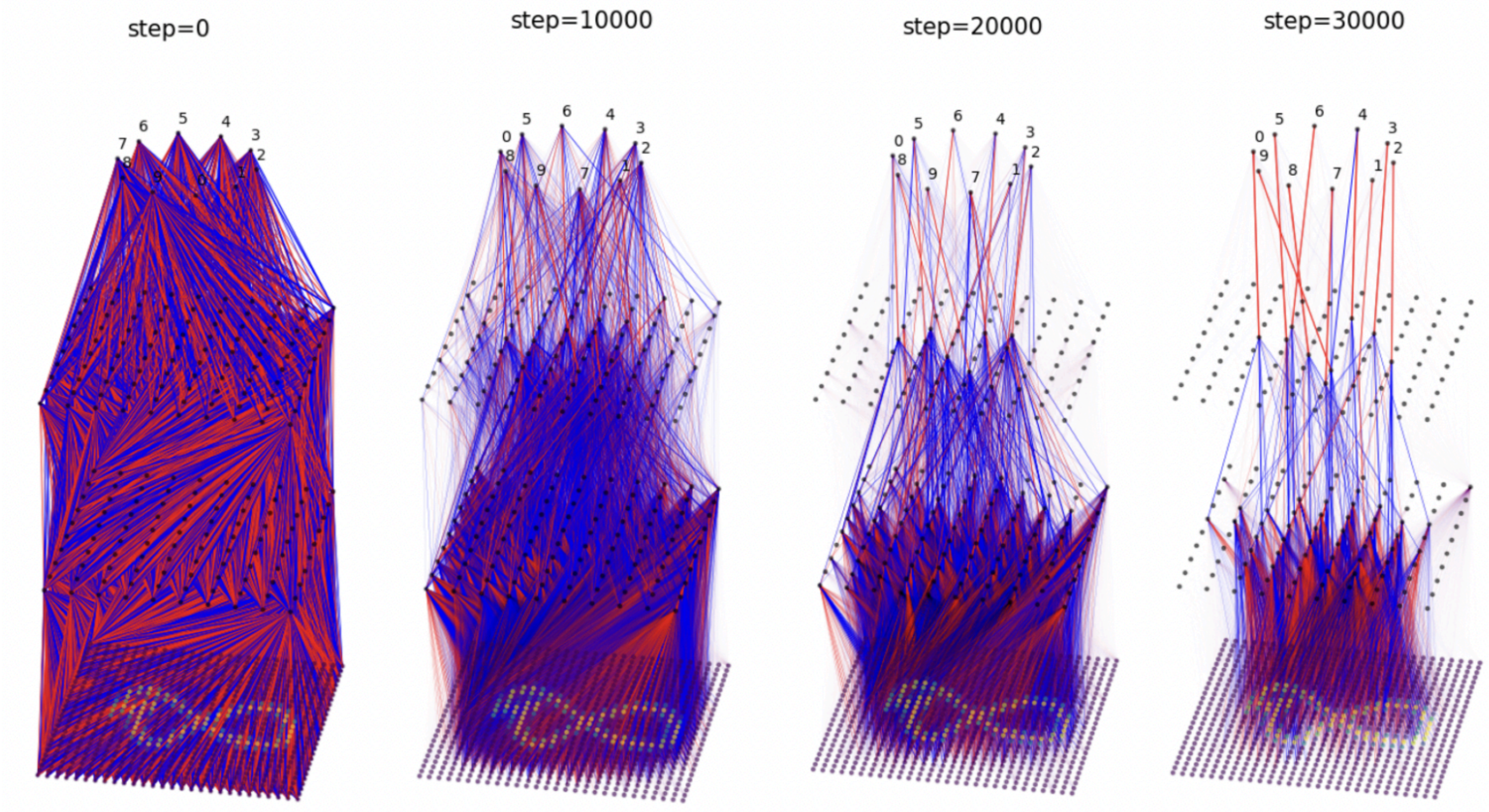


(a) independence

(b) feature sharing

(c) compositionality

Locality/Modularity



Human scientists can survive in the time of AI

- The Survivors（幸存派）： admit AI has limitations, but reject AI. Humans can work in areas where AI has limitations. Do research areas that AI cannot solve.
- The Collaborators (合作派)：admit AI has limitations, and accept AI. Only collaboration between AI and humans can lead to faster scientific development. Find research topics that neither AI nor human can solve.
- The Redemptionists (拯救派)：admit AI has limitations, and improve AI. Use scientific tools to develop better AI.
- The Self-rescuers (自救派)：concern about AI safety. Humans need to focus on AI safety so that we can better understand and finally control AI. Research on AI interpretability.
- The worshipers (崇拜派)：AI is omnipotent, and embrace AI. Humans needn't and cannot understand AI. Just take advantage of AI power.
- The Adventists (降临派)：AI is omnipotent, embrace AI and become AI. Align human values to AI, not the other way around. I don't know…

Note: Some of these beliefs are quite extreme. Most people lie on the continuous spectrum rather than extreme points.