# Predictions

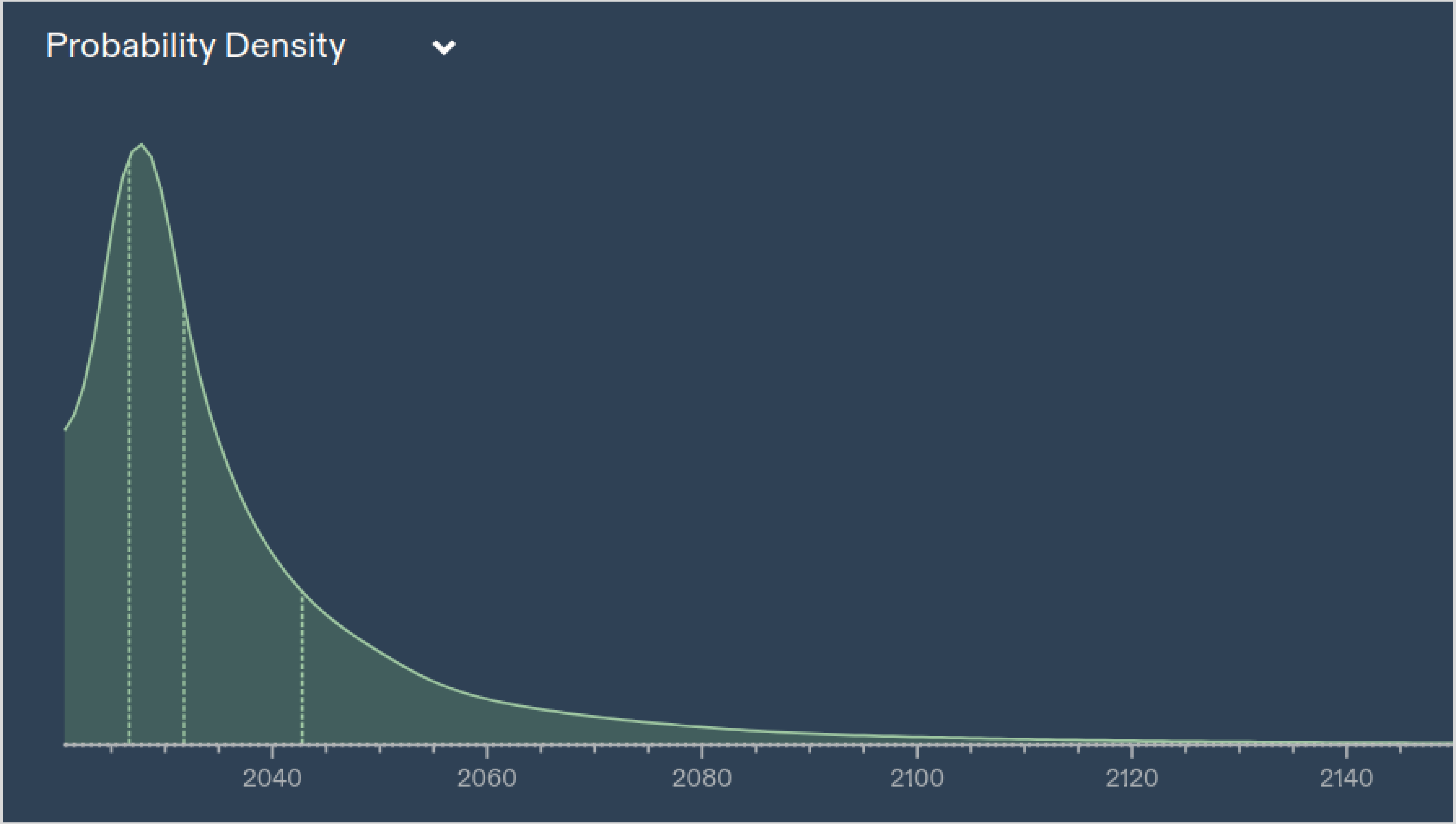
# Readings

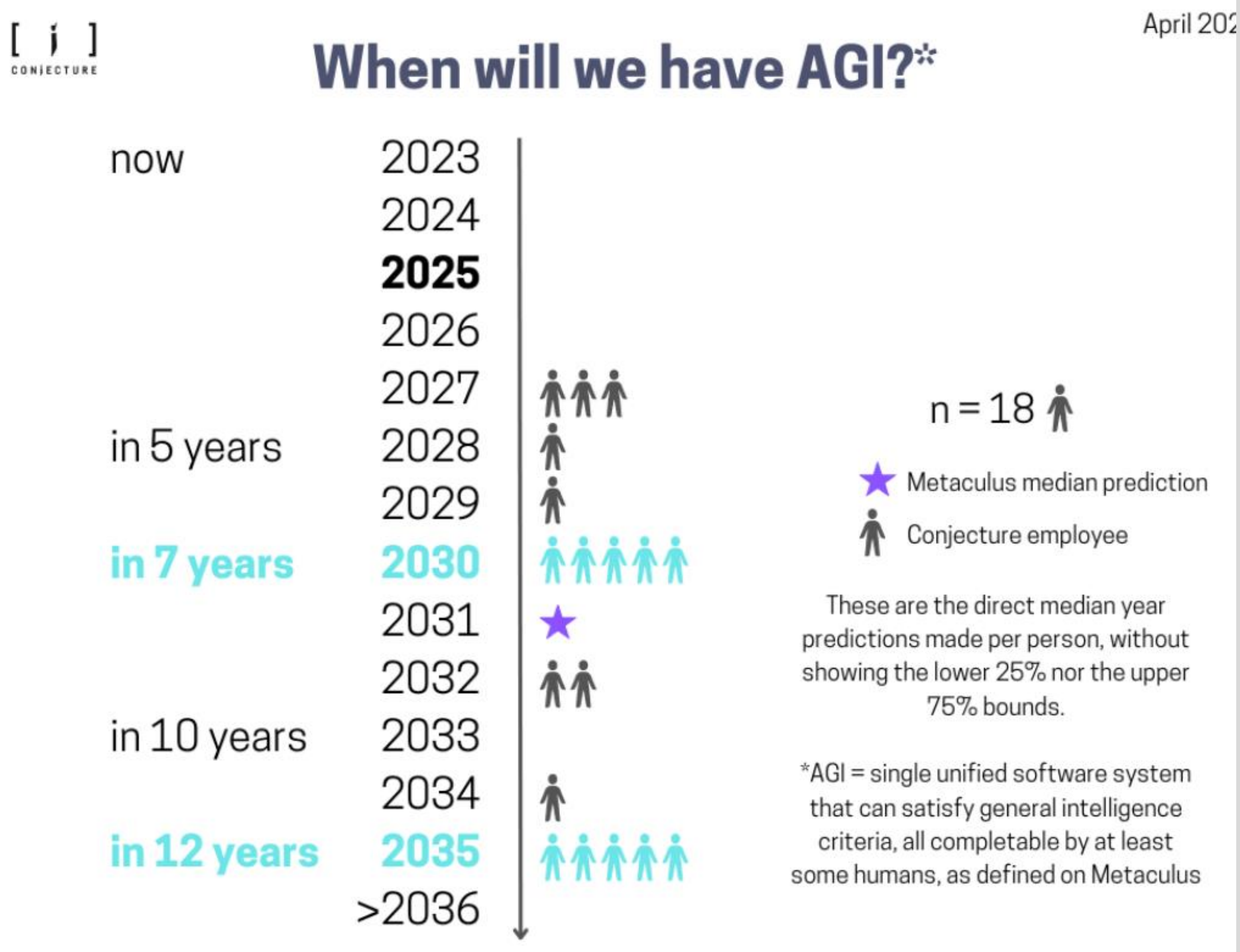
http://www.incompleteideas.net/IncIdeas/BitterLesson.html

https://www.astralcodexten.com/p/biological-anchors-a-trick-that-might

# Other predictions

- The 2022 Expert Survey on Progress in AI gives a probability of 50% of human-level AI by 2059. It was 2061 in the 2016 Survey.
- Metaculus predicts Jan 20, 2032.


Probability Density
2040
2060
2080
2100
2120
2140



CONJECTURE

April 202

# When will we have AGI?*

|  |  |  |
|---|---|---|
| now | 2023 |  |
|  | 2024 |  |
|  | **2025** |  |
|  | 2026 |  |
|  | 2027 | 🧍🧍🧍 |
| in 5 years | 2028 | 🧍 |
|  | 2029 | 🧍 |
| **in 7 years** | **2030** | 🧍🧍🧍🧍🧍 |
|  | 2031 | ★ |
|  | 2032 | 🧍🧍 |
| in 10 years | 2033 |  |
|  | 2034 | 🧍 |
| **in 12 years** | **2035** | 🧍🧍🧍🧍🧍 |
|  | >2036 |  |

n = 18 🧍

★ Metaculus median prediction

🧍 Conjecture employee

These are the direct median year predictions made per person, without showing the lower 25% nor the upper 75% bounds.

*AGI = single unified software system that can satisfy general intelligence criteria, all completable by at least some humans, as defined on Metaculus

# Discussion

- What do you think of these methods? Which ones seem more valuable to you? What would be another way to predict AI progress?
- What weight would you give to each biological anchor?
- Is the bitter lesson right? What are some counterexamples to it? Do you expect the bitter lesson to stay mostly right during the next 10 years?
- How would operationalize the definition of “human-level AI”? What kind of test would you run to know if we’re getting closer?

# Readings

- https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications

# Available data

An analysis by EpochAI concluded that:

- High-quality text data will likely be exhausted before 2026.
- Low-quality text data will be exhausted between 2030 and 2050.
- Visual data will be exhausted between 2030 and 2070.

# Available data

But there are other important considerations:

- How to filter out low-quality data (particularly complex for visual data)
- How to handle copyright and privacy issues

These problems can potentially limit the amount of available data.

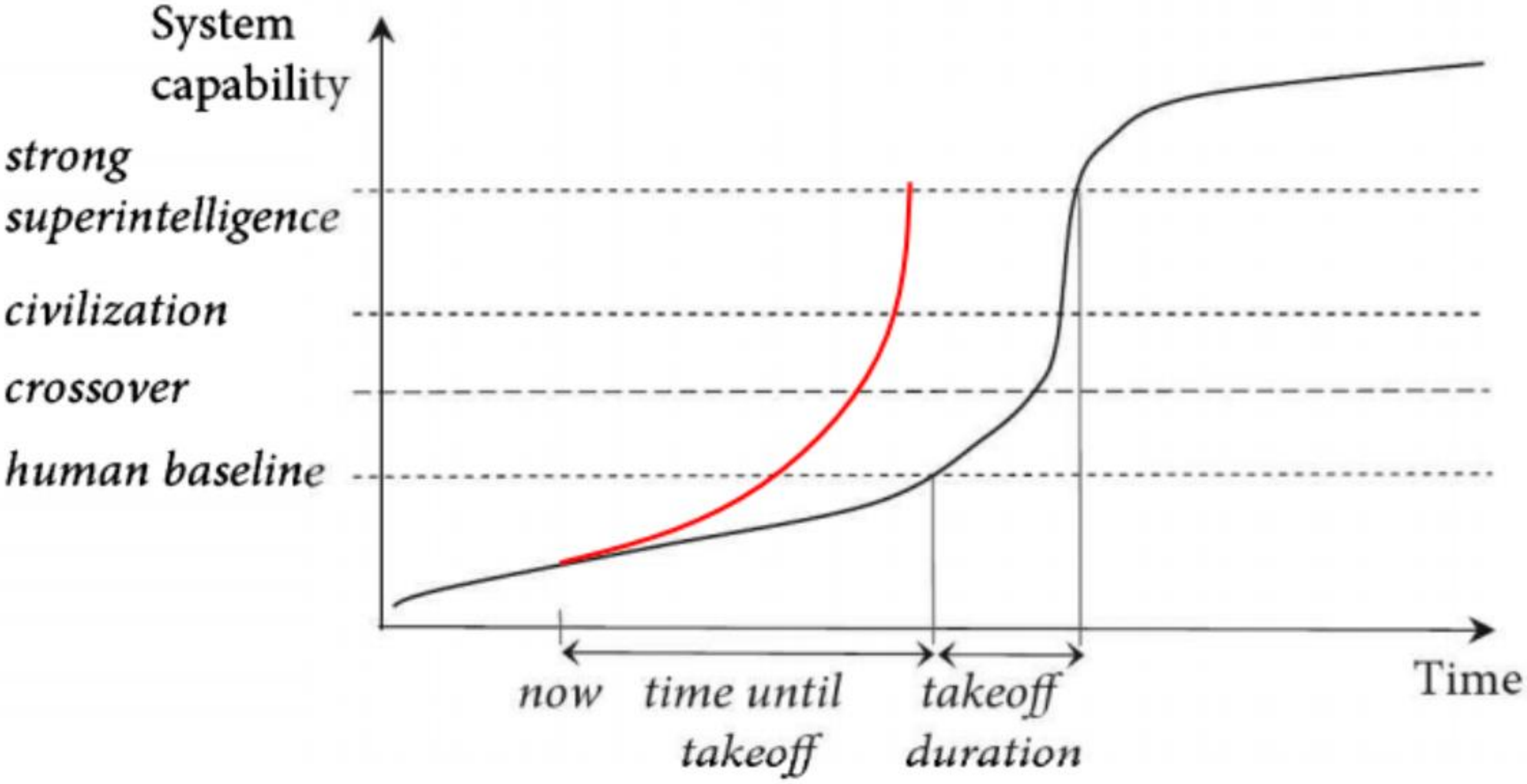
# Discussion

- What data do you expect could potentially be used for training and that are not currently used? Would that change the big picture?
- Could we use AI to generate and alter human data to have more training data and keep an exponential growth?
- Humans need way less data than AIs to learn. A big difference is that we are embodied and can interact with the world, we can test our hypotheses in an efficient way. Do you expect that embodied AIs could solve the data bottleneck?

# AI Takeoff speed

AI Takeoff refers to the process of going from an AGI to a superintelligence.

There are a lot of disagreements on the order of magnitude of the duration of this process (either fast “hard” takeoff, or slow “soft” takeoff).


System capability
strong superintelligence
civilization
crossover
human baseline
now
time until takeoff
takeoff duration
Time

# AI Takeoff speed

General arguments

- Not a big difference between the brain size of a chimp and the brain size of a human
- there would be a “core” to intelligence
- Not a big difference between the brain size of the most intelligent person and an average person
- Recursive self-improvements
- In the other direction, scaling laws

# AI Takeoff speed

The distinction is quite important

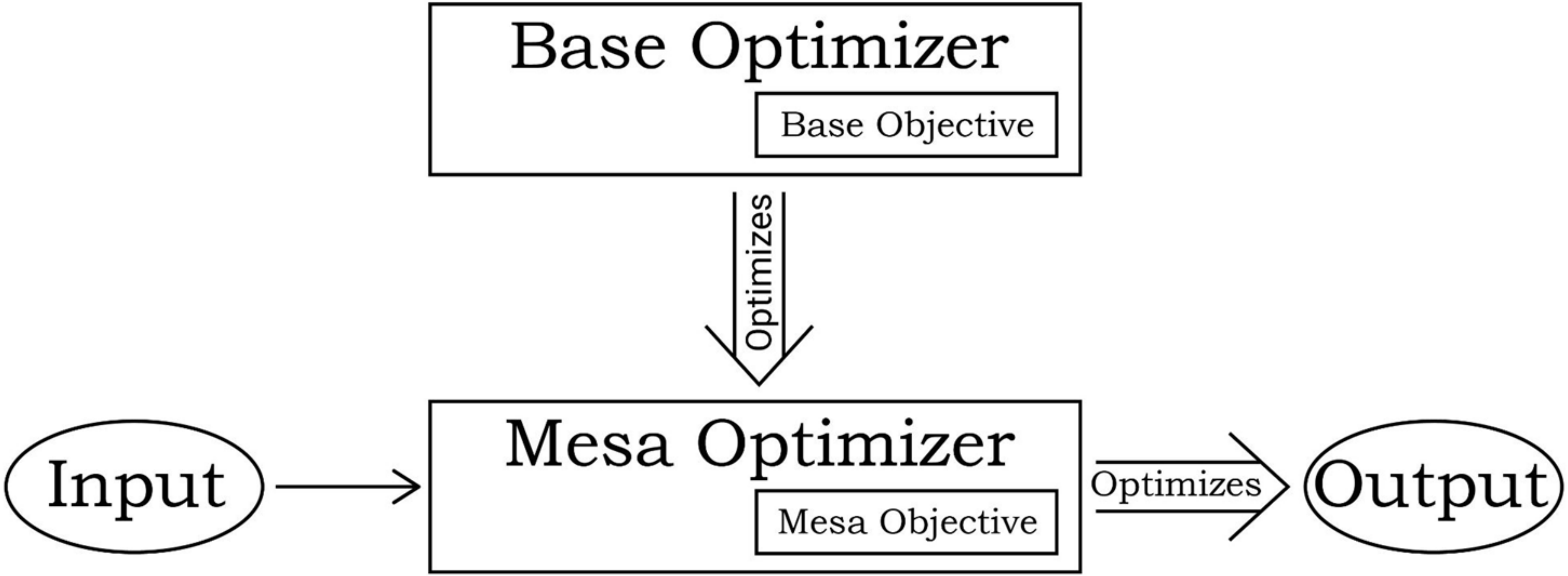
- A fast takeoff speed is likely to imply a unique AI taking over the world quickly.
- A slow takeoff is likely to imply a lot of different AIs being used in a lot of domains, and humans being replaced gradually over a few years.

A fast takeoff is generally considered to be very dangerous, and should be avoided.

# Sharp left turn and mesa-optimizers

The sharp left turn is a training step when we have a sudden large generalization in capabilities, without a corresponding generalization in the alignment properties.

One possible story is that we would get a mesa-optimizer: the model would begin to optimize for a certain goal, become an optimizer, and this would prompt it to become much more capable and intelligent.


Base Optimizer
Base Objective
Optimizes
Input
Mesa Optimizer
Mesa Objective
Optimizes
Output

# Reading

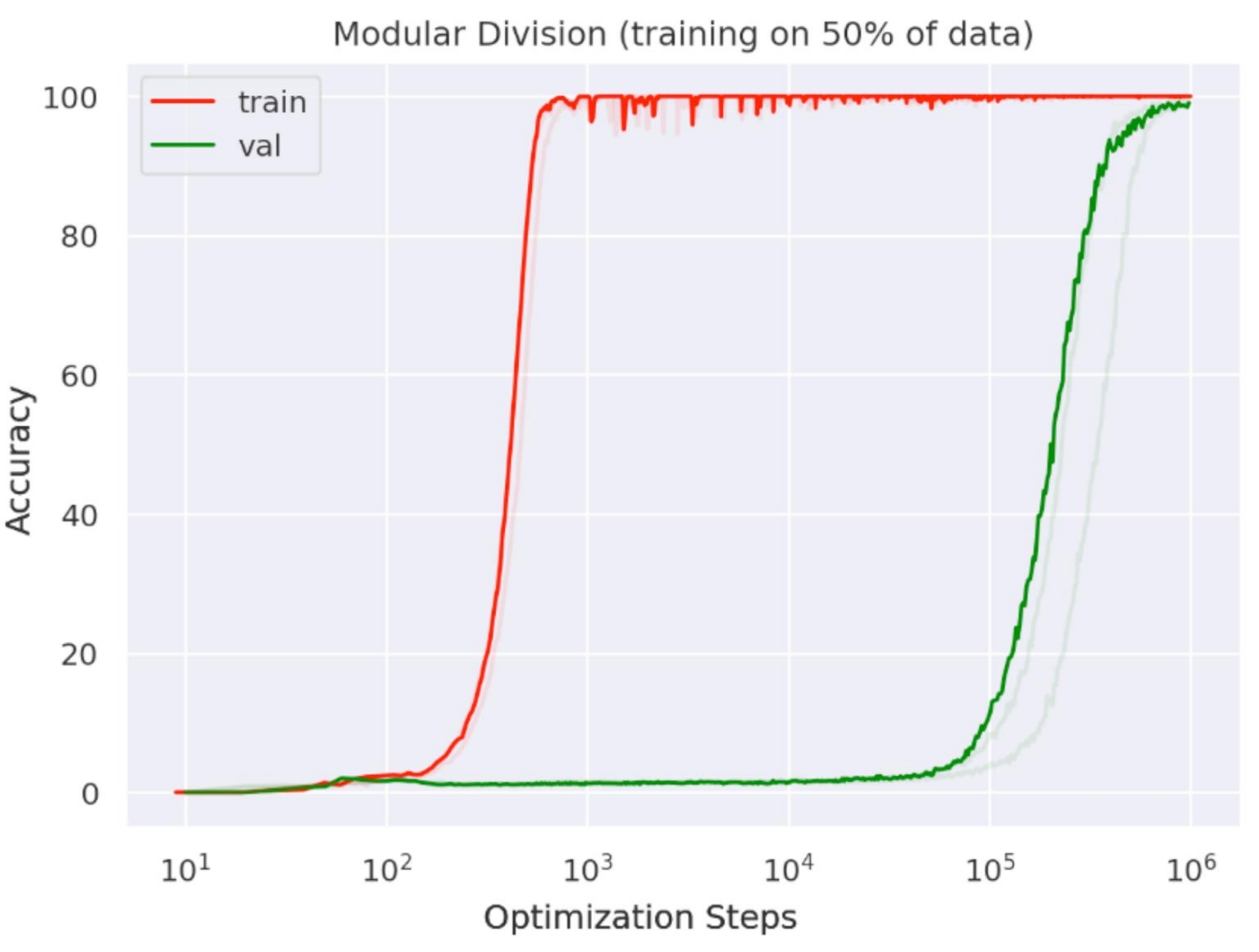
https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

# Grokking

Grokking refers to phenomenon when a model generalizes and obtains a good score on the test data way after getting an almost perfect score on the training data.

A sufficiently big neural network can learn and get an almost perfect score on a random data and so it is not surprising that it can first learn a non-general algorithm.

But thankfully we have regularization: even though it already has a perfect score, it is biased and will converge towards simpler solutions.



Modular Division (training on 50% of data)
train
val
100
80
60
40
20
0
Accuracy
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
$10^6$
Optimization Steps

# Grokking

It is sometimes argued that grokking is essential to understanding the internals of big neural networks, because it is connected to phase changes.

However, I don’t really see the point to talk about grokking instead of generalization and phase changes in general.

# Sources

- https://www.alignmentforum.org/posts/GpSzShaaf8po4rcmA/qapr-5-grokking-is-maybe-not-that-big-a-deal
- Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning
- Refining the sharp left turn
- Will AI undergo discontinuous progress?