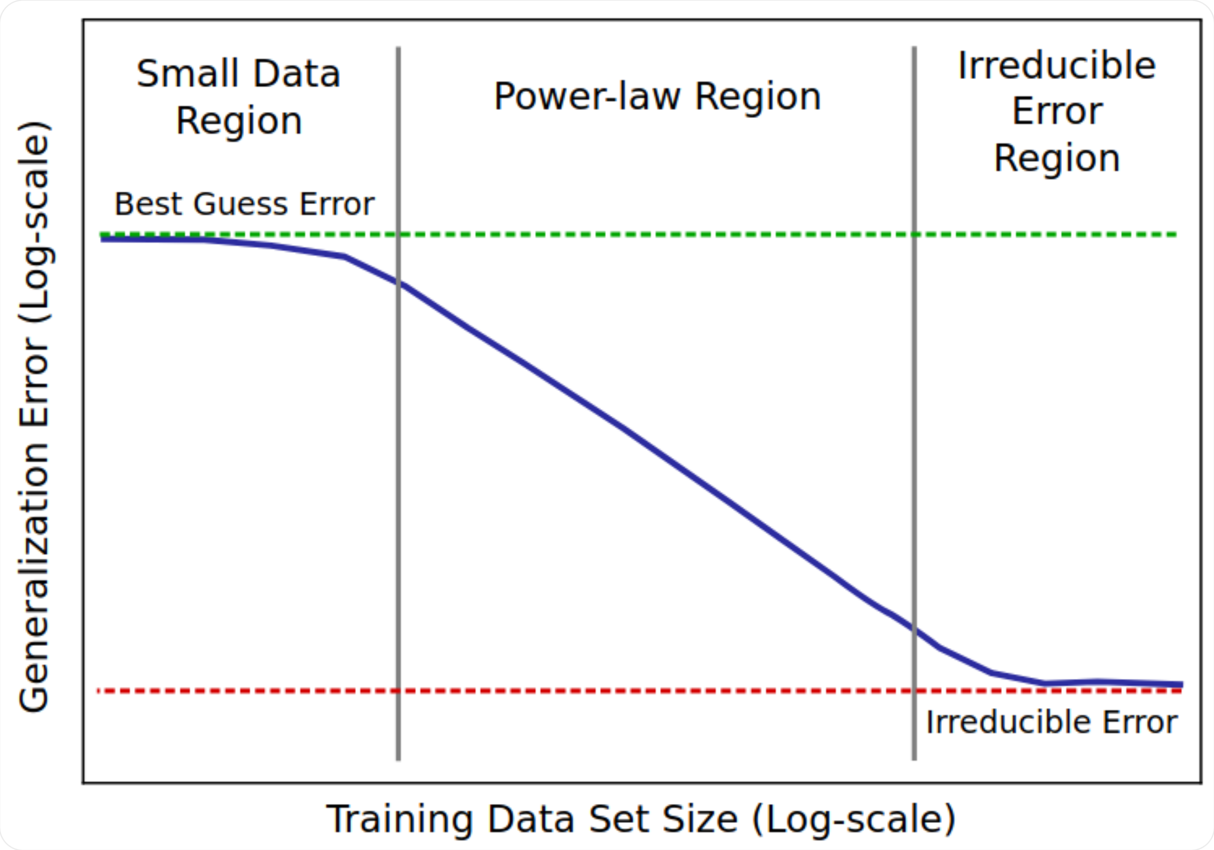
# Scaling Laws Literature Review

I have collected a database of scaling laws for different tasks and architectures, and reviewed dozens of papers in the scaling law literature.



Published

Jan 26, 2023

Authors

Pablo Villalobos

Resources

Cite

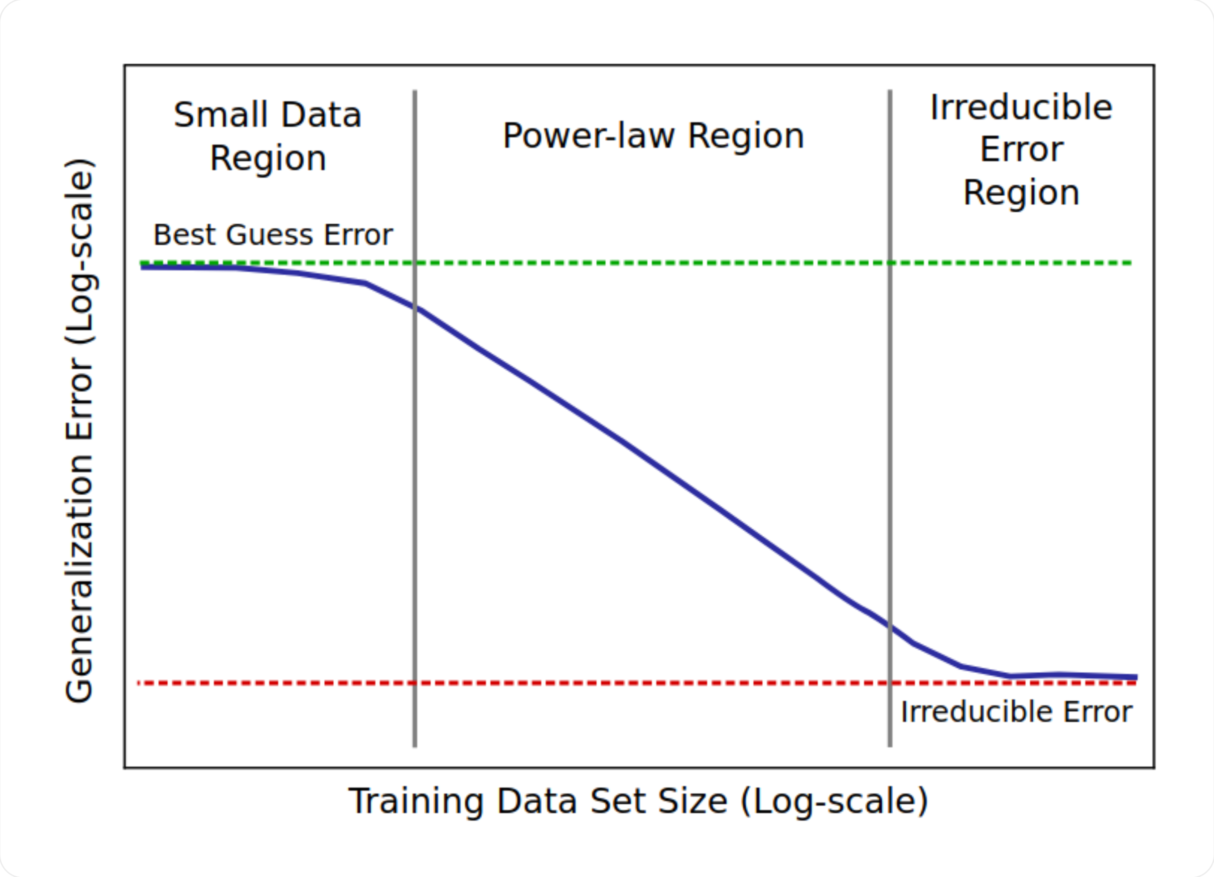
## Contents

Executive summary
Introduction
Overview
Takeaways
Upstream loss
Transfer and downstream accuracy

Theoretical analyses

Appendix: Paper reviews

Bibliography



Common shape of a scaling law, taken from Hestness et al. (2017)

# Executive summary

- Scaling laws are predictable relations between the scale of a mode and performance or other useful properties.
- I have collected a database of scaling laws for different tasks and architectures, and reviewed dozens of papers in the scaling law literature.
- My main takeaways are:
  - **Functional forms**: a basic power law can effectively model the scaling behavior in the power-law region but not the transitions

to the other two regions. For this, either the M4 estimator or the BNSL estimator introduced below seem to be the best options right now.

**Transfer learning**: there is not a simple universal scaling law for transfer learning between arbitrary tasks. When the tasks are similar enough, upstream loss and downstream performance are closely related, but when tasks are very different, the details of the architecture and hyperparameters become very relevant.

**See the full table of scaling laws here.**

# Introduction

The term “scaling laws” in deep learning refers to relations between functional properties of interest (usually the test loss or some performance metric for fine-tuning tasks) and properties of the architecture or optimization process (like model size, width, or training compute). These laws can help inform the design and training of deep learning models, as well as provide insights into their underlying principles.

In this document, I present a detailed table outlining the known scaling laws and a summary of each paper’s contributions to the understanding of scaling in deep learning.

My main goal for this document is to serve as a comprehensive and up-to-date resource for information on scaling laws in deep learning. By presenting a curated list of papers that have explored various scaling laws, and providing a detailed table outlining the known scaling laws and a summary of each paper’s contributions, I aim to make it easier to access and understand the current state of knowledge on scaling in deep learning.

I selected the papers using the following criteria:

- Publication date: I only focus on recent papers, and exclude work prior to 2015
- Contribution: I include papers which contribute in one of these ways:
  - Reporting empirical scaling laws
  - Proposing better functional forms or fitting methods
  - Proposing theoretical models to explain scaling behavior
  - Connecting scaling behavior with lower-level properties of models

# Overview

While the scaling behavior of machine learning models has been studied for a long time (e.g., Statistical Mechanics of Learning from Examples), it was only recently that empirical research into scaling deep learning models became widely known and incorporated into practice.

Previous theoretical analyses often predicted that the test loss would decrease as a power law of training data, $L = BD^{-b} + E$, with exponent $b = 1$ or $1/2$. However, this clashes with empirical results, in which the scaling exponent is usually smaller than $1/2$.

The modern study of scaling laws arguably started with Hestness et al. (2017), who empirically identified power-law scaling of the test loss with respect to training data size in several different domains. In Hestness et al. (2019) this previous result was used to predict the increases in model and dataset sizes that would be needed to reach important performance milestones.

Shortly after, Rosenfeld et al. (2020) constructed a joint error function with respect to model and data sizes, given by $L = AN^{-a} + BD^{-b} + E$ and showed that this form could accurately reproduce the empirical error landscape.

During 2020 and 2021, our understanding of scaling laws was greatly expanded:

- Henighan et al. (2020) found scaling laws for more tasks and architectures.
- Kaplan et al. (2020) tested them at much larger scales.
- Sharma et al. (2020), Hutter (2021) and Bahri et al. (2021) proposed theoretical mechanisms behind the values of the scaling exponents.
- Hernandez et al. (2021), Mikami et al. (2021) and Abna et al. (2021) studied scaling laws in the setting of transfer learning.

More recently, there have been several surprising results in the field that seem to suggest that the previously found scaling laws are less universal than one might have expected, particularly for downstream performance:

- Hoffmann et al. (2022) revealed that the scaling laws found by Kaplan et al. (2020) were suboptimal, after finding better hyperparameter choices and testing at larger scales.
- Sorscher et al. (2022) showed that, at least for some tasks, loss can scale exponentially with dataset size, rather than as a power-law.
- Tay et al. (2022) were able to drastically improve downstream performance with a small amount of extra pretraining on a different objective.
- Caballero et al. (2022) found that trend breaks in scaling curves can’t be easily predicted from smaller scale data.

In addition, scaling laws have been found for other properties, such as Goodharting (Gao et al. 2022), and there have been more connections between high-level scaling and low-level model features.

# Takeaways

## Upstream loss

For a vast collection of architectures and tasks, test loss for direct training scales predictably, transitioning from an initial random-guessing loss, to a power law regime, to a plateau at the irreducible loss. This is true for parameter, data, and compute scaling, and is relatively independent from the choice of architecture and hyperparameters, with some exceptions such as the learning rate schedule.

Except for the initial transition from random guessing to the power-law regime, and the transition from the power-law regime to the final plateau, the loss is well approximated by the functional form $L(N, D) = AN^{-a} + BD^{-b}$. There are several possible ways of modeling the transitions, but empirically it seems the best ones are the M4 estimator from Alabdulmohsin et al. (2022):

$$\frac{L - E}{(I - L)^{\alpha}} = AN^{-a} + BD^{-b}$$

and the BNSL estimator from Caballero et al. (2022), although it has significantly more parameters:

$$L(D) = E + (b \cdot D^{-c_0}) \prod_{i=1}^{n} (1 + (\frac{D}{d_i})^{\frac{1}{f_i}})^{-c_i f_i}$$

## Transfer and downstream accuracy

While ‘scale is all you need’ seems mostly true for direct training, when it comes to transfer learning, the downstream performance critically depends on the tasks at hand as well as the choice of architecture and hyperparameters (Tay et al., 2022). When the upstream and downstream tasks are similar, downstream loss can be reasonably well predicted from upstream loss, but this is not the case when the two tasks are substantially different (Abnar et al., 2021).

## Theoretical analyses

So far the most convincing theoretical analyses I’ve found are Sharma et al. (2020) and Bahri et al. (2021). In them, the magnitude of the scaling exponents are shown to be inversely proportional to the intrinsic dimension of the data manifold, both empirically and theoretically.

# Appendix: Paper reviews

For a summary of all the papers reviewed, see here.

# Bibliography

1. Abnar, Samira, et al. *Exploring the Limits of Large Scale Pre-Training*. arXiv, 5 Oct. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2110.02095.
2. Alabdulmohsin, Ibrahim, et al. *Revisiting Neural Scaling Laws in Language and Vision. 2,* arXiv, 1 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2209.06640.
3. Ardalani, Newsha, et al. *Understanding Scaling Laws for Recommendation Models*. arXiv, 17 Aug. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2208.08489.
4. Bahri, Yasaman, et al. *Explaining Neural Scaling Laws*. arXiv, 12 Feb. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2102.06701.
5. Bansal, Yamini, et al. *Data Scaling Laws in NMT: The Effect of Noise and Architecture*. arXiv, 4 Feb. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2202.01994. 6 Caballero, Ethan, et al. *Broken Neural Scaling Laws*. arXiv, 10 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2210.14891.
6. Droppo, Jasha, and Oguz Elibol. *Scaling Laws for Acoustic Models*. arXiv, 11 June 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2106.09488.
7. Gao, Leo, et al. *Scaling Laws for Reward Model Overoptimization*. arXiv, 19 Oct. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2210.10760.
8. Ghorbani, Behrooz, et al. *Scaling Laws for Neural Machine Translation*. 1, arXiv, 16 Sept. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2109.07740.
9. Gordon, Mitchell A., et al. "Data and Parameter Scaling Laws for

Neural Machine Translation.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 5915–22. ACLWeb, https://doi.org/10.18653/v1/2021.emnlp-main.478.

10. Henighan, Tom, et al. *Scaling Laws for Autoregressive Generative Modeling*. arXiv, 5 Nov. 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.2010.14701.
11. Hernandez, Danny, et al. *Scaling Laws for Transfer*. arXiv, 1 Feb. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2102.01293.
12. Hestness, Joel, et al. *Deep Learning Scaling Is Predictable, Empirically*. arXiv, 1 Dec. 2017. *arXiv.org*, https://doi.org/10.48550/arXiv.1712.00409.
13. Hoffmann, Jordan, et al. *Training Compute-Optimal Large Language Models*. arXiv, 29 Mar. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2203.15556.
14. Hutter, Marcus. *Learning Curve Theory*. arXiv, 8 Feb. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2102.04074.
15. Jones, Andy L. *Scaling Scaling Laws with Board Games*. arXiv, 15 Apr. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2104.03113.
16. Kaplan, Jared, et al. *Scaling Laws for Neural Language Models*. arXiv, 22 Jan. 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.2001.08361.
17. Mikami, Hiroaki, et al. *A Scaling Law for Synthetic-to-Real Transfer: How Much Is Your Pre-Training Effective?* arXiv, 8 Oct. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2108.11018.
18. Neumann, Oren, and Claudius Gros. *Scaling Laws for a Multi-Agent Reinforcement Learning Model*. arXiv, 29 Sept. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2210.00849.
19. Ramasesh, Vinay Venkatesh, et al. *Effect of Scale on Catastrophic Forgetting in Neural Networks*. 2022. *openreview.net*, https://openreview.net/forum?id=GhVS8_yPeEa.
20. Rosenfeld, Jonathan S., et al. *A Constructive Prediction of the Generalization Error Across Scales*. arXiv, 20 Dec. 2019. *arXiv.org*, https://doi.org/10.48550/arXiv.1909.12673.

21. Rosenfeld, Jonathan S. *Scaling Laws for Deep Learning*. 1, arXiv, 17 Aug. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2108.07686.
22. Seung, H. S., et al. “Statistical Mechanics of Learning from Examples.” *Physical Review A*, vol. 45, no. 8, Apr. 1992, pp. 6056–91. APS, https://doi.org/10.1103/PhysRevA.45.6056.
23. Sharma, Utkarsh, and Jared Kaplan. *A Neural Scaling Law from the Dimension of the Data Manifold*. arXiv, 22 Apr. 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.2004.10802.
24. Sorscher, Ben, et al. *Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning*. arXiv, 15 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2206.14486.
25. Tay, Yi, Mostafa Dehghani, et al. *Scaling Laws vs Model Architectures: How Does Inductive Bias Influence Scaling?* arXiv, 21 July 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2207.10551.
26. Tay, Yi, Jason Wei, et al. *Transcending Scaling Laws with 0.1% Extra Compute*. arXiv, 16 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2210.11399.
27. Zhai, Xiaohua, et al. *Scaling Vision Transformers*. arXiv, 20 June 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2106.04560.

About the authors

**Pablo Villalobos** has a background in Mathematics and Computer Science. After spending some time as a software engineer, he decided to pivot towards AI. His interests include the economic consequences of advanced AI systems and the role of algorithmic improvements in AI progress.

Share

Twitter

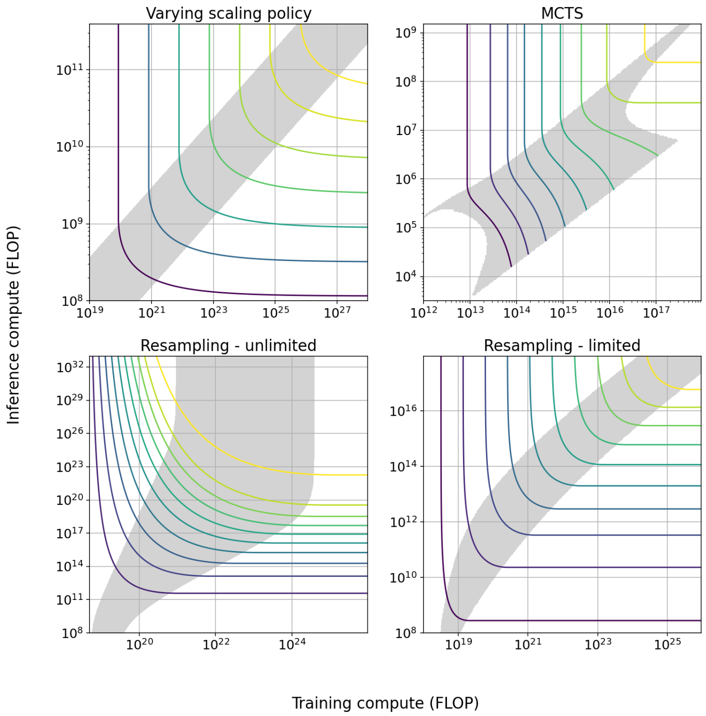
Tags

Literature reviews

LinkedIn

Scaling laws

## Related posts

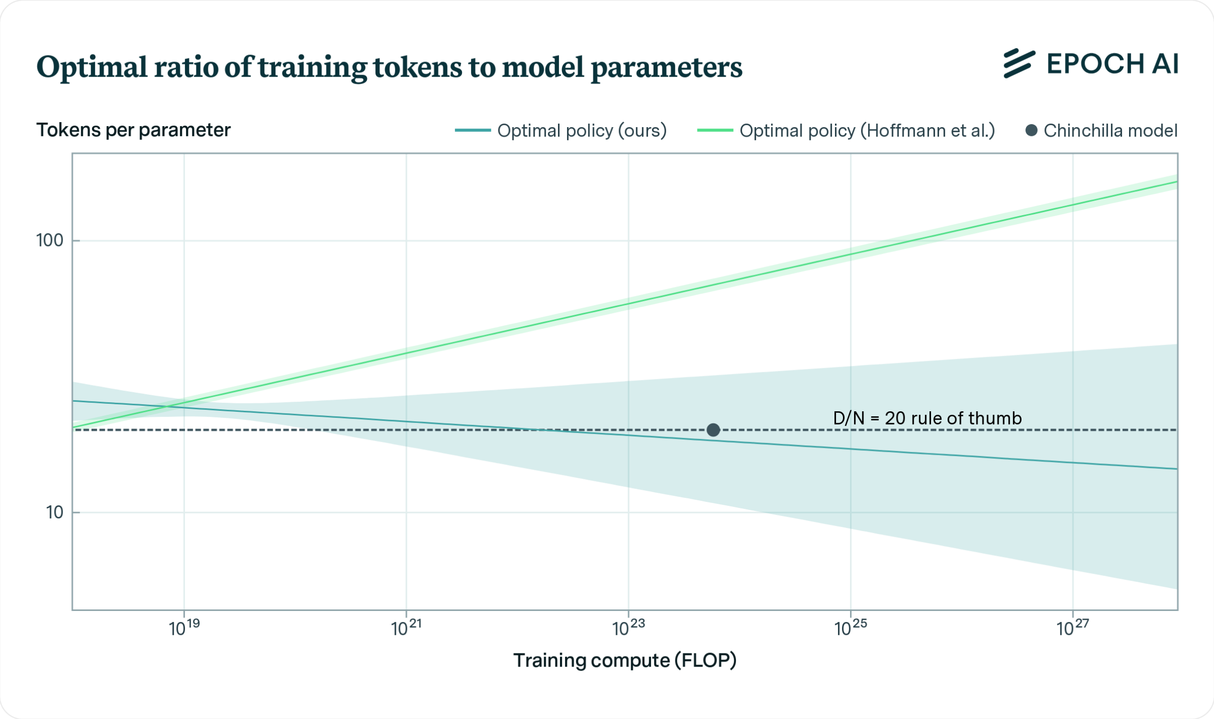


REPORT · 27 MIN READ

## Trading Off Compute in Training and Inference

We explore several techniques that induce a tradeoff between spending more resources on training or on inference and characterize the properties of this tradeoff. We outline some implications for AI governance.

Jul 28, 2023 · By Pablo Villalobos and David Atkinson

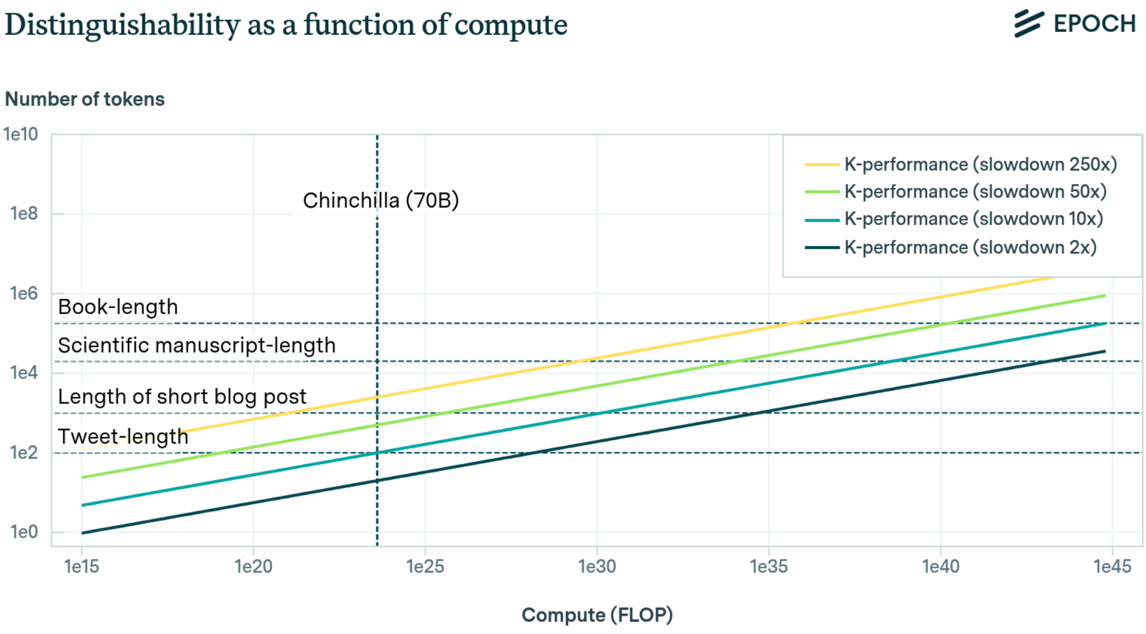


PAPER · 4 MIN READ

## Chinchilla Scaling: A Replication Attempt

We replicate Hoffmann et al.'s estimation of a parametric scaling law and find issues with their estimates. Our estimates fit the data better and align with Hoffmann's other approaches.

Apr 17, 2024 · By Tamay Besiroglu, Ege Erdil, Matthew Barnett and Josh You



REPORT · 10 MIN READ

## The Direct Approach

Empirical scaling laws can help predict the cross-entropy loss associated with training inputs, such as compute and data. However, in order to predict when AI will achieve some subjective level of performance, it is necessary to devise a way of interpreting the cross-entropy loss of a model. This blog post provides a discussion of one such theoretical method, which we call the Direct Approach.

Apr 25, 2023 · By Matthew Barnett and Tamay Besiroglu

Excited about our work?

Talk to us

Support our research

RESEARCH

Blog

Publications

Machine Learning Trends

Data

ORGANIZATION

About Epoch AI

Careers

Support us

Contact us

Privacy Notice

Epoch AI is fiscally sponsored by Rethink Priorities.