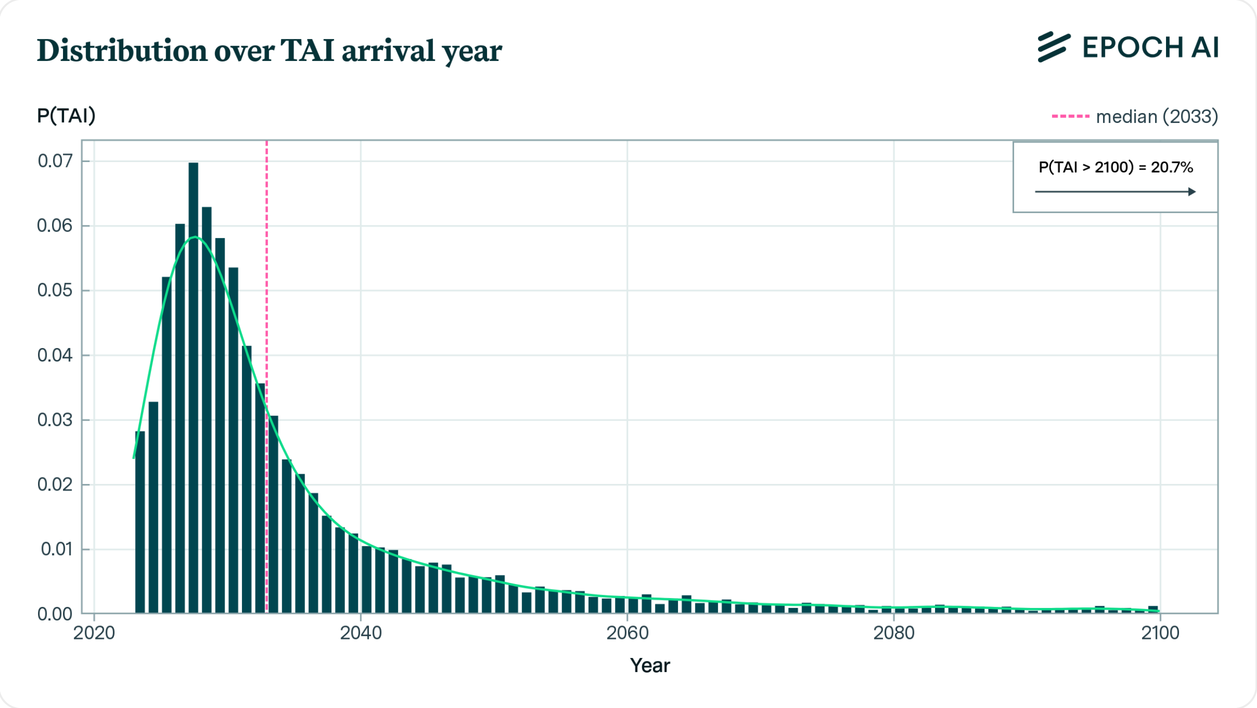
Articles Direct Approach Interactive Model

Direct Approach Interactive Model

We combine the Direct Approach framework with simple models of progress in algorithms, investment, and compute costs to produce a user-adjustable forecast of when TAI will be achieved.



Distribution over TAI arrival year

Published

May 31, 2023

Last updated Jan 09, 2024 ›

Authors

David Atkinson ›

Matthew Barnett ›

Edu Roldán ›

Ben Cottier ›

Tamay Besiroglu ›

Resources

Source Code

Cite

Menu

Summary: This post presents an interactive model for forecasting transformative AI, by which we mean AI that if deployed widely, would precipitate a change comparable to the industrial revolution. In addition to showcasing the results of the Direct Approach, we present a simple extrapolative model of key inputs (algorithmic progress, investment, hardware efficiency) that produce a user-adjustable forecast over the date transformative AI will be deployed. This model contains four parts:

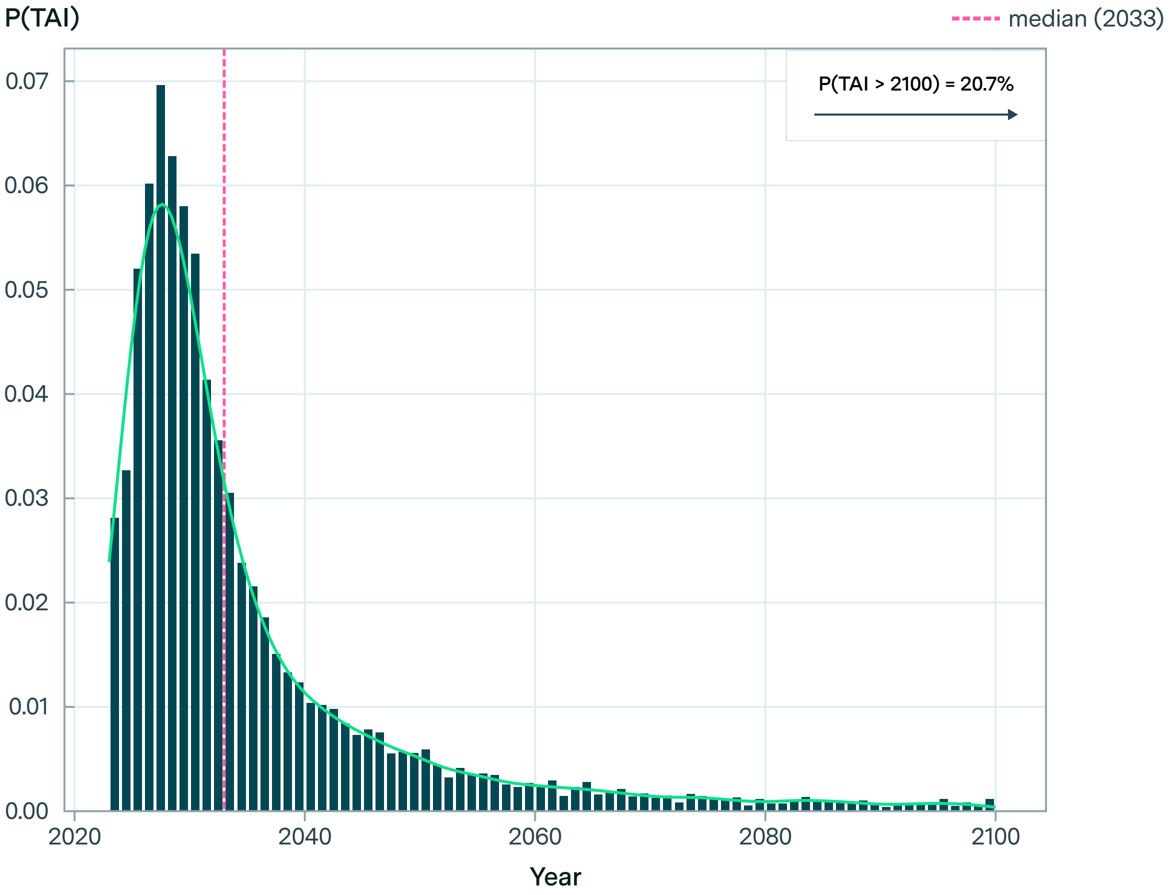
- Compute requirements estimated using the Direct Approach framework
- Projected algorithmic progress
- Projected investment in training transformative AI models
- Projected compute availability and cost

These components are combined to estimate the probability that the compute needs for transformative AI will be met in a given future year. Under default parameter values calibrated on historical estimates, the simple extrapolative model assigns a high chance of the development of transformative AI by 2050.

We take this to mean that current trends of algorithmic progress and compute scaling, if continued, will likely lead to AI systems within several decades that have dramatic effects on scientific progress and economic growth. The outputs of this model should not be construed as the authors' all-things-considered views on the question of when transformative AI will arrive. Instead, the model is better seen as potentially illustrating the predictions of well-informed extrapolative models.

Distribution over TAI arrival year

Distribution over the year at which the number of effective FLOP available will exceed the FLOP requirements predicted by the Direct Approach, making a transformative training run possible.

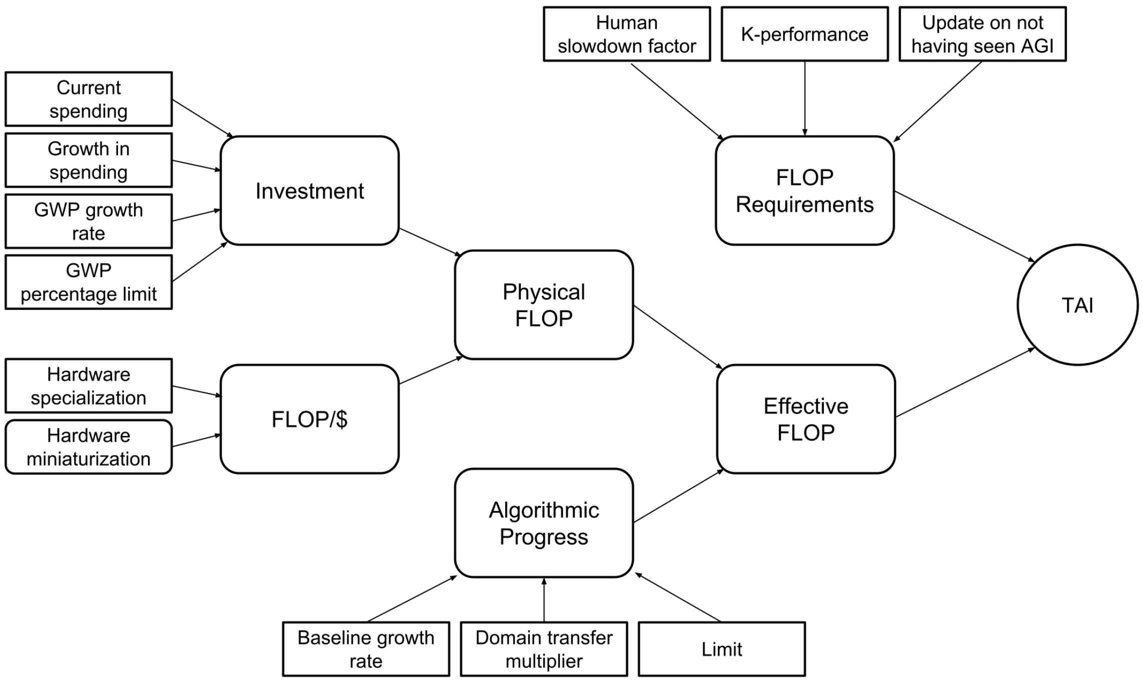


Show:

TAI arrival year probabilities

There have been many attempts to forecast the date at which AI models will achieve human level performance on a transformative task.[1] [2] We contribute another, making use of our Direct Approach framework (Barnett and Besiroglu, 2023), which uses neural scaling laws to bound the compute needed to train a transformative model. Our work is most similar to *Forecasting TAI with biological anchors* (Cotra, 2020).

Our forecast model has four user-adjustable components, pictured in the figure below: investment in model training, hardware price performance, algorithmic progress, and the compute requirements produced by the Direct Approach. Together, they produce a probability of TAI being trained during each year. Investment and hardware price performance combine to produce an estimate (which we call "physical FLOP") of the total FLOP available during a given year. When the amount of physical FLOP is adjusted to account for algorithmic progress, we get an estimate of the amount of "effective FLOP"—the amount of physical FLOP that would be required to reach a given level of performance if algorithms did not improve. Finally, this effective FLOP estimate is compared to the FLOP requirements produced by the Direct Approach. When the number of effective FLOP exceeds those requirements, we consider TAI to have been achieved.



Overview of the model's components, and how they relate to each other.

Collected model parameters

Compute requirements

Human slowdown factor:

6.60 – 433

K-performance:

1.99e+3 – 2.22e+5

Update on not having seen TAI:

□

Scale TAI requirements:

□

Algorithmic progress

Baseline growth rate (OOM/year):

0.244 – 0.775

Domain transfer multiplier:

0.355 – 1.24

Limit (OOM):

2.25 – 12.1

Investment

Current spending ($, millions):

60.0

Yearly growth in spending (%):

146 – 246

GWP growth rate (%):

0.400 – 4.75

Max % of GWP spent on training (%):

0.0137 – 1.46

Compute

Growth in FLOP/s/$ from hardware specialization (OOM/year):

0.0400 – 0.250

Samples:

2000

Regenerate timeline

Compute requirements

The compute requirements for human-level AI are generated using our Direct Approach framework, which is an approach based on extrapolating a neural scaling law. This approach requires three parameters, which are explained in more detail in (Barnett and Besiroglu, 2023). Our estimates are intended to provide illustrative and highly speculative guesses for the parameters, taking into account the fact that the Chinchilla scaling law was derived from models trained on internet data, rather than scientific data.

Scaling law. We employ the scaling law estimated by Hoffmann et al., 2022, which takes the form of $L = E + AN^{-\alpha} + BD^{-\beta}$. The central estimates of the parameters are: A = 306, B = 411, α = 0.34 and β = 0.28. We model these as the means of normal distributions with standard deviation = 25 for parameters A and B, and standard deviation = 0.05 for α and β.

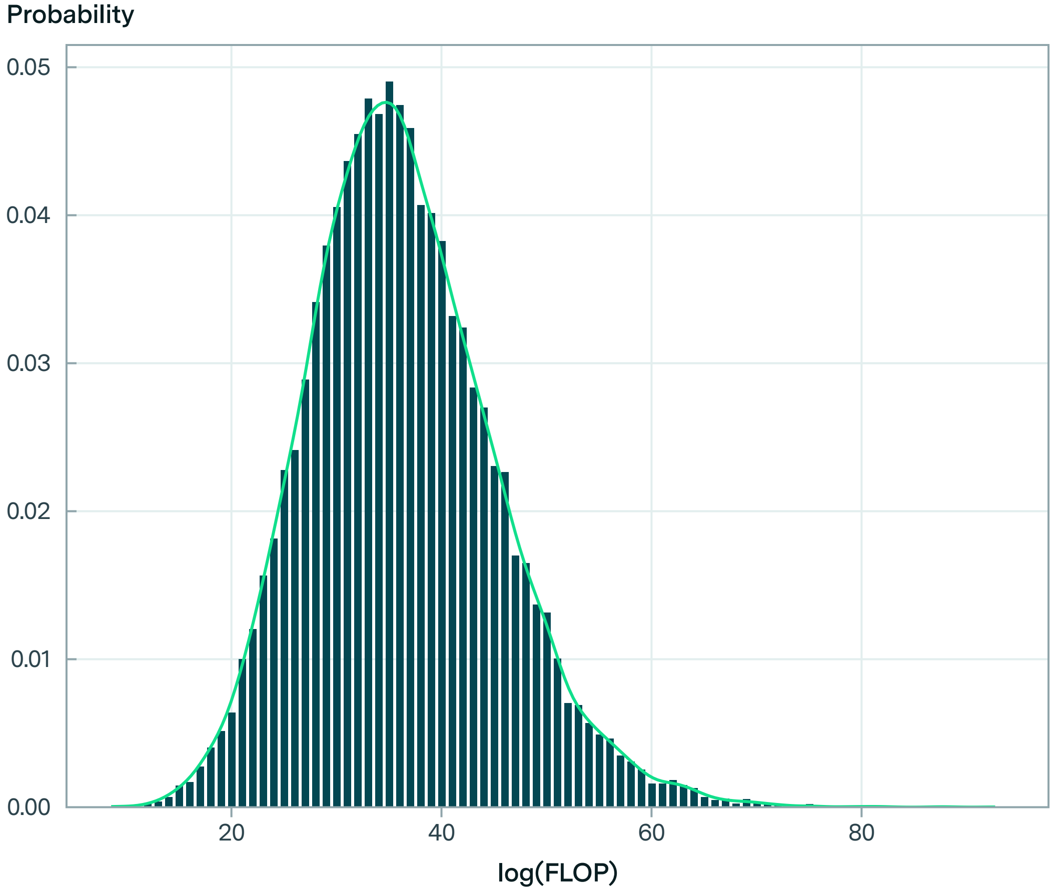
***K*-performance.** This refers roughly to the average number of tokens generated by a model that a human judge requires in order to reliably distinguish the model's outputs from a human expert's outputs, on some task. Our default distribution, generated by an internal poll, is lognormal with an 80% confidence interval of 2,000 to 222,000. This was anchored to the number of tokens in a typical scientific manuscript.[3] This reflects our view that producing scientific work indistinguishable from actual scientific work is a strong candidate for a task, which, when mastered by AI, is likely to be transformative.

Slowdown factor. This parameter represents the multiplier for the number of tokens human discriminators will need to see compared to an ideal Bayesian predictor, when distinguishing between human- and computer-generated texts. Informed by an internal poll, we enforce a lognormal distribution with a median of 53.1, a 15th percentile estimate of 9.84, and an 85th percentile of 290. This parameter is perhaps the most subjective of all the parameters considered, as we do not yet have reliable data on how quickly trained human judges can discriminate between real and model-generated texts.

These distributions produce the following distribution over the effective FLOP requirements.[4]

Distribution over effective FLOP required for TAI before adjustment

Distribution over the number of effective FLOP needed to reach a transformative level of performance, before adjusting for not having seen transformative AI.



Human slowdown factor:

6.60 – 433

K-performance:

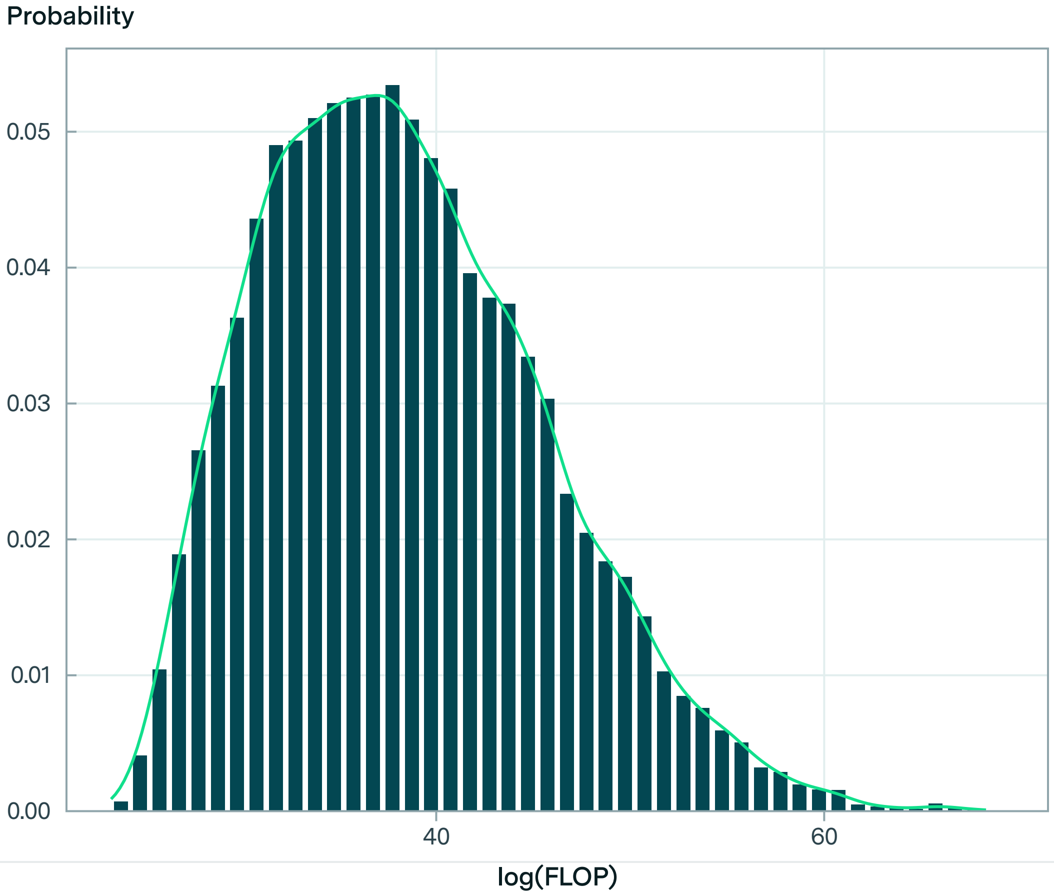
1.99e+3 – 2.22e+5

Regenerate timeline

We apply a further adjustment to the distribution to reflect our observation of having not achieved transformative AI at a level of O(1e25) FLOP, which is commensurate with the compute-intensity of the largest training run to date.[5]

Distribution over effective FLOP required for TAI after adjustment

Distribution over the number of effective FLOP needed to reach a transformative level of performance, after adjustment.



Show:

Probability density function

Apply update on not having seen TAI:

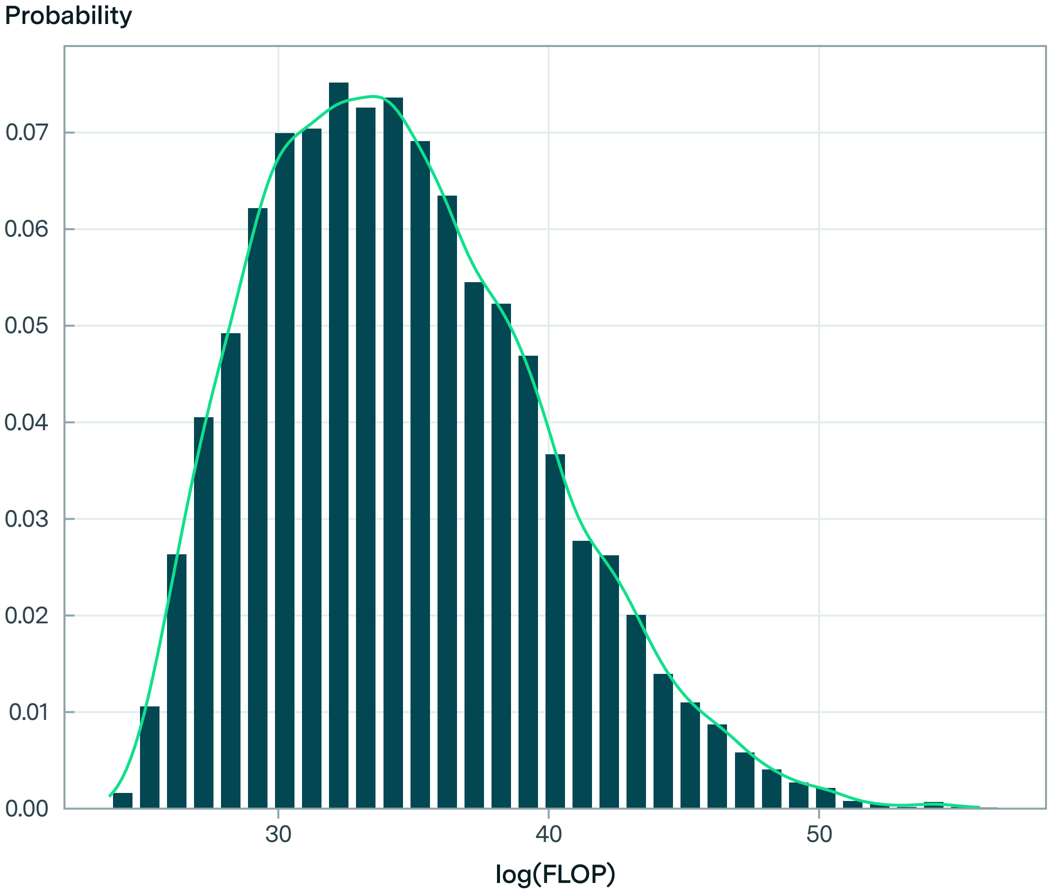
□

Regenerate timeline

The Direct Approach framework yields a soft-upper bound, meaning that we can be reasonably confident that the effective compute requirements will not exceed the bound for a given set of parameter values. The reason is that a more efficient means of automating high-quality reasoning may be developed than the method of training large models to emulate human reasoning. However, we think this bound is unlikely to be more than a few orders of magnitude larger than what is required if we assume that deep-learning based autoregressive models remain the dominant approach in AI. To adjust for the fact that this framework produces an upper bound, rather than a central estimate, we have automatically scaled the compute requirements distribution by 5/7 around the point 10^25 to put more probability on lower compute values.[6]

Distribution over effective FLOP required for TAI after adjustment and scaling

Distribution over the number of effective FLOP needed to reach a transformative level of performance, after adjustment and scaling.



Show:

Probability density function

Apply scaling:

□ ⓘ

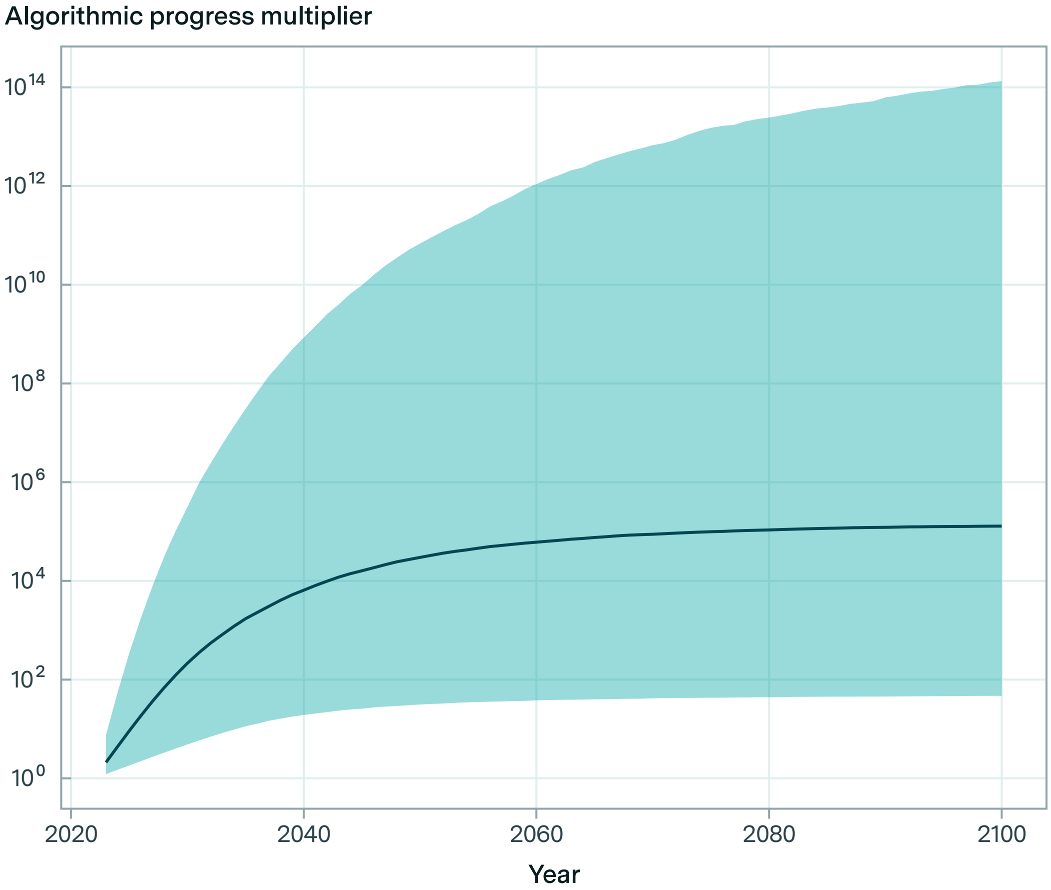
Regenerate timeline

Algorithmic Progress

Algorithmic progress refers to improvements in the algorithms and architectures of machine learning models that reduce the amount of compute necessary to reach a given level of performance.

The extent to which progress in algorithms increases the number of effective FLOP. Error bars contain 90% of the distribution for that year.

EPOCH AI



Baseline growth rate (OOM/year):

0.244 – 0.775 ⓘ

Domain transfer multiplier:

0.355 – 1.24 ⓘ

Limit (OOM):

2.25 – 12.1 ⓘ

Regenerate timeline

We use three parameters to model algorithmic progress:

Baseline rate of efficiency improvements. Erdil and Besiroglu (2022) estimate that the historical rate of algorithmic progress in computer vision models has been 0.4 orders of magnitude per year (80% CI: 0.244 to 0.775).

Domain transfer multiplier. We might well expect the pace of progress towards transformative models to be meaningfully different from that of progress in computer vision algorithms. To account for that, we multiply the baseline growth rate with a parameter—lognormally distributed, with values elicited via an internal poll—representing the extent to which we expect progress in potentially-transformative models to be faster or slower than the rate in computer vision.

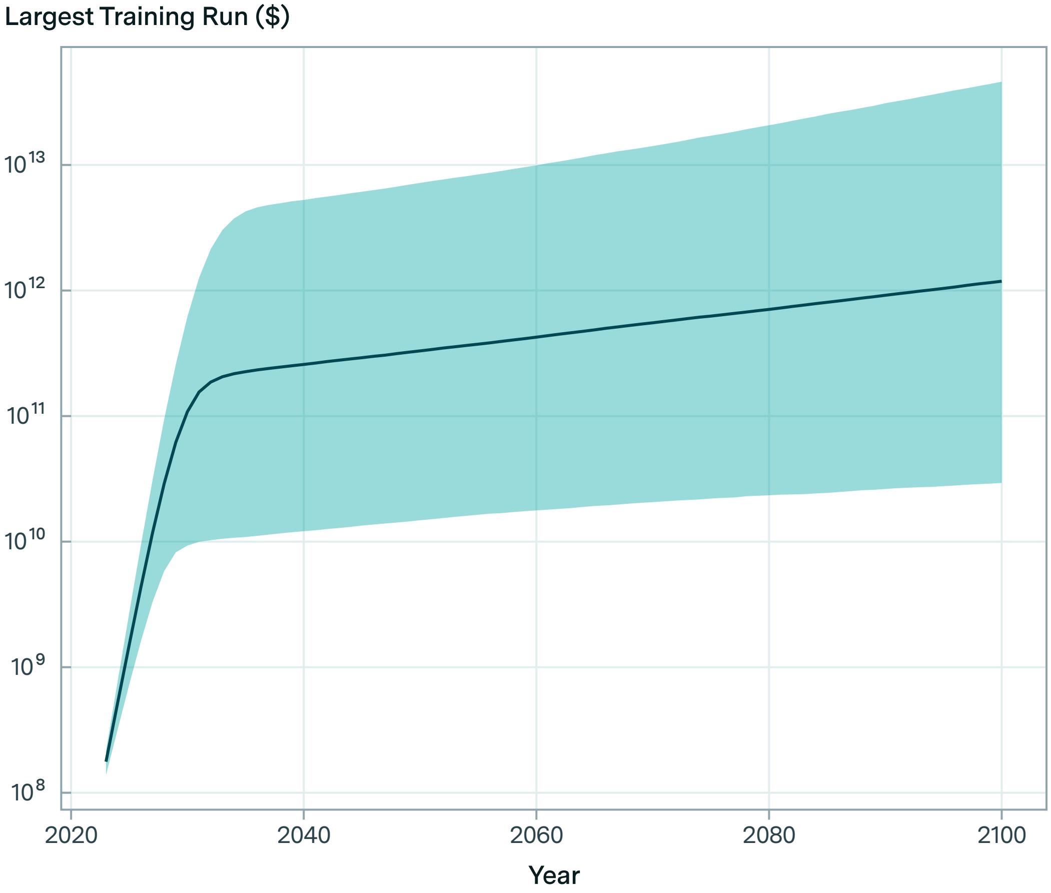
Algorithmic progress limit. It's reasonable to expect that the previously-seen exponential growth in algorithmic efficiency will not continue indefinitely. This parameter addresses this concern by representing the maximum possible performance multiple relative to 2023. The parameter follows a lognormal distribution, with its default interval sourced from an internal poll.

At each timestep, algorithmic efficiency increases at the baseline growth rate times the domain transfer multiplier. As the resulting multiplier reaches the limit, the rate of growth approaches zero.

Investment

Historically, investment in large-scale training runs has grown at between 25.9% and 151.3% per year, over the 2015 to 2022 period (Cottier, 2023a). We model the investment in potentially-transformative training runs by combining estimates of the dollar cost of current training runs, projections of future GWP growth, and estimates for the maximum percentage of GWP that might be spent on such a training run.

The projected dollar cost of potentially-transformative training runs. Error bars contain 90% of the distribution for that year.



Current spending ($, millions):

60.0

Yearly growth in spending (%):

146 – 246

GWP growth rate (%):

0.400 – 4.75

Max % of GWP spent on training (%):

0.0137 – 1.46

Regenerate timeline

Current spending. We estimate that the current largest training run in 2023 will cost $61 million, using our tentative guess at GPT-4's training cost (Cottier, 2023b), and updating it in line with previous growth rates found in Cottier, 2023a.

Yearly growth in spending. Following Cottier, 2023a, we assume spending will increase between 146% and 246% per year (80% CI).

GWP growth rate. OECD, 2021 projects that world GWP will be $238 trillion in 2060, implying a 1.9% per year growth rate. We use this as the center of 80% CI of 0.400% to 4.75% per year.

Maximum percentage of GWP spent on training. Since spending on training runs is growing faster than GWP, a naive extrapolation suggests that we will eventually see training runs which cost many multiples of GWP. To account for this, we include a parameter that represents the maximum possible share of GWP that can be spent on a training run. We tentatively put this at 0.0137% to 1.46% (80% CI).

As spending approaches the limit defined by the GWP and the GWP percentage limit, growth in spending slows. Eventually, spending growth is entirely limited by the growth of GWP.

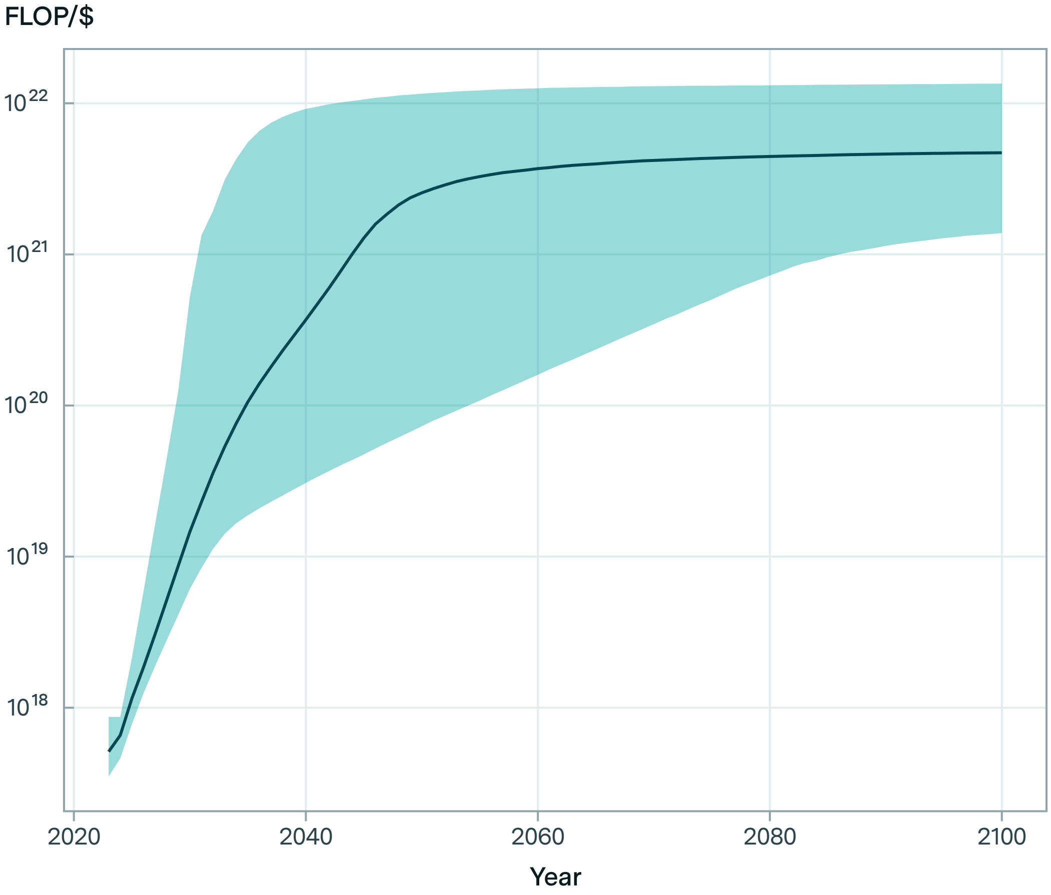
Compute

The first aspect of compute we consider is the progress of FLOP/s per dollar. We modify projections of the growth in FLOP/s in top GPUs

(Hobbhahn and Besiroglu, 2022) to incorporate an estimate of the benefits from **hardware specialization**.

Projected cost of compute. Error bars contain 90% of the distribution for that year.

EPOCH AI



Growth in FLOP/s/$ from hardware specialization (OOM/year):

0.0400 – 0.250

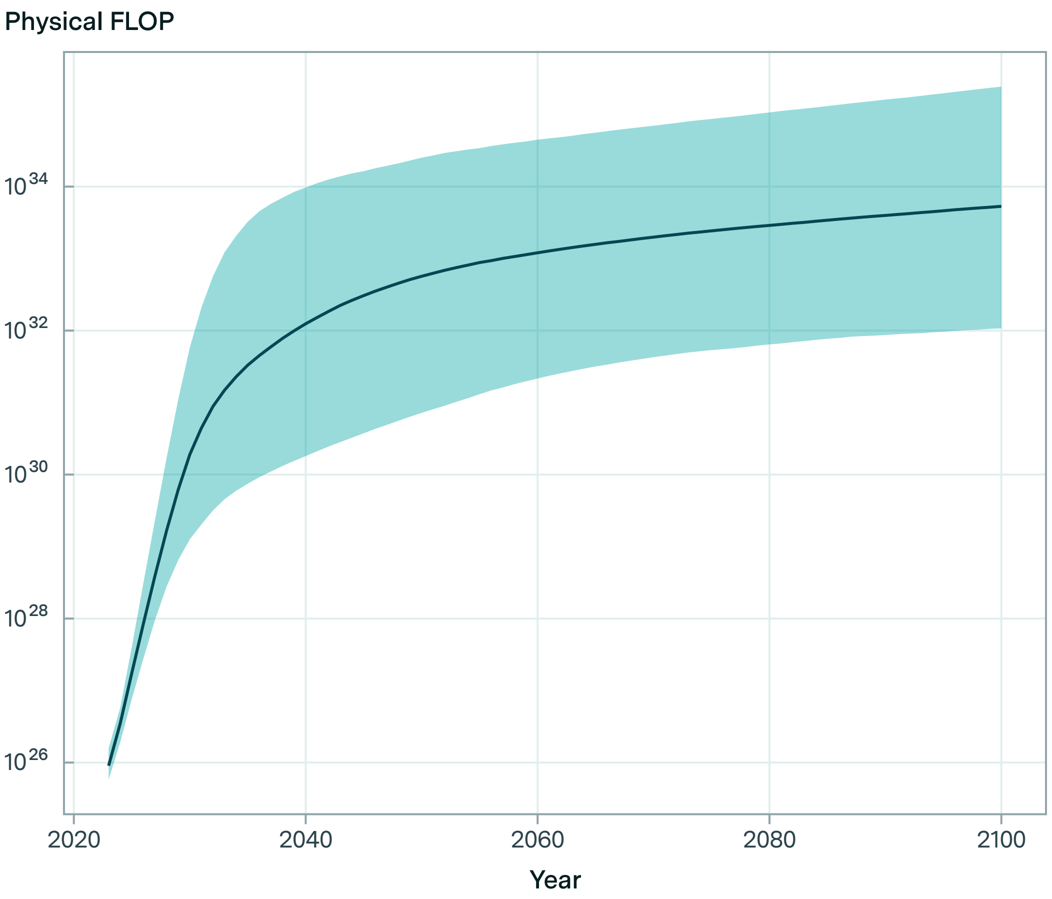
Regenerate timeline

The hardware specialization parameter is designed to account for future adjustments for specific workloads, including changes to parallelism, memory optimization, data specialization, quantization, and so on. We assume that gains from such specialization will accrue up until a limit,[7] and model the per-year growth of hardware specialization gains with a lognormal distribution, using values elicited from an internal poll.

At each timestep, we use the current rate of FLOP/s growth to estimate the lifetime of a typical GPU. This lifetime, combined with an estimate of the cost of a typical GPU, produces an estimate of the FLOP/$ available at that timestep.[8]

Projected number of FLOP available for potentially-transformative training runs. Error bars contain 90% of the distribution for that year.

EPOCH AI



Using these FLOP/$ figures, along with the projection of spending described above, we can estimate the number of FLOP available each year (“Physical FLOP”).

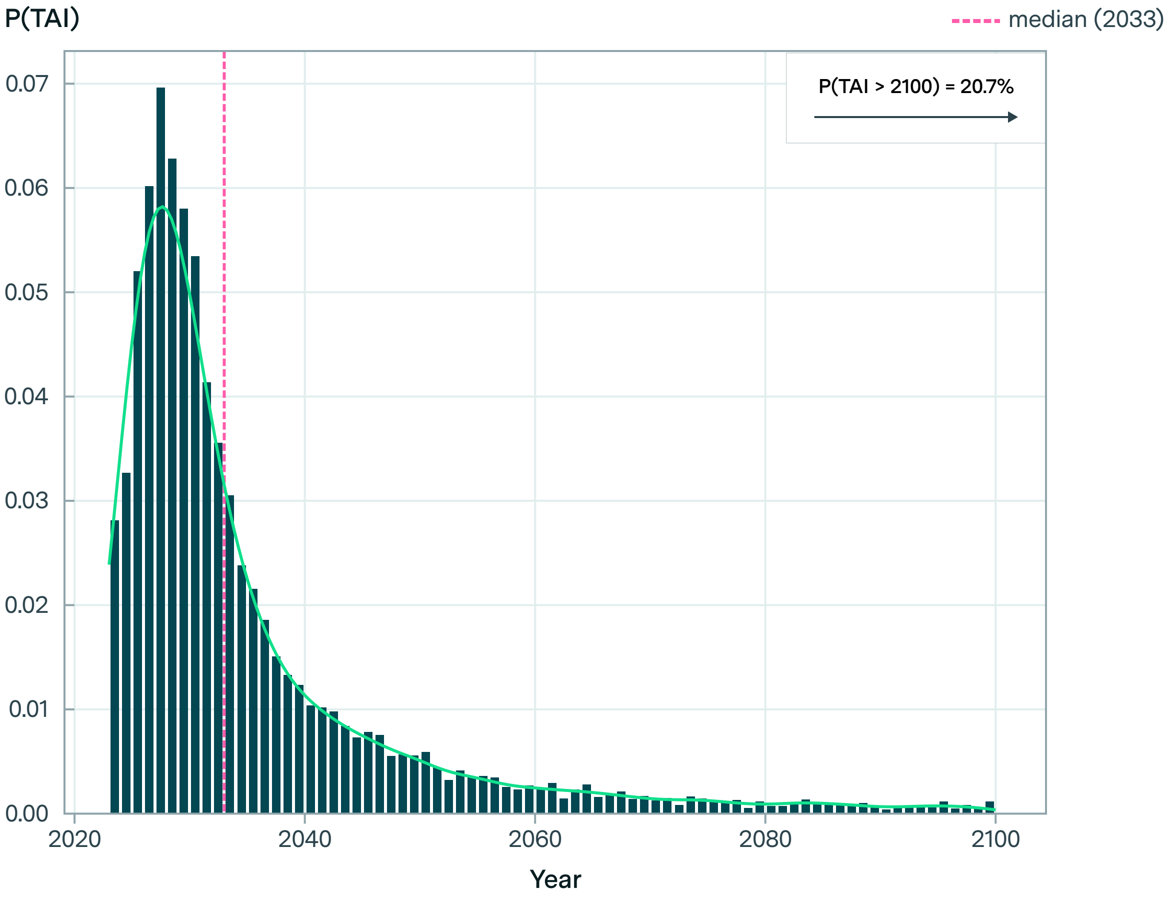
Conclusion

We combine the estimate of physical FLOP with the previous estimate of the algorithmic progress multiplier to get the number of effective FLOP available by year. This is compared to our estimate of the effective FLOP requirements to compute the probability of TAI per year.

Distribution over TAI arrival year

EPOCH AI

Distribution over the year at which the number of effective FLOP available will exceed the FLOP requirements predicted by the Direct Approach, making a transformative training run possible.



Show:

TAI arrival year probabilities

Samples:

2000

Regenerate timeline

Probability of TAI by...		
2030	**2050**	**2100**
42%	71%	79%

Quantile		
10%	**Median**	**90%**
2025	2033	>2100

In this model, for each year, we consider Transformative AI (TAI) to have arrived if the number of effective FLOP surpasses the estimate generated by the Direct Approach.

The model implemented here is fundamentally extrapolative, meaning that is based on projections of progress in hardware and software and expansions in investment based on historical data. This is in contrast to more sophisticated models that ‘endogenize’ such processes by having explicit models of the economic processes that drive these trends (such that of Davidson, 2022, implemented by Epoch AI here). While evaluating the differences between these modeling approaches is beyond our scope, it should be noted that extrapolative models could produce wider distributions than endogenous models, and that extrapolative models can tend to overstate the likelihood of especially long timelines.

We thank Daniel Kokotajlo, Carl Shulman and Charlie Giattino for their feedback on a draft of this report. We also thank Ege Erdil, Jaime Sevilla, Pablo Villalobos and the rest of the Epoch AI team for their suggestions and previous research.

Appendix: Simulation specification

To generate our results, we run the model described below many times. Each run, or “rollout,” produces an estimate for the value of each component of the model. The results plotted above report the distribution of the values over the rollouts.

In what follows:

- $\lambda(D, l, u)$ is a function which renormalizes a probability distribution D to lie in (l, u)
- The notation $x \leftarrow D$ is used to indicate that x is sampled from D.
- The function κ, defined as $\kappa(v, l) = l(1 - \exp(-v/l))$, constrains the value of v to be no higher than l.

The concrete numbers are chosen to produce distributions which match the default confidence intervals provided by the interactive tool. Lognormal distributions are parameterized with the mean and standard deviation of the underlying normal distribution.

Investment

Algorithmic Improvements

Compute

Combined Model

Notes

1. By a transformative task we imagine tasks like scientific research, which if cheaply automated, we think will likely drastically accelerate economic growth rates. ↩

2. See Wynroe, et al, 2023 for a review. ↩

3. Fire and Guestrin, 2019 analyze 120M scientific papers and find that the average number of pages in 2014 was 8.4. Assuming that there are 500 words per page, and 0.75 tokens/word, this amounts to 5600 tokens/publication. ↩

4. Note that this is effective FLOP instead of physical FLOP, as this is the estimate for the amount of FLOP needed to reach transformative levels of automation with 2023 algorithms. We will address algorithmic progress in a later section of the model. ↩

5. This adjustment was modeled as a Bayesian update for parameter values of a scaling law that defined the arrival rate of transformative AI in a poisson process. The power law has two parameters: a coefficient and an exponent. The exponent is fixed at 0.3, as it merely determines the smoothing of the update. The prior over the coefficient is set such that the prior FLOP requirement distribution defines the scale over which you expect transformative AI to arrive in 1 year. The parameter values were updated on the basis of seeing a series of sequential training runs at various FLOP scales, with the amount of time between training runs. The training run sequences that produced the distribution seen below were: (10^17 FLOP, 8 years), (10^23 FLOP, 2 years), (10^25 FLOP, 1 year). ↩

6. In case the reader does not wish to apply this ad-hoc adjustment, they can disable this scaling by unselecting the "Apply scaling" box in the compute requirements section. ↩

7. The TPU-V1 was on average about 15X-30X faster than its contemporary GPU lacking architectural optimizations for AI workloads (Jouppi et al., 2017), and later versions of the TPU, such as the V3 provide an additional order of magnitude of performance in similar floating point representations per chip. While a large part of the improvements in accelerators is often explained by CMOS scaling (see Fuchs and Wentzlaff, 2018) rather than architectural improvements alone, it is not unlikely that an increase in tensor cores and a reduction

in relevant precision could together provide an increase of performance equivalent to increasing FLOP/s by 1 or 2 orders of magnitude in some given floating point precision. In light of this, given that improvements “at the top” are likely to be exhausted in the short-to-medium term, we set a cap of 250x on the additional performance improvements from a continuation of similar architectural improvements. ↩

8. Specifically, at each year, we treat the lifetime of a typical GPU as 1.2/(*r*+0.1), where *r* is the value of that year’s growth in FLOP/s. (If hardware is improving quickly at 30% each year, for example, replacement times will be a relatively short three years. In the limit, if hardware improvements stop, replacement times will be longer: 12 years). The GPU lifetime multiplied by the FLOP/s tells us how many FLOP the GPU will make available over its lifetime. Finally, we divide that by the cost of a top GPU to get the FLOP/$, setting the cost of a GPU to $5000 so that it provides 4e17 FLOP/$ in 2023, which is in line with the realized price performance of today’s GPUs used in large-scale training runs. ↩

Updates

About the authors

Former employee
David Atkinson studied at Deep Springs college, and graduated from the University of Colorado, Boulder, with a degree in Computer Science and Mathematics. Before coming to Epoch AI, he did some NLP research and then worked as a software engineer. He’s interested in model interpretability and forecasting.

Matthew Barnett received a degree in computer science from UC Berkeley. He’s currently interested in getting a clearer picture of how AI will impact the world in the next few decades.

Edu Roldán is a software developer at Epoch AI. He maintains the website and assists with other programming tasks, helping the team to delve into research.

Ben Cottier's research interests include the diffusion of AI capabilities among actors, and measuring the effects of different inputs to AI progress. Previously, he was a Research Fellow at Rethink Priorities, and spent time as a software engineer. Ben has a background in machine learning.

Tamay Besiroglu is the associate director at Epoch AI. His work focuses on the economics of computing and big-picture trends in machine learning. Previously, he was a researcher at the Future Tech Lab at MIT, led strategy for Metaculus, consulted for the UK Government, and worked at the Future of Humanity Institute.

Share

Twitter

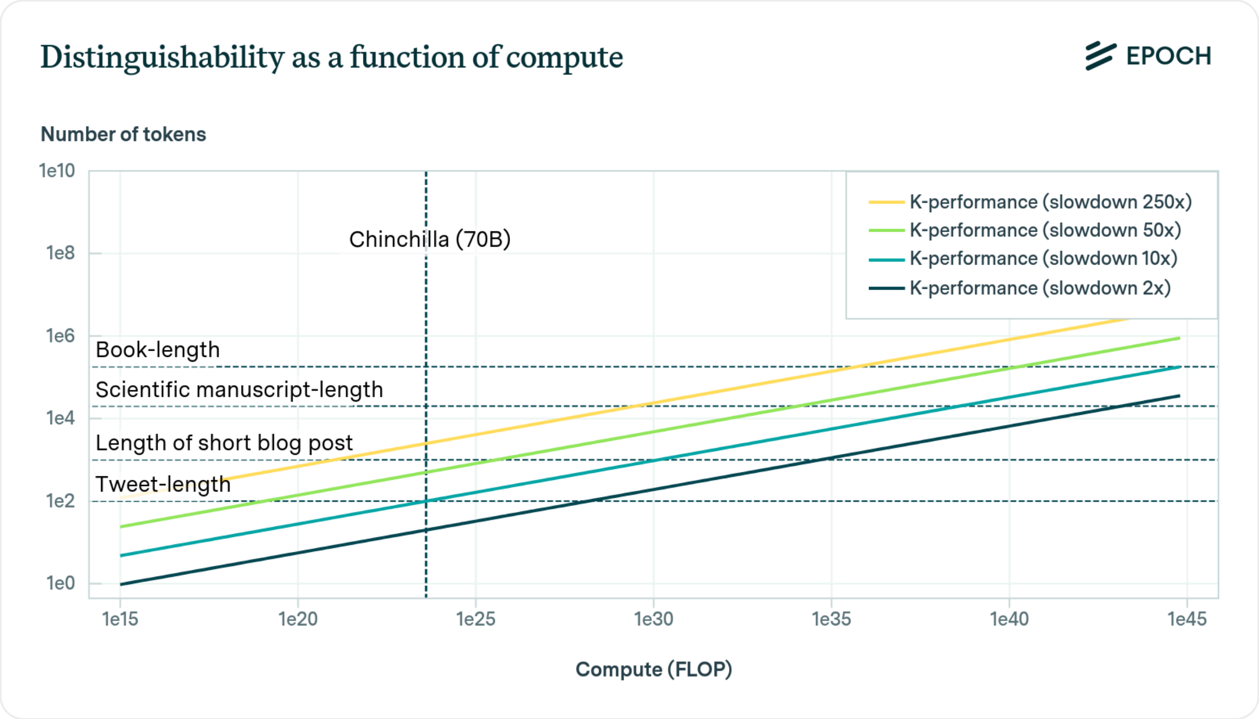
LinkedIn

Tags

Interactive models

Dashboards

Related posts

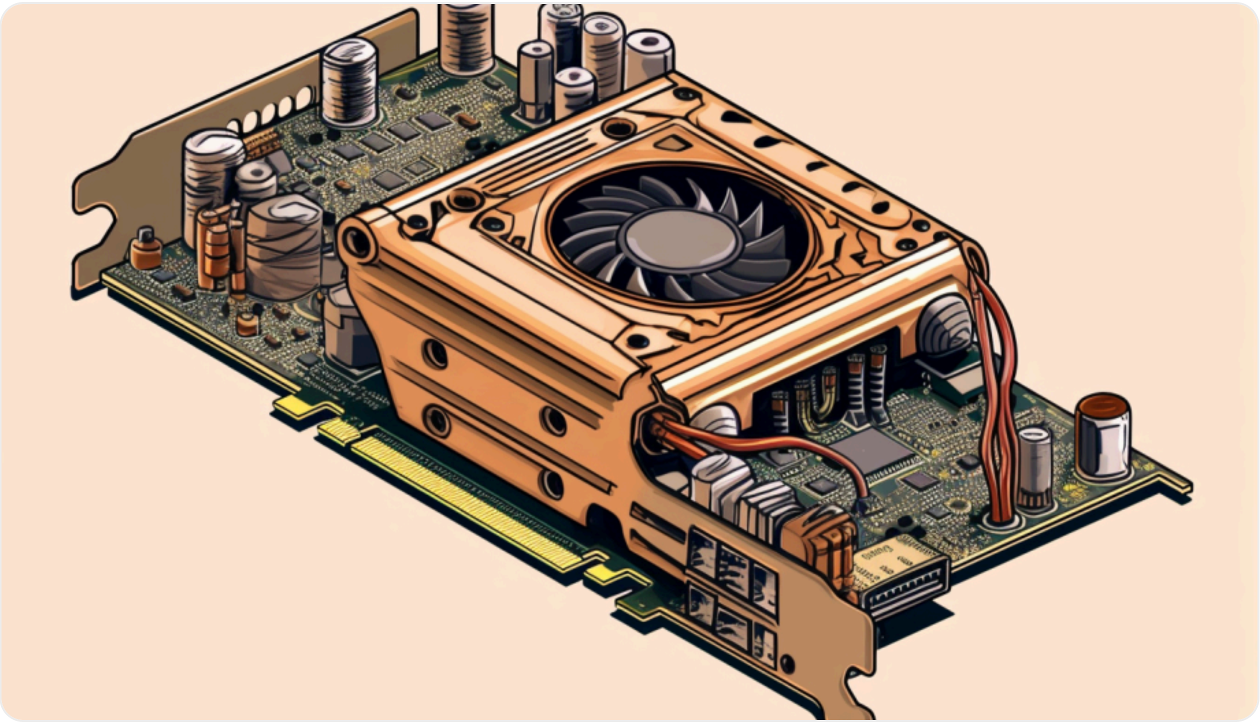


REPORT · 10 MIN READ

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

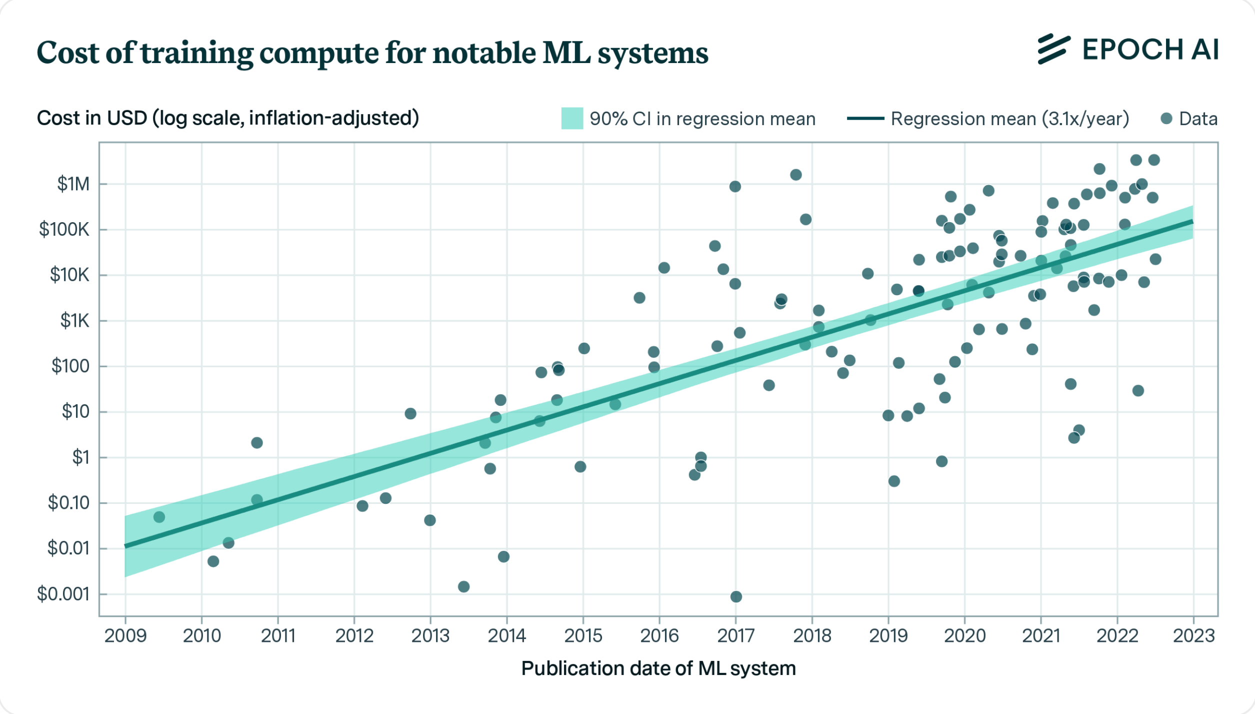


VIEWPOINT · 26 MIN READ

A Compute-Based Framework for Thinking About the Future of AI

AI's potential to automate labor is likely to alter the course of human history within decades, with the availability of compute being the most important factor driving rapid progress in AI capabilities.

May 31, 2023 · Updated Aug 10, 2023 · By Matthew Barnett



REPORT · 66 MIN READ

Trends in the Dollar Training Cost of Machine Learning Systems

I combine training compute and GPU price-performance data to estimate the cost of compute in US dollars for the final training run of 124 machine learning systems published between 2009 and 2022, and find that the cost has grown by approximately 0.5 orders of magnitude per year.

Jan 31, 2023 · By Ben Cottier

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.