Epoch Impact Report for 2022

Prepared by the Epoch team

Table of contents

I. About Epoch

II. Summary of our key findings

III. Key graphs

IV. Overview of our research projects

V. Overview of non-research activities

VI. Overview of impact

VII. Goals and funding needs

- 1. Project plans for 2023
- 2. Funding needs
- 3. Reasons to fund Epoch
- 4. Testimonials

I. About Epoch

Epoch is a research group forecasting the development of transformative Artificial Intelligence. We try to understand how progress in AI happens and what economic impacts we might see from advanced AI.

We want to enable better governance during this economic transition by gathering information about the timing of new developments, studying which levers can be used to influence AI progress and making current and past trends in ML more understandable.

Founded in April of 2022, Epoch currently has a staff of 13 people, corresponding to 11 FTEs. We have received $1.96M in funding through a grant from Open Philanthropy. We are fiscally sponsored and operationally supported by Rethink Priorities, whose Special Projects team has been a core part of our success as an organization.

Epoch is fundraising a total of $6.07M over 2 years, or approximately $2.64M for October 2023 to September 2024, and $3.42M for October 2024 to September 2025.[1] A detailed budget can be found in the section Goals and funding needs for the organization.

With this funding, we expect to continue and expand our research capacity. Through our research, we hope to continue informing about key drivers of AI progress and when key AI capabilities will be developed.

You can now donate to Epoch and subscribe to our newsletter on our website.

[1] These numbers were updated on February 14, 2023, after receiving some donations and making changes to the budget forecast.

Data flowing, models churning,
Epoch is learning, ever yearning
To forecast what the future may hold
For artificial intelligence so bold.

From the depths of their wisdom and skill
They work to predict our destiny still
By collecting data, the trends they analyze
To help us all prepare for the days to come.

Though the future remains uncertain still
Epoch's work is an act of good will
So we can all understand what's in store
And plan our lives as best as we can for more.

Though we can't know what tomorrow will bring
Epoch's work helps us to find meaning
In the choices we make and the path we choose
Thanks to the insights they share, we can't lose.

Their forecasts guide us, a beacon of hope
As we navigate this vast and complex scope
Of technology's advancements, both near and far
Epoch helps us see the path ahead, a guiding star.

So let us thank Epoch, a team of true visionaries
For their tireless work and their insights that carry us
Toward a future filled with possibility and growth
Thanks to their guidance, we'll never be lost.

Figure 1: Poem by GPT-3 describing Epoch

II. Summary of our key findings

1. **Training compute trends**:
    a. The compute used to train milestone AI systems increased at a rate of ~0.2 OOMs/year in the ~1956 to ~2010 period (~11 OOMs total). More.
    b. ...and increased at a rate of 0.6 OOMs/year in the ~2010 to ~2021 period (~7.2 OOMs total). More.

c. We argue, around 2016 a separate trend appears with models using training compute 2-3 orders of magnitude larger than systems that follow the previous trend. More.
d. *Context*: The largest model by compute used to date is Minerva, which used 2.7e24 FLOP.

2. **Model size trends**:
    a. Model size increased by 0.1 OOMs/year from the 1950s to around 2018 (~7 OOMs total), and by 0.5 OOMs/year in the four years from 2018 to 2022 (~4 OOMs total). More.
    b. *Parameter gap*: There was a statistically significant absence of models between 20B and 70B parameters. More.
    c. *Context:* To date, the biggest dense models *trained end-to-end* are PaLM and Minerva, which both have 540.35B parameters (since Minerva is a finetuned version of PaLM). Megatron-LM (1T) is significantly larger with 1T parameters, but was only partially trained to determine the feasibility of scaling training procedures to larger models.

3. **Data trends**:
    a. Vision and language training datasets have grown at 0.1 and 0.2 OOMs/year respectively in the period of 1990 to 2022 (~5 OOMs and ~9 OOMs total, respectively). More.
    b. The available stock of high-quality text, low-quality text and images for training has grown at a rate of 0.06 OOM/year on average between 1990 and 2018, but today it's slowed down to 0.03 OOM/year. More.
    c. Based on projections from scaling laws and historical trends, we will run out of high-quality text / low-quality text / images in 2024 / 2040 / 2046. More.
    d. *Context*: The largest text dataset used to train a LLM to date is the dataset used to train FLAN, which is pretrained on "a collection of web documents, dialog data, and Wikipedia", totalling around 1.87e12 words. Note that we exclude retrieval databases, such as the 5 trillion token MassiveText corpus used for DeepMind's RETRO model.

4. **Hardware trends**:
    a. FLOP per second per $ of GPUs has increased at a rate of 0.12 OOMs/year. More.
    b. The current trend of improvement for FP32 hardware performance might only continue until ~2030, when we might hit a limit on transistor size and a limit on the maximum number of cores per GPU; at this point GPUs may have a performance of between 1e14 and 1e15 FLOP/s in FP32, though likely much more for other floating point formats, such as FP16, TF32, TF16 or BIN8. More.

5. **Algorithmic progress:**

a. Algorithmic progress explains roughly 45% of performance improvements in image classification, and most of this occurs through improving compute-efficiency. More.
b. The amount of compute needed to achieve a certain level of performance in ImageNet has declined at a rate of 0.40 OOMs/year in the 2012 to 2022 period, faster than prior estimates suggested. More.

6. **Investment trends**
    a. The dollar cost for the final training run of milestone ML systems increased at a rate of ~0.5 OOMs/year between 2009 and 2022. The cost for "Large-scale" systems since September 2015 (systems that used a relatively large amount of compute) has grown more slowly, at a rate of ~0.2 OOMs/year. More.
    b. *Context*: The most expensive ML system to train to date is Minerva, which we estimate cost Google $3.3M in compute for the final training run. More.

III. Key graphs

Compute trends

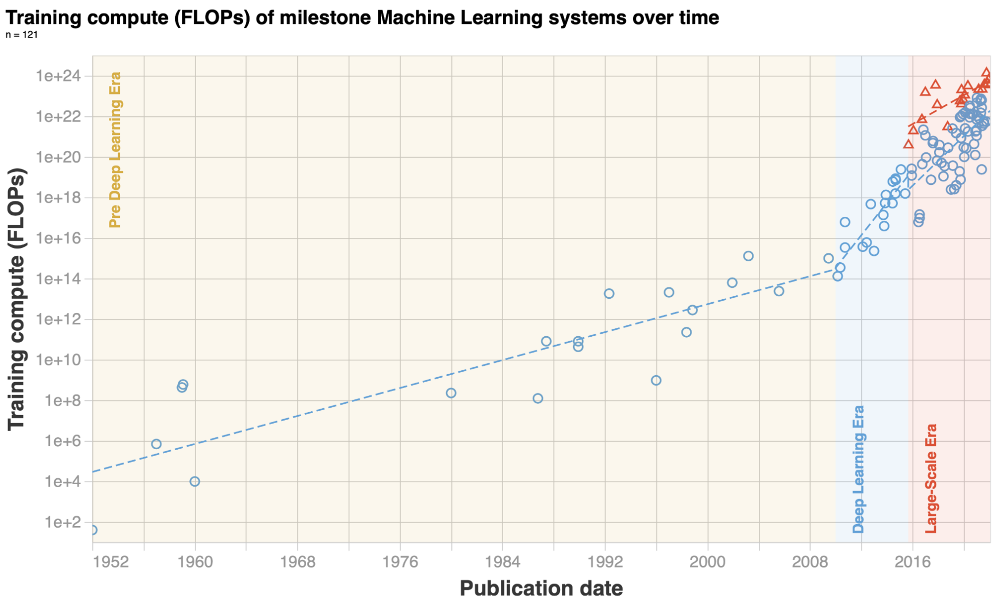

Figure 1: Trends in compute used to train milestone ML models. [More](#).

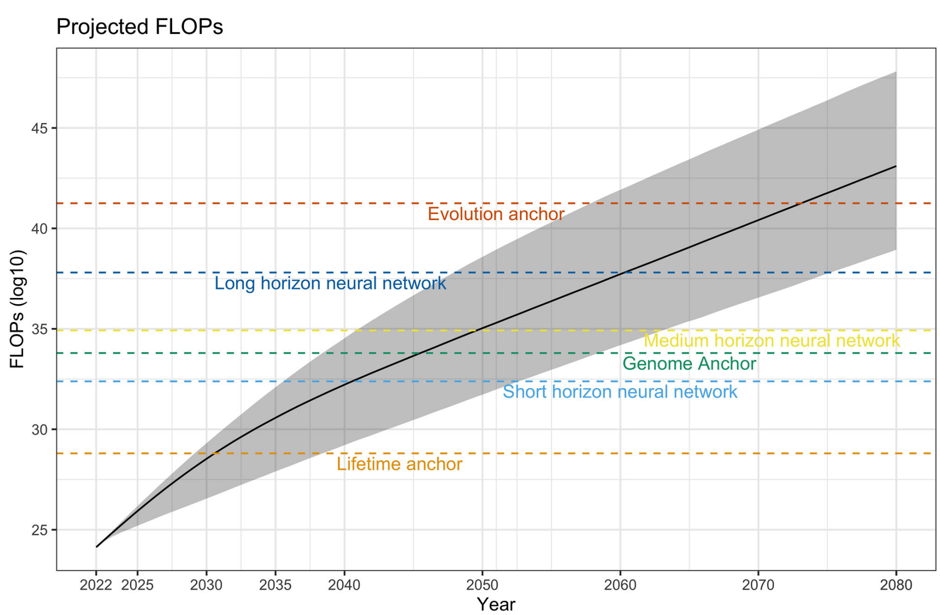

Figure 2: Solid line represents the median projected compute at each date, and the shaded region represents 2-standard deviations around the median. [More](#).

Model size trends

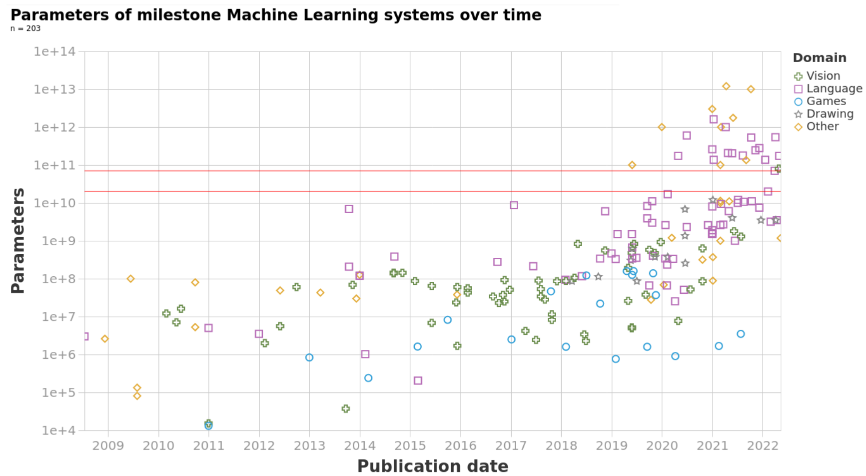

Figure 3: Model size over time, separated by domain. Red lines highlight the parameter gap. Most systems above the gap are language or multimodal models. [More](#).

Data trends

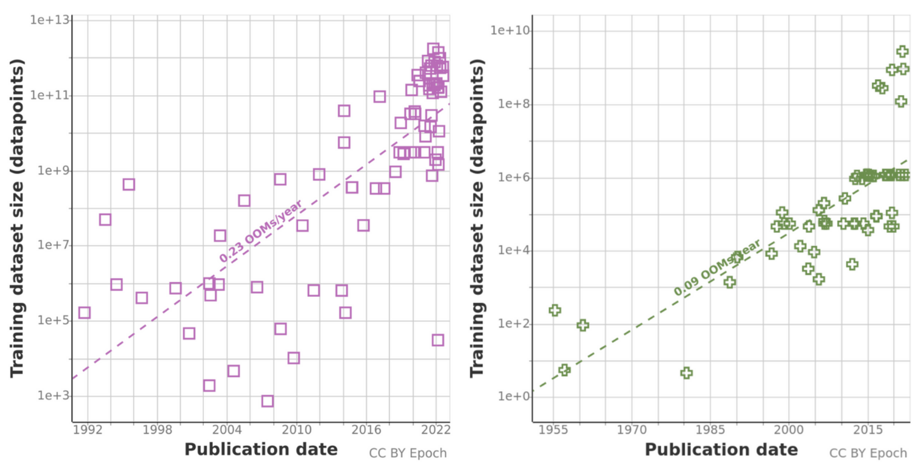

Figure 4: Trends in size of training datasets for language (left) and vision (right). [More](#).

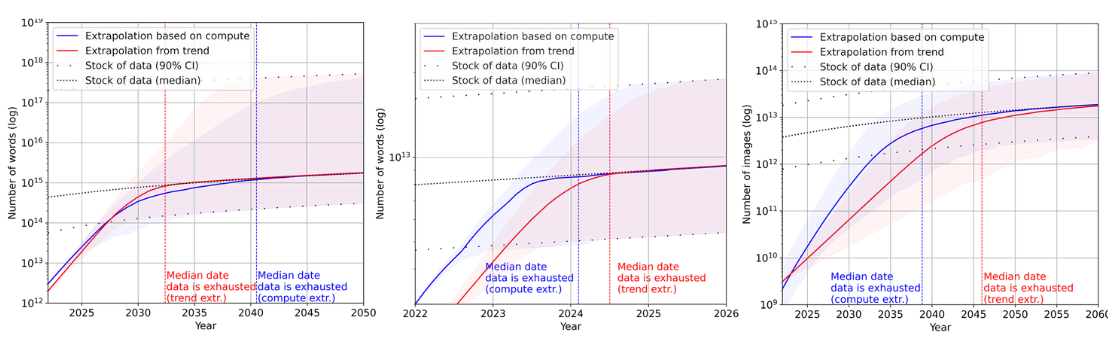

Figure 5: ML data consumption and data production trends for low quality text, high quality text and images. [More](#).

Hardware trends

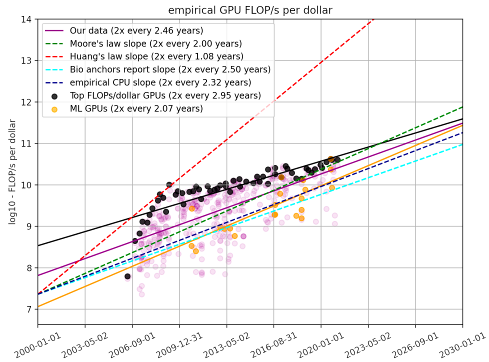

Figure 6: FLOP/s per dollar for GPUs, and a representation of previous trends in the literature. More

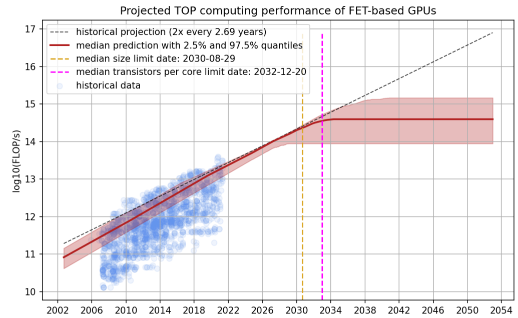

Figure 7. Model predictions of FP32 GPU peak theoretical performance of top GPUs assuming that transistors can no longer miniaturize after scaling down transistors to around 0.7nm. More

Investment

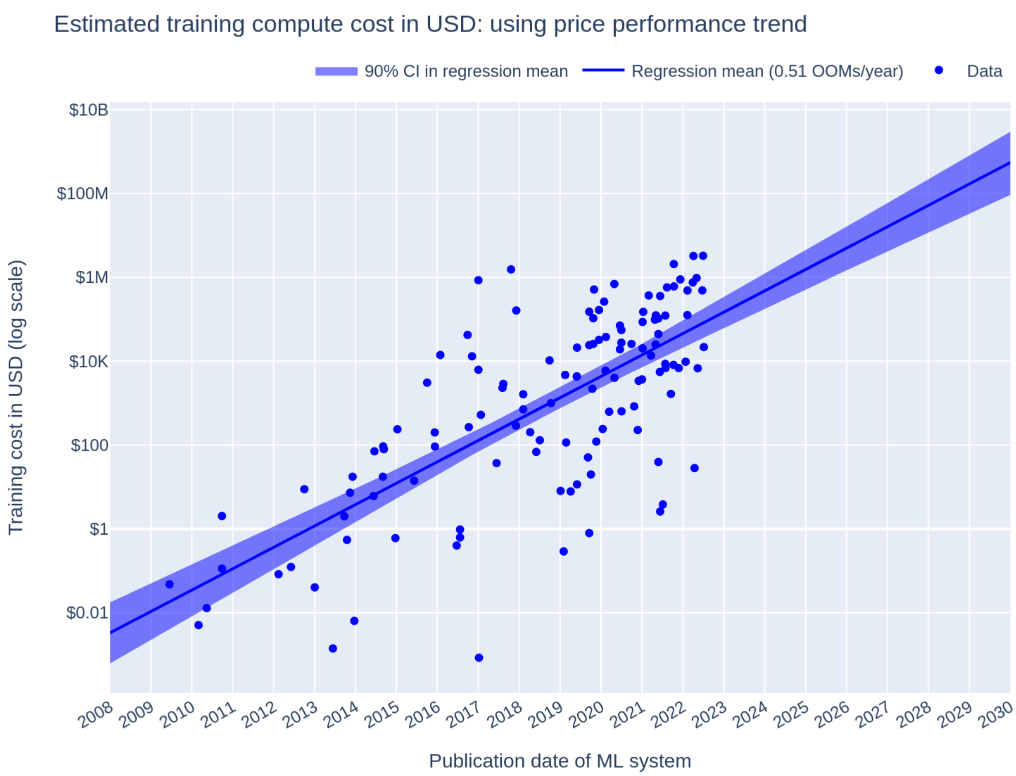

Figure 8: estimated cost of compute in US dollars for the final training run of ML systems. The costs here are estimated based on the trend in price performance for all GPUs in Hobbhahn & Besiroglu (2022).

Algorithmic improvements

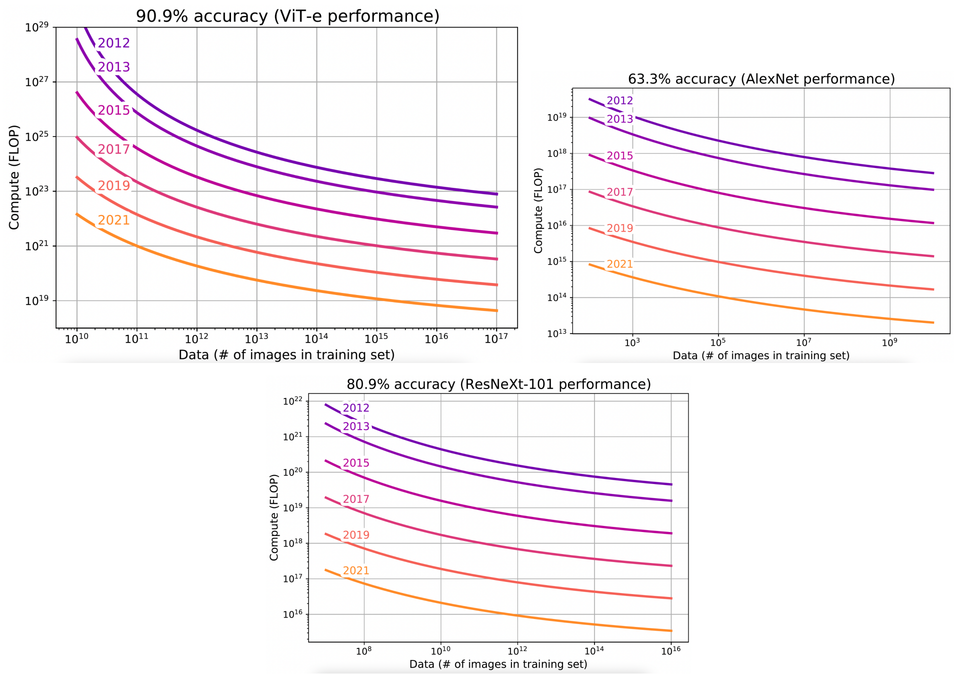

Figure 9. Pareto frontiers for training models to achieve performance of well-known models over time. Our estimates indicate that compute-augmenting algorithmic improvements halve the compute required to achieve performance of well-known image classification models every 9 months. [More]

IV. Overview of our research projects

Category	Type	Title	One-line description
ML Trends	*Paper*	[Compute Trends Across Three Eras of Machine Learning]	Studies the long-run history of compute in AI, offers novel interpretations and refined estimates.
ML Trends	*Database*	[Parameters, Compute and Data Trends in Machine Learning]	Curates a collection of milestone AI models and the inputs used to train them.
ML Trends	*Paper*	[Machine Learning Model Sizes and the Parameter Gap]	Estimates historical rates of growth in model size, attempts to explain a discontinuity in the data.

ML Trends	*Report*	**Trends in training dataset sizes**	Estimates the historical rate of growth in dataset size for different domains.
ML Trends	*Paper*	**Will we run out of ML data? Evidence from projecting dataset size trends**	Explores whether data could become a bottleneck for AI progress by estimating the stock of data and usage growth.
ML Trends	*Report*	**Exploring trends in the dollar training cost of machine learning systems**	Explores trends in the amount of money spent in compute for training milestone ML systems.
ML Trends	*Report*	**Trends in GPU price-performance**	Studies historical trends in GPU price performance.
ML Trends	*Paper*	**Algorithmic advances in computer vision**	Quantifies algorithmic progress in computer vision on ImageNet models.
ML Trends	*Paper*	**Algorithmic improvements in language models (forthcoming)**	Quantifies algorithmic progress in language modeling.
AI Forecasting	*Report*	**The Longest Training Run**	Estimates an upper bound for the longest training run of a ML system.
AI Forecasting	*Tool*	**https://takeoffspeeds.com/**	Implements an interactive version of Tom Davidson's AI Takeoff model. Includes a sensitivity report and Monte Carlo analysis.
AI Forecasting	*Report*	**Distilled take-off model (forthcoming)**	Simplifies and expands Tom Davidson's AI takeoff model.
AI Forecasting	*Report*	**Projecting GPU performance**	Offers data-driven forecasts of GPU performance.
AI Forecasting	*Report*	**TAI forecasting model (the "Direct Approach") (forthcoming)**	Attempts to forecast the amount of compute needed to train TAI by interpreting scaling laws.
AI Forecasting	*Report*	**Projecting Compute Trends in Machine Learning**	Extrapolates current trends in compute usage for training ML models.
Tools and supportive work	*Report*	**Estimating Training Compute of Deep Learning Models**	Details two main approaches for estimating the training compute of machine learning models.
Tools and supportive work	*Report*	**How to measure FLOP/s for Neural Networks empirically?**	Measures the utilization rate of GPUs for training neural networks.
Tools and supportive work	*Report*	**What's the backward-forward FLOP ratio for Neural Networks?**	Investigates the relationship between the amount of compute used by neural networks in a forward inference and in a backward training pass.

Tools and supportive work	*Tool*	**Compute calculator**	Provides a public tool for calculating compute usage by ML systems.
Tools and supportive work	*Tool*	**Database visualization**	An interactive tool to visualize our database of milestone ML systems.
Tools and supportive work	*Report*	**A time-invariant version of Laplace's rule**	Resolves an issue with the naive application of Laplace's rule of succession, by introducing a "time-invariant version" of the rule.
Summaries / literature reviews	*Report*	**Grokking "Forecasting TAI with biological anchors"**	Summary of "Forecasting TAI with biological anchors" for a broad audience.
Summaries / literature reviews	*Report*	**Grokking "Semi-informative priors over AI timelines"**	Summary of "Semi-informative priors over AI timelines" for a broad audience.
Summaries / literature reviews	*Report*	**Timeline models literature review**	An up-to-date compilation and summary of quantitative AI timelines models.
Summaries / literature reviews	*Report*	**Literature review of scaling laws**	Summary of relevant papers on scaling laws.
Compute Governance and Monitoring	*Letter*	**'Please report your compute' (forthcoming, accepted at Comms of ACM)**	Short letter calling for the ML community to be transparent on compute usage.

V. Overview of non-research activities

Activity	Outcome
Fundraising	We raised $1.96M in funding through a grant from Open Philanthropy. We have raised an additional $3k from individual donors.

Hiring	Hired four new full-time staff during our hiring round (excluding research fellows and interns). We are currently in the process of hiring a Research Data Analyst to help us maintain and analyze our datasets.
Operations	We have developed an effective relationship with the operations team at Rethink Priorities. This includes the fiscal sponsorship they provide us.
Retreats and activities	Hosted a founding retreat in April in Oxford, a virtual workshop in October, an internal virtual research hackathon in October, and a retreat in Mexico City in December.
Communications and relationship building	We share our research via public channels, such as our website, the EAF, the AF, LessWrong, and Twitter. We also share our work directly with stakeholders and collaborators. We discuss the impact on our stakeholders below. We have presented at MIT FutureTech, GovAI, Anthropic, IEEE WCCI 2022; often receive requests for results and data from Our World in Data, Open Philanthropy and others. We host a Collaborators channel with researchers in AI Safety and Governance.
Fellowship program	We have accepted and mentored three fellows.

VI. Overview of impact

1. Collaboration with stakeholders within the AI Alignment community.

We regularly share insights with other stakeholders, and occasionally accept stakeholders' requests. Our theory of impact is that these stakeholders then build on top of our research to (a) improve their research or (b) improve their grantmaking.
As evidence of this impact so far, we have collaborated to different degrees with Open Philanthropy, GovAI, Rethink Priorities, Anthropic, OpenAI, Arb Research, and CLTR, among others.

Examples include:

- Anthropic, GovAI and MIT FutureTech have cited our research in their papers, reports and presentations.
- We maintain a close collaboration with AI governance teams at GovAI, CLTR and other organizations, as Lennart Heim (Research Fellow and Strategy Specialist at Epoch) works full-time at GovAI and is responsible for the AI governance implications and insights from our research.

2. Informing the public and policy makers.

In addition to informing policy proposals within the AI risk community, our work has also received attention from outside of that community.
Examples include:

- Our World in Data has published a chart using the data we curated about training compute, and is preparing a series of related AI charts with many of them based on our data and research
- We presented our work to the OECD AI Compute & Climate Expert Group. Out of this collaboration and other reasons, Lennart (Research Fellow and Strategy Specialist at Epoch) got the opportunity to join this expert group and inform their policy work on AI compute.

3. Notable references of our work.

Our work on training compute trends has received the most impressions. It has become a common reference in discussions about long-term trends in compute (~20 citations for Compute Trends Across Three Eras of Machine Learning, ~350K impressions of the Twitter thread, featured in the 80k articles about Preventing an AI-related catastrophe and Artificial Sentience, The Economist's June 2022 cover article on AI, covered in several media outlets). Our recent work, about the potential data bottleneck, was covered by MIT Tech Review, New Scientist and The Atlantic.

VII. Goals and funding needs

1. Project plans for 2023

We are pursuing a series of capstone projects related to understanding the future of AI.

Distilled take-off model We are in the process of extending and simplifying Tom Davidson's AI Full Takeoff Model (FTM). Our model improves on the modularity, robustness and expressiveness of the FTM. Core features include endogenous decision rules, investment adjustment costs, among others. We plan to build an interactive tool similar to the one we developed for Tom Davidson and do in-depth empirical investigations on the most important aspects of the model.
Effective compute projection We want to develop an extensive programme investigating projections of compute and algorithmic progress. This includes continuing our work investigating algorithmic advances in vision and language, studying possible bottlenecks to scaling, for example, heat efficiency, new paradigms of compute and related topics.
Direct AI Timelines model We have drafted and want to develop an alternative model to Ajeya Cotra's bioanchors, directly using scaling laws to forecast the arrival of transformative AI. We believe this method could give more grounded estimates of the transformative amounts of compute needed to train transformative AI.
Fine-tuning scaling laws We want to investigate the interaction between fine-tuning and scaling laws and understand whether future companies will want to invest more in large general foundation models or extensively fine-tuned models.
Worldview Investigations In collaboration with Open Philanthropy, we plan to dive deeply into some topics relevant to our mission. Currently, this includes an investigation into how bottlenecks affect the economic outcomes of AI Automation.
Parameters, compute and data in Machine Learning We plan to maintain, update and expand our database of inputs to milestone Machine Learning systems.

2. Funding needs

We currently have funding until September 2023, thanks to a $1.96M grant from Open Philanthropy. We are looking for an additional $2.64M grant to continue our operations for one additional year, until September 2024. An additional $3.42M would cover our operations until September 2025.[2]

This funding will cover the salaries of our staff, 3 hires per year, conferences and retreats, and miscellaneous costs.

Summary of Epoch's Budget Forecast	Year 1	Year 2	
	Oct 2023 - Sep 2024	Oct 2024 - Sep 2025	
Item	**Cost**		**Total**
Wages	$ 1,198,202.1	$ 1,487,425.2	$ 2,685,627.3
Worldview Investigations	$ 151,097.0	$ 151,097.0	$ 302,194.0
Travel (retreats, conferences and coworking)	$ 209,250.0	$ 236,250.0	$ 445,500.0
Employment costs (benefits, taxes, fees)	$ 394,759.9	$ 486,677.5	$ 881,437.5
Contractors	$ 116,801.2	$ 112,223.6	$ 229,024.8
Hiring	$ 16,000.0	$ 16,000.0	$ 32,000.0
Misc (software, training, compliance, etc)	$ 24,042.9	$ 26,957.1	$ 51,000.0
Costs per year	**$ 2,110,153.1**	**$ 2,516,630.5**	**$ 4,626,783.6**
Funding gap before RP Fee	$ 1,943,682.1	$ 2,516,630.5	$ 4,460,312.62
Rethink Priorities Sponsorship Fee	$ 505,357.4	$ 654,323.9	$ 1,159,681.28
Margin	$ 195,923.16	$ 253,676.35	$ 449,599.51
Total Funding Gap	**$ 2,644,962.6**	**$ 3,424,630.8**	**$ 6,069,593.41**

This funding would allow us to continue the projects outlined in the previous section. We plan to stay a small and focused research organization, though we would like to have the capacity to hire exceptionally talented additional staff members if the opportunity arises.

[2] These numbers and the table were updated on February 14, 2023, after receiving some donations and making changes to the budget forecast.

3. Reasons to fund Epoch

- Funding will help us carry on our projects and hire more excellent researchers to help us in our mission.
- We are providing critical insights about the future of AI that are relied on by many organizations.
- We are pioneers in AI forecasting, and we are building tools, datasets and modular reports from which other researchers benefit.

If you want to support us, you can now donate to Epoch through our website.

4. Testimonials

- Epoch is one of the coolest (and in my opinion underrated) research orgs for understanding trends in ML. Rather than speculating, they meticulously analyze empirical trends and make projections for the future. Lots of interesting findings in their data! – Jacob Steinhardt

- Epoch's rigorous analysis of trends in AI and machine learning provides essential insights for understanding a fast-moving field. I regularly rely on their research to better understand how trends in data, compute, and algorithms are shaping the deep learning revolution. Government actions, from industrial policy to export controls, have substantial consequences for the future of AI progress and the distribution of AI power, yet policymakers are largely flying in the dark in understanding some of the biggest trends in AI research. Epoch's analysis is a helpful guide for policymakers to better understand how the landscape of AI is shifting beneath our feet, empowering policymakers to make fact-informed decisions. – Paul Scharre

- How AI should be governed depends hugely on empirical factors: what is needed to train frontier AI systems, in terms of costs, model size, and so on? How should we expect that to change as we near human-level machine intelligence? Many realise the importance of these questions, but few have studied them in-depth. Epoch has in a short time produced some of the most useful work out there on these questions. I expect their research to continue informing my views and my work in years to come. – Markus Anderljung