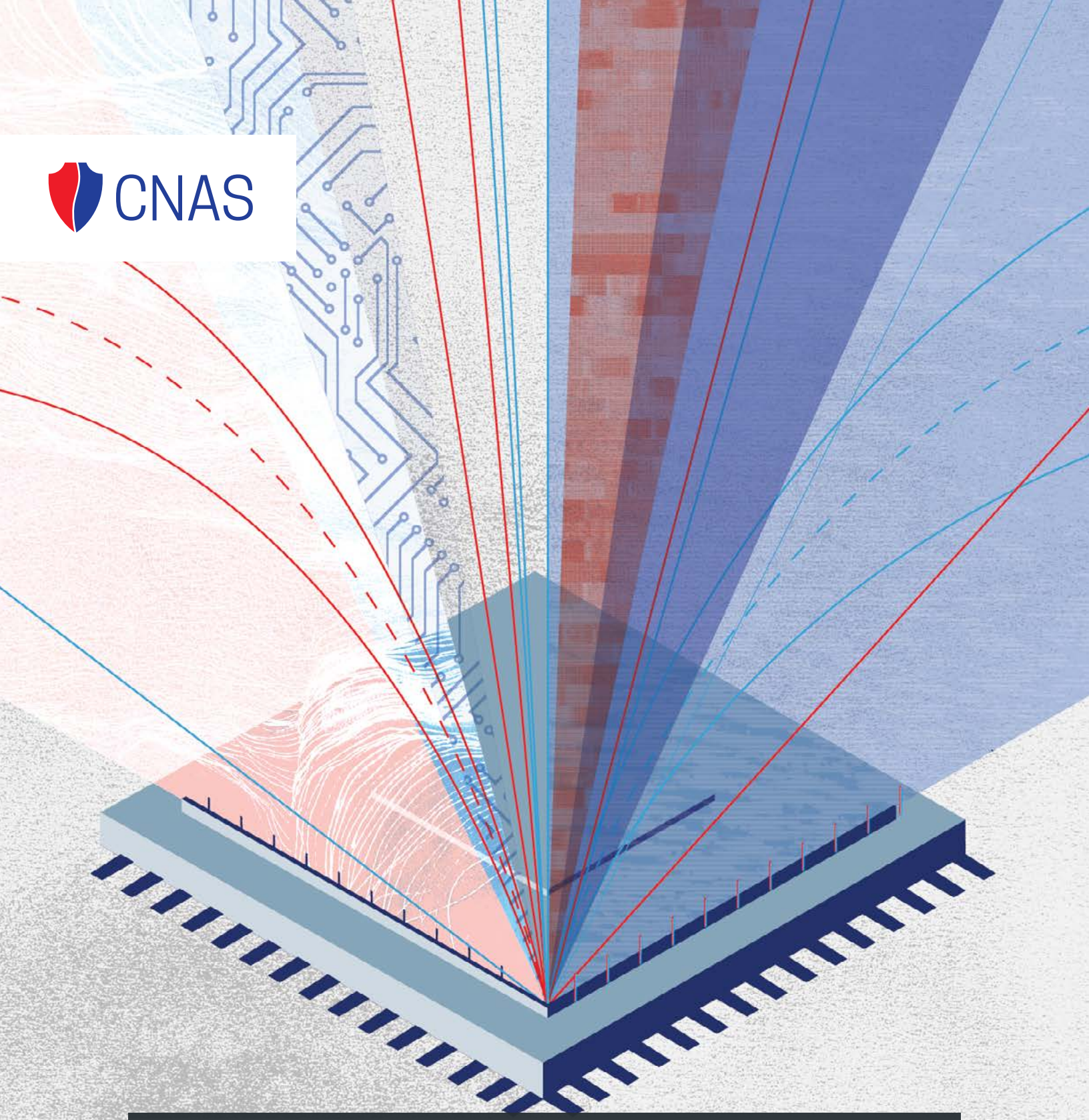


CNAS

MARCH 2024

# Future-Proofing Frontier AI Regulation

## Projecting Future Compute for Frontier AI Models

Paul Scharre

## About the Author

**Paul Scharre** is the executive vice president and director of Studies at the Center for a New American Security (CNAS). He is the award-winning author of *Four Battlegrounds: Power in the Age of Artificial Intelligence*. His first book, *Army of None: Autonomous Weapons and the Future of War*, won the 2019 Colby Award, was named one of Bill Gates's top five books of 2018, and was named by *The Economist* as one of the top five books to read to understand modern warfare. Scharre worked in the Office of the Secretary of Defense in the Bush and Obama administrations, where he played a leading role in establishing policies on unmanned and autonomous systems and emerging weapons technologies. He led the Department of Defense (DoD) working group that drafted DoD Directive 3000.09, establishing the department's policies on autonomy in weapon systems. He holds a PhD in war studies from King's College London, an MA in political economy and public policy, and a BS in physics, cum laude, from Washington University in St. Louis. Prior to working in the Office of the Secretary of Defense, Scharre served as an infantryman, sniper, and reconnaissance team leader in the Army's 3rd Ranger Battalion and completed multiple tours in Iraq and Afghanistan. He is a graduate of the Army's Airborne, Ranger, and Sniper Schools and honor graduate of the 75th Ranger Regiment's Ranger Indoctrination Program.

## About the Technology & National Security Program

The CNAS Technology & National Security program explores the policy challenges associated with emerging technologies. A key focus of the program is bringing together the technology and policy communities to better understand these challenges and together develop solutions.

## About the Artificial Intelligence Safety & Stability Project

The CNAS AI Safety & Stability Project is a multiyear, multiprogram effort that addresses the established and emerging risks associated with artificial intelligence. The work is focused on anticipating and mitigating catastrophic AI failures, improving the U.S. Department of Defense's processes for AI test and evaluation, understanding and shaping opportunities for compute governance, and understanding Chinese and Russian decision-making on AI and stability.

## Acknowledgments

I owe an enormous debt of gratitude to the researchers at Epoch, whose analysis this report relies upon. Their study of AI trends is a valuable resource for understanding the AI revolution, and I hope this report helps to further expose policymakers to their insights. This report builds on prior research and analysis by numerous scholars to whom I am indebted, including Dario Amodei, Markus Anderljung, Tamay Besiroglu, Tom B. Brown, Ryan Carey, Ajeya Cotra, Ben Cottier, Ege Erdil, Tim Fist, Lennart Heim, Danny Hernandez, Anson Ho, Marius Hobbhahn, Saif M. Khan, Andrew J. Lohn, Alexander Mann, Micah Musser, Jaime Sevilla, and Pablo Villalobos. I am especially grateful to Markus Anderljung, Tamay Besiroglu, Ben Cottier, Lennart Heim, and Jaime Sevilla for their valuable feedback on earlier drafts of this report. CNAS Research Associate Michael Depp provided valuable background research, feedback, editing, and assistance with citations and the bibliography. Thank you to CNAS colleagues Maura McCarthy, Melody Cook, Rin Rothback, and Anna Pederson for their roles in the review, production, and design of this report. Any errors are the responsibility of the author alone. This report was made possible with the generous support of Open Philanthropy.

# TABLE OF CONTENTS

## Executive Summary

Policymakers should prepare for a world of significantly more powerful AI systems over the next decade. These developments could occur without fundamental breakthroughs in AI science simply by scaling up today's techniques to train larger models on more data and computation.

The amount of computation (compute) used to train frontier AI models could increase significantly in the next decade. By the late 2020s or early 2030s, the amount of compute used to train frontier AI models could be approximately 1,000 times that used to train GPT-4. Accounting for algorithmic progress, the amount of effective compute could be approximately one million times that used to train GPT-4. There is some uncertainty about when these thresholds could be reached, but this level of growth appears possible within anticipated cost and hardware constraints.

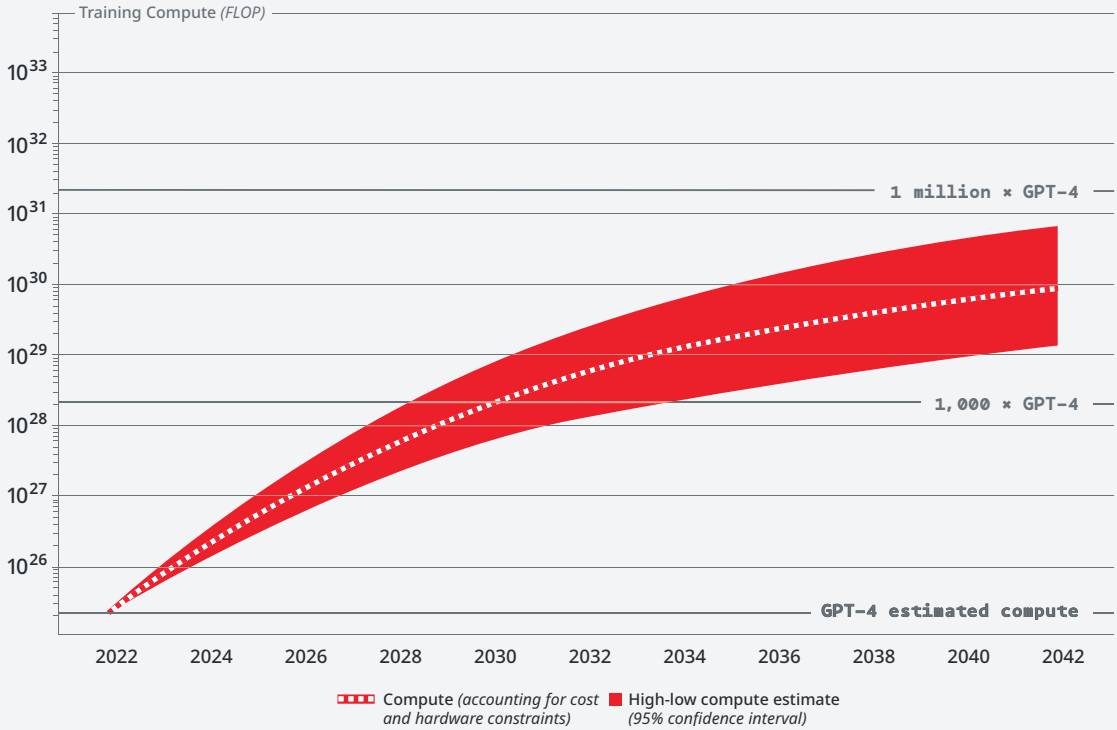
Improvements of this magnitude are possible without government intervention, entirely funded by private corporations on the scale of large tech companies today. Nor do they require fundamental breakthroughs in chip manufacturing or design. Increased spending beyond the limits of private companies today or fundamentally new computing paradigms could lead to even greater compute growth.

Rising costs to train frontier AI models may drive an oligopoly at the frontier of research, but capabilities are likely to proliferate rapidly. At present, algorithmic progress and hardware improvements quickly decrease the cost to train previously state-of-the-art models. Within five years at current trends, the cost to train a model at any given level of capability decreases roughly by a factor of 1,000, or to around 0.1 percent of the original cost, making training vastly cheaper and increasing accessibility.

The U.S. government has placed export controls on advanced AI chips destined for China, and denying actors access to hardware improvements creates a growing gap in relative capability over time. Actors denied access to hardware improvements will be quickly priced out of keeping pace with frontier research. By 2027, using older, export-compliant chips could result in a roughly tenfold cost penalty for training, if export controls remain at the current technology threshold and are maximally effective.

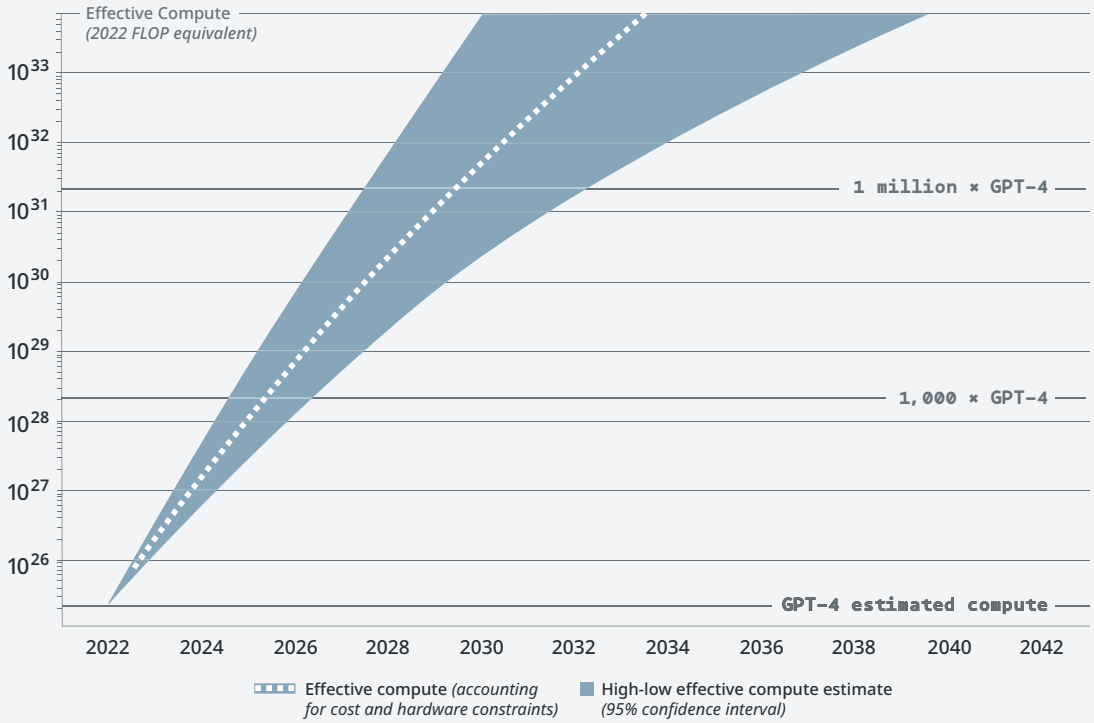
However, proliferation of any given level of AI capabilities will be delayed only a few additional years. At present, the cost of training models at any given level of AI capabilities declines rapidly due to algorithmic progress alone. If algorithmic improvements continue to be widely available, hardware-restricted actors will be able to train models with capabilities equivalent to once-frontier models only two to three years behind the frontier.

**COMPUTE USED TO TRAIN A FRONTIER AI MODEL RISES OVER TIME**



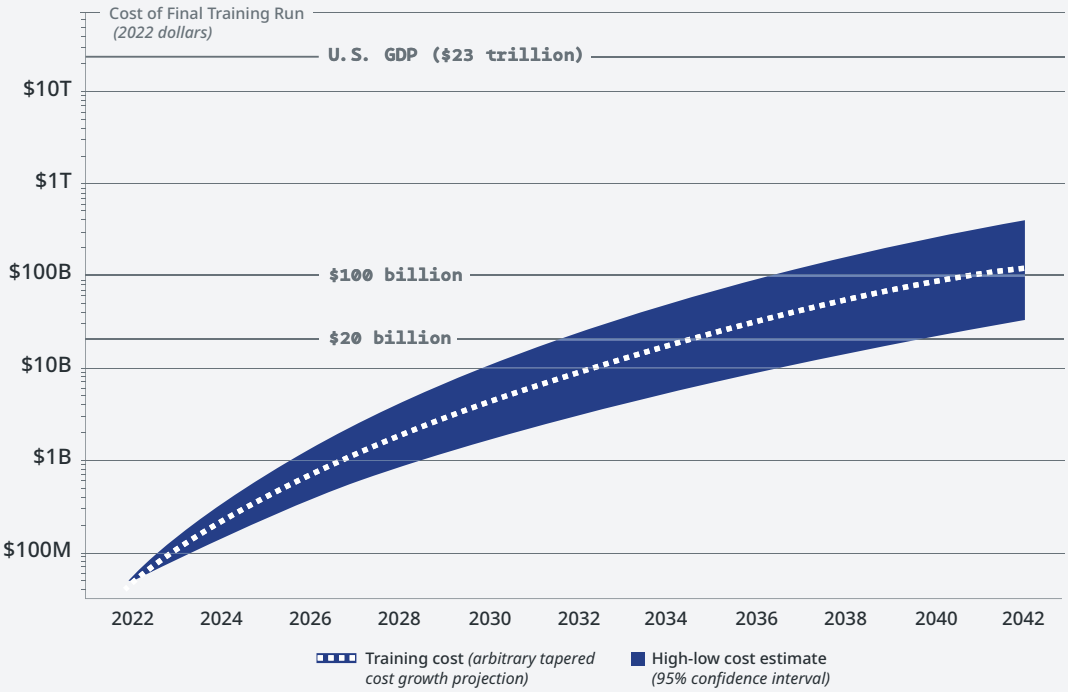

*The amount of compute used to train frontier AI models could increase to around 1,000 times GPT-4 by the late 2020s or early 2030s, even accounting for cost and hardware constraints.*

## WITH ALGORITHMIC IMPROVEMENTS, EFFECTIVE COMPUTE GROWS OVER TIME




*Accounting for algorithmic progress, the amount of effective compute used to train frontier AI models could be one million times GPT-4 by the late 2020s or early 2030s.*

## THE COST TO TRAIN A FRONTIER AI MODEL RISES OVER TIME




*The cost to train a frontier AI model is currently doubling approximately every 10 months. Cost growth is assumed to slow as costs approach the limit for private companies, currently in the tens of billions of dollars. In this projection, the cost doubling period is arbitrarily assumed to increase by 1.5 months per year, slowing the rate of cost growth.*

Access to compute and algorithmic improvements both play a significant role in driving progress at AI's frontier and affecting how rapidly capabilities proliferate and to whom. At present, the amount of compute used to train large AI models is doubling every seven months, due to a combination of hardware improvements and increased spending on compute. Algorithmic efficiency—the ability to achieve the same level of performance with less compute—is doubling roughly every eight to nine months for large language models. Improved performance comes from both increased compute and algorithmic improvements. If compute growth slows in the 2030s due to rising costs and/or diminishing hardware performance gains, future progress in frontier models could depend heavily on algorithmic improvements. At present, fast improvements in algorithmic efficiency enable rapid proliferation of capabilities as the amount of compute needed to train models at any given level of performance quickly declines. Recently, some leading AI labs have begun withholding information about their most advanced models. If algorithmic improvements slow or become less widely available, that could slow progress at AI's frontier and cause capabilities to proliferate more slowly.

While there is significant uncertainty in how the future of AI develops, current trends point to a future of vastly more powerful AI systems than today's state of the art. The most advanced systems at AI's frontier will be limited initially to a small number of actors but may rapidly proliferate. Policymakers should begin to put in place today a regulatory framework to prepare for this future. Building an anticipatory regulatory framework is essential because of the disconnect in speeds between AI progress and the policymaking process, the difficulty in predicting the capabilities of new AI systems for specific tasks, and the speed with which AI models proliferate today, absent regulation. Waiting to regulate frontier AI systems until concrete harms materialize will almost certainly result in regulation being too late.

The amount of compute used to train models is likely to be a fruitful avenue for regulation if current trends continue. Massive amounts of compute are the cost of entry to train frontier AI models. Compute is likely to increase in importance over the next 10 to 15 years as an essential input to training the most capable AI systems. However, restrictions on access to compute are likely to slow, but not halt, proliferation of capabilities, given the ability of algorithmic advances to enable training AI systems with equivalent performance on less compute over time. Regulations on compute will be more effective if paired with regulations on models themselves, such as export controls on certain trained models.

## Introduction

Policymakers and industry leaders have increased their attention on regulations for highly capable general-purpose AI models, sometimes called "frontier" models. Examples of current frontier AI models include GPT-4 (OpenAI), Claude 3 (Anthropic), and Gemini Ultra (Google). Companies already are training larger, more capable next-generation models using ever-larger amounts of data and computing hardware.

The computation used to train frontier AI systems is growing at an unsustainable rate. The amount of computation, or compute, used to train state-of-the-art machine learning models increased ten billionfold from 2010 to 2022 and is doubling every six months.[1] For the largest models, the amount of compute used for training is doubling approximately every seven months. This rapid increase in compute exceeds the pace of hardware improvements and is in part driven by increased spending on training. Costs for training the largest models are doubling roughly every 10 months.[2] Training current frontier models costs on the order of tens of millions of dollars just for the final training run. The full cost for training frontier models today, accounting for earlier training runs and experiments, could cost around $100 million.[3] As training costs continue to rise, they could reach hundreds of millions of dollars or even billions of dollars.

**Current trends point to a future of vastly more powerful AI systems than today's state of the art.**

In the near term, growth in large-scale training runs at AI's frontier is likely to continue. Leading AI labs already are reportedly training next-generation models or raising funds to do so.[4] Nvidia is shipping hundreds of thousands of new chips, which will enable more powerful future training runs. In the long run, however, cost and possibly hardware limitations are likely to constrain future compute growth.[5] The current exponential pace of compute growth cannot continue indefinitely. How long it continues, at what pace, and how much compute grows before leveling off has important implications for the future of AI progress. The role of cost and access to hardware as barriers to entry for training highly capable AI systems also has policy implications, such as for export controls and some regulatory proposals.

## Research Questions

This paper aims to answer several questions about how trends in cost and compute could affect the future of AI:

1. **Cost and compute projections:** If current trends were to continue, how would the amount of compute used to train frontier AI models and the cost of training rise over time? Accounting for algorithmic progress, how would the amount of effective compute increase over time?
2. **Limits on cost growth:** How much could compute increase before reaching the spending limits of private companies, and when would that occur? If the rate of cost growth slows as costs rise, how might that affect the amount of compute used for training frontier models?
3. **Limits on hardware improvements:** How might limits on continued hardware improvements affect future compute growth?
4. **Proliferation:** How might improvements in hardware and algorithmic efficiency affect the availability of AI capabilities over time?
5. **Costs for hardware-restricted actors:** How might constraints on hardware availability (for example, due to export controls) affect cost and compute growth for actors denied access to continued improvements in AI hardware?
6. **Compute regulatory threshold:** How might improvements in hardware and algorithmic efficiency impact the effectiveness of training compute as a regulatory threshold for frontier models over time?

The answers to these questions have important bearing on policy-relevant decisions today, such as the anticipated effect of export controls or other proposed regulations that would limit access to compute-intensive AI models. On the one hand, trends in rising costs are consolidating access to frontier AI models among a handful of leading AI labs. On the other hand, countervailing trends in hardware improvements and algorithmic efficiency are lowering barriers to capabilities, enabling proliferation. Some regulatory and policy interventions may be more or less feasible or desirable depending on how compute and cost change over time and the consequences for access to frontier AI models and the proliferation of capabilities. This paper aims to answer these questions with the goal of informing policymakers' understanding of possible scenarios for future AI development.

## Approach

Using current trends as a baseline, this paper projects cost and compute growth under various scenarios. The paper projects compute growth due to increased spending and hardware improvements. Additionally, it accounts for algorithmic improvements by projecting effective compute over time. The paper then estimates when training costs are projected to reach current limits for large corporations and move into the realm of what have historically been government-level expenditures. Additional scenarios explore how limits in hardware improvements may affect the availability of future compute. Since the cost to train a model with any given level of capabilities will decrease over time due to improvements in hardware and algorithmic efficiency, the paper also estimates how costs will decline over time, making capabilities more accessible to a wider array of actors, enabling greater proliferation. The paper then estimates how training costs change for actors that are restricted from continued improvements in AI hardware, such as U.S. government export controls on advanced AI chips destined for China. Finally, this paper estimates how future improvements in hardware and algorithmic efficiency may increase the accessibility of compute and capabilities relative to the U.S. government's notification threshold established in the October 2023 executive order. The paper concludes by assessing the policy implications of these projections.

# PART I: BACKGROUND

## Cost and Access to AI Models

Rising costs have important implications for determining who is able to access the most capable AI models. Compute costs are already an obstacle to academic researchers, who are priced out of training the largest, state-of-the-art models.[6] Training costs for current large language models are estimated to be in the tens of millions of dollars just for the final training run. Total costs, accounting for earlier training runs and experiments, could be around $100 million for the largest models to date.[7] Many frontier AI labs are backed by large corporations with deep pockets. Google DeepMind is owned by Alphabet. OpenAI secured a $10 billion investment from Microsoft. Google and Amazon have invested $2 billion and $4 billion, respectively, in Anthropic.[8] Academics and start-ups do not have the financial resources to compete at this scale. The U.S. government is launching a pilot program of the National AI Research Resource to provide compute and data resources to academics, although the effort is not yet fully funded.[9] If established, a national research cloud could help mitigate the effects of, but will not fundamentally alter, the trends driving increased training costs.

Current trends are pushing the frontier of AI research toward an oligopoly, where only a handful of well-funded actors can afford training the most capable AI models. Since 2019, many frontier AI labs have shifted increasingly to more limited release of their models, allowing other researchers or the public to interface with the model via an application programming interface (API) or only releasing models to a small number of vetted researchers. This trend toward the concentration of AI capability in the hands of a small number of corporate actors reduces the number and diversity of AI researchers able to engage with the most capable models. In response to this trend, many AI researchers and some leading companies, such as Meta, have pushed for more open-source releases of foundation models to help level the playing field and democratize AI capabilities to researchers who are not able to afford tens of millions of dollars to train their own foundation models.[10]

### Potential Harms

Research labs that have limited release of their models have pointed to potential harms from widespread proliferation.[11] Foundation models have well-documented problems of bias and toxicity due to their training data.[12] Large language models can be prompted to generate propaganda, hate speech, or misinformation at scale. Generative image models can be used to generate deepfakes and non-consensual pornography, including of minors. Generative image models also can show gender and ethnic/racial biases in how they present images, such as sexualizing female images, generating non-consensual nudes, or changing skin tone.[13] Additionally, large language models are known to "hallucinate" facts, presenting misleading information. The largest and most capable large language models have dual-use capabilities, including identifying vulnerabilities in computer code or scientific knowledge that could be used to enable chemical or biological attacks.[14]

**The largest and most capable large language models have dual-use capabilities.**

The accessibility of models has important implications for their potential to cause harm. For some generative models, researchers have developed filters or other safeguards to reduce the likelihood of the model generating harmful content. For example, Stability AI included in their model Stable Diffusion a content filter to prohibit the generation of harmful images.[15] Additionally, some generative image models have included embedded watermarking to make it possible to identify AI-generated images and reduce the likelihood of the model being used to generate misleading deepfake images. Current state-of-the-art large language models, such as OpenAI's GPT-4, Anthropic's Claude 3, Meta's Llama 2, and Google's Gemini Ultra, use fine-tuning with reinforcement learning to reduce their likelihood of producing problematic content.[16] When access to the model is limited via an API, model owners have some degree of control over the content the model produces, although many models nevertheless still generate concerning content.[17] Once models are open-sourced, however, they rapidly proliferate, and researchers can easily remove or disable filters. Following Stable Diffusion's release, researchers quickly disabled the content filter and removed the watermarking.[18] Similarly, open-source large language models can be easily fine-tuned at relatively low cost to remove safeguards. It took 19 hours of training at marginal cost to create to create an "uncensored" version of Llama 2 that was then posted online for anyone to download.[19] Open-sourcing models increases their accessibility to academic researchers and start-ups but also to malign actors who may use the models for harmful applications.

### Geopolitical Implications of Model Access

Given their dual-use capabilities, access to state-of-the-art models also has important implications for geopolitical and economic power. U.S. officials have expressed concern about the Chinese Communist Party's use of AI for military development, human rights abuses, and internal repression. In October 2022, the U.S. Commerce Department established export controls on advanced semiconductors and semiconductor manufacturing equipment destined for China, rules that have subsequently been updated and further refined.[20] The rules prohibit the export to China of the most advanced graphics processing units (GPUs) used for machine learning, even when the chips are produced outside the United States. U.S. export controls also limit the transfer of U.S. semiconductor manufacturing equipment, software, and tooling to Chinese semiconductor fabrication plants (fabs), restricting China's ability to advance domestic chip production. Japan and the Netherlands, leading producers of semiconductor manufacturing equipment, have adopted similar export controls on chip-making equipment and tooling. These export controls aim to deny Chinese firms the ability to access the most advanced AI chips, restricting their ability to conduct large-scale training runs.

**Once the model is open-sourced, it rapidly proliferates.**

U.S. officials have stated that the threshold for banned chips will remain constant over time, even as new chips are released.[21] If these controls are effective, over time they will widen the hardware gap between China and the rest of the world, as today's leading-edge chips become tomorrow's legacy chips. Without access to the most advanced AI chips, Chinese labs would face higher costs to train models and, in some cases, might be priced out entirely of accessing the largest and most capable models.

China continues to pursue efforts to grow its indigenous chip-making capabilities, however. Recent breakthroughs suggest that export controls may merely slow Chinese indigenous chip fabrication, not stop it completely. In September 2023, Huawei announced that its latest phone, the Mate 60 Pro, contained 5G technology powered by HiSilicon's new Kirin 9000S chip. Independent experts have assessed that the chip was produced using SMIC's 7 nanometer (nm) foundry, an advanced chip fabrication process restricted under U.S. export controls.[22] Questions remain about China's ability to produce advanced chips cost-effectively and at scale and how Chinese indigenous chip manufacturing will evolve over time.

Presently, Chinese labs can access state-of-the-art AI models open source, negating the effectiveness of chip export controls. Chinese labs do not need to train their own large foundation models if they can simply download trained open-source models directly from the internet. Safeguards on models can be easily fine-tuned away, undermining the U.S. government's efforts to keep advanced American AI technology from empowering the Chinese military or enabling Chinese government human rights abuses. The accessibility of future frontier AI models will have important implications for the effectiveness of U.S. export controls.

### Countervailing Trends That Increase Model Access

While compute costs are rising, countervailing trends in hardware improvements and algorithmic efficiency are reducing the costs for training a model at any given level of capability over time. As costs decline, an increasing number of actors can afford to train models with equivalent capabilities, enabling proliferation. At lower costs, it becomes increasingly likely that the cost of training a model will be affordable to an actor willing to open-source the model, as some companies such as Meta and Stability AI have done in the past. Once the model is open-sourced, it rapidly proliferates. Recent experience with generative language and image models suggests that the time lag from an initial breakthrough to an open-source version can be brief, as little as approximately seven months.[23]

## Implications for Policymakers

The economic and strategic value of current state-of-the-art AI systems, such as those based on large language models or multi-modal models, is highly uncertain. Some early studies have suggested that the most capable large language models today could be used to automate a significant portion of tasks currently done by white-collar employees.[24] Increasingly capable AI systems could be used to improve productivity and accelerate scientific discovery. They also could have dangerous dual-use applications, such as enabling the development of chemical, biological, or cyber weapons.

Policymakers face the difficult challenge of making AI models as accessible as possible for beneficial uses while restricting their access for harmful applications. This paper does not seek to resolve this dilemma. Rather, it aims to present policymakers with a greater understanding of how trends in AI progress may affect the accessibility of AI systems over time. Rising costs have

important implications for the quantity and diversity of AI researchers using state-of-the-art models, proliferation and potential harms, and geopolitical and economic power. This paper projects cost and compute trends under various scenarios to better understand cost growth and when different actors may be able to access models at different levels of computational power. Ideally, as a result of this analysis, policymakers will be better able to "future-proof" the policies they adopt today, taking into account exponential growth trends in cost, compute, and algorithmic progress.

## Understanding Cost and Compute Growth

The amount of computation, or compute, used to train machine learning models, and the associated costs for training, have been rapidly increasing during the deep learning revolution. This revolution pairs machine learning techniques, many of which date back decades, with increased computational performance that only became more recently available to train large artificial neural networks.[25]

Using machine learning, algorithms are trained on data using computing hardware.[26] The result is a trained model that is a representation of patterns in the underlying training data. This trained model has various applications, such as classifying new data or generating synthetic (AI-generated) data.

While machine learning theories are decades old, for many applications machine learning requires large amounts of computation in order to turn raw data into a useful trained model, and this only began to become available in the late 2000s. By refining machine learning techniques and combining them with advancements in computing hardware, algorithms, and increased data availability, scientists have generated significant advancements in computer vision, image generation, language processing, gaming, and other areas.[27]

Improvements in machine learning models can come from any of the three technical inputs into machine learning: the training data, the computing hardware used for training, and the algorithms used for training. Progress during the deep learning revolution, which began around 2010 to 2012, has come from improvements in all three technical inputs. Researchers, in fact, have found remarkably predictable "scaling laws" that capture the relationship between model performance and growth in model size (the size of the neural network), dataset size, and the amount of compute used to train a model.[28] These empirically derived scaling laws demonstrate that model loss—a measurement of model inaccuracy on test data—has an inverse relationship with model size, dataset size, and training compute. Larger models, datasets, and training compute lead to reduced model loss, or improved accuracy on test data. This negative scaling is remarkably smooth and is not affected as much by model architecture or other factors. In short, without any fundamental advances in the science of AI or understanding of intelligence, training larger models with greater amounts of compute and more training data yields improved performance.

AI researchers typically measure model performance using standardized benchmarks. The language model benchmarks Massive Multitask Language Understanding (MMLU) and Beyond the Imitation Game (BIG-bench) cover a diverse array of language-based tasks, from coding to learning U.S. history.[29] Owen (2023) found that while large language model performance on individual tasks was highly variable, aggregate performance on benchmarks showed "a fairly smooth relationship between overall performance and scale, consistent with an S-curve." The amount of compute used to train models, adjusted for optimal scaling, was a "fairly predictable" gauge of benchmark performance. Owen concluded, "This supports the idea that higher-level model capabilities are predictable with scale, and gives support to a scaling-focused view of AI development."[30] Over time, benchmarks eventually become saturated as models reach 80 or 90 percent accuracy and further attempts at improvement show diminishing returns.[31] AI researchers then simply invent new benchmarks to tackle harder problems.

**Training larger models with greater amounts of compute and more training data yields improved performance.**

AI researchers have used these scaling laws to train ever-larger models, with continued performance improvements. As of around 2021 to 2022, the most capable large language models had hundreds of billions of parameters and were trained on hundreds of gigabytes of data using thousands of advanced chips.[32] Leading AI labs such as OpenAI, Anthropic, and Google DeepMind more recently have begun to restrict publicly available information about their most capable models. Leaked information about GPT-4 suggests it is a very large model (an approximately 1.8 trillion parameter mixture-of-experts model), trained on a massive dataset (approximately 13 trillion tokens of training data),

using massive amounts of computing hardware (approximately 25,000 GPUs).[33] Even while public details about the most advanced models have become more scarce, indications are that leading AI labs are continuing to pursue ever-larger models.[34]

### Compute

The amount of computation used for training is a key metric for AI models. Compute is a function of the number of chips used for training, the type of chip, the hardware utilization rate, and the amount of time the chips are used for training. The total amount of compute used for training can be captured in a single metric measured in floating-point operations (FLOP).[35] AI researchers can increase the amount of compute used for training by using more chips, using better chips, or increasing the training time. AI hardware continues to improve, with machine learning GPU price-performance, or performance per dollar (FLOP per second per dollar), doubling roughly every 2.1 years. These hardware improvements alone would lead to greater compute over time as AI researchers use better chips for future training runs. However, AI labs also are increasing their spending, buying tens of thousands of advanced chips for large-scale training runs. Thus, growth in compute is a function of both spending more money on chips and hardware improvements that increase the computations per dollar that chips can perform.

$$\Delta \text{ compute (FLOP)} = \Delta \text{ expenditures (dollars)} \times \Delta \text{ hardware improvements (FLOP per second per dollar)}$$

Future compute growth is likely to be driven by both increased spending and improved hardware. This paper will project compute growth due to both factors, as well as estimating the effects of possible limits on spending and hardware improvements.

### Algorithmic Efficiency

Algorithms are also improving over time, reducing the amount of compute needed to train a model to the same level of performance. Gains in performance can come from increases in compute and/or increases in algorithmic efficiency, which allows compute to be used more effectively.

$$\Delta \text{ performance (effective compute)} = \Delta \text{ compute (FLOP)} \times \Delta \text{ algorithmic efficiency}$$

Algorithmic progress factors into the projections used in this paper in two ways: (1) to decrease the amount of compute needed to train a model at any given level of capability over time; and (2) to increase performance for frontier models.[36]

First, algorithmic progress decreases the amount of compute needed to train a model at any given level of performance over time. This effect has been measured for a variety of machine learning domains, including image classifiers, reinforcement learning algorithms, and large language models.[37] Improved algorithmic efficiency plays a role in increasing accessibility by lowering compute (and therefore cost) as a barrier to entry to training a model with any given level of capabilities. (Other factors, such as dataset availability, may still be a barrier for some actors.) As the amount of compute required to train a model at a given performance level decreases (in addition to compute per dollar increasing due to hardware improvements), costs decline over time, making previously unaffordable models accessible to a wider array of actors. The Proliferation section of this paper applies improvements in algorithmic efficiency and hardware performance to estimate how training costs decline over time for any given level of performance, enabling wider proliferation of capabilities.

Second, overall improvements in algorithms increase the performance of frontier models. Future frontier models will use more compute due to hardware improvements and increased expenditures on compute. Algorithmic improvements allow researchers to use this compute more effectively, leading to better performance. The combined effect of increased compute and algorithmic improvements is shown in this paper as "effective compute," which is represented in 2022 FLOP equivalent (the amount of effective computational power equivalent to FLOP in 2022).[38]

## Related Work

The amount of computing power used to train state-of-the-art machine learning models has exploded during the deep learning revolution. As researchers have collected data on this trend, they have sometimes used this historical data to generate future projections of compute and cost growth, including estimating how spending limits might affect compute growth. This paper builds on that prior analysis, updating projections using more recent historical estimates for compute and cost growth.

Amodei and Hernandez (2018) assessed that the amount of compute used for training the largest AI models increased 300,000-fold from 2012 to 2018,

**FIGURE 1.1 | COMPUTE GROWS DUE TO INCREASE SPENDING AND HARDWARE IMPROVEMENTS**

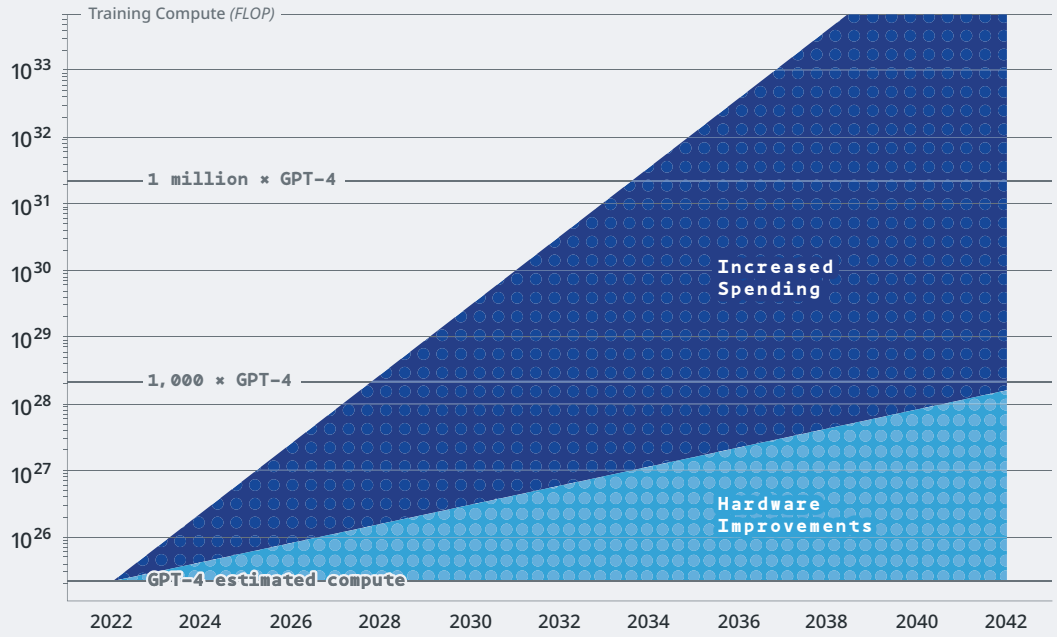
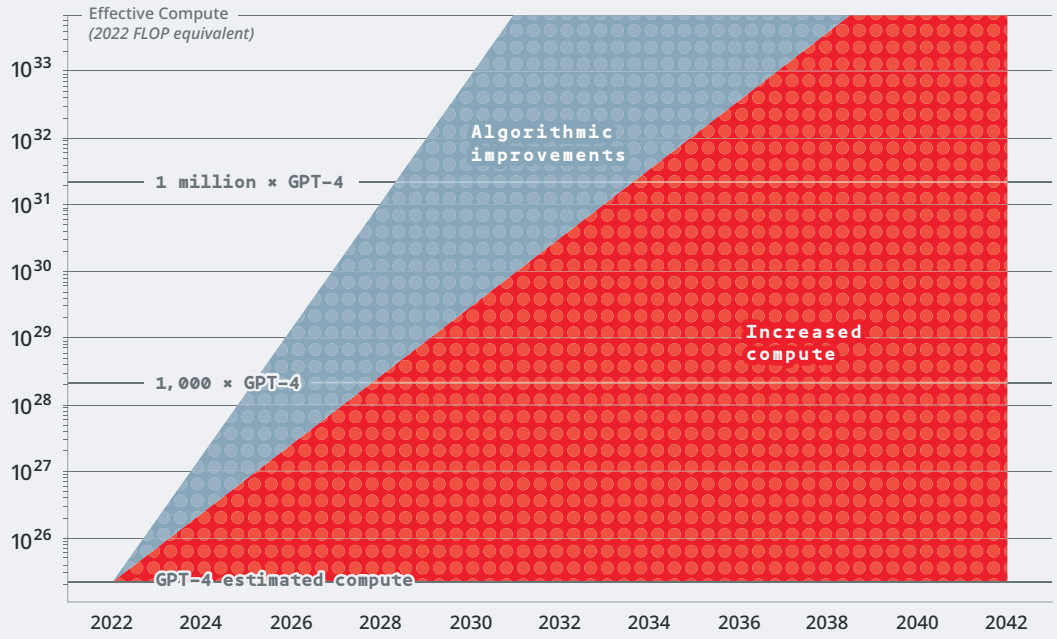



**FIGURE 1.2 | PERFORMANCE IMPROVES DUE TO INCREASED COMPUTE AND ALGORITHMIC IMPROVEMENTS**



doubling every 3.4 months.[39] While they assessed that cost eventually would limit compute growth, they estimated that this trend would continue in the short term. Subsequent analysis has attempted to more precisely project cost growth and when cost would become a limiting factor if compute were to continue to rise at current rates.

Carey (2018) estimated cost growth in training compute and determined that the current rate of compute growth was not sustainable beyond a few years.[40] Using a 3.5-month compute doubling period and estimating that the per unit cost of compute drops an order of magnitude every 4 to 12 years due to hardware improvements, Carey estimated that the cost of the largest training runs would increase an order of magnitude every 1.1 to 1.4 years. Carey assumed that the maximum spending capacity of private corporations was approximately $20 billion and for governments approximately $200 billion based on spending from the Manhattan and Apollo projects. Using a starting estimate of $10 million per training run in 2018, Carey estimated that training costs should reach the maximum spending limit for private corporations in 2021 to 2022 and for governments in 2023 to 2024. Carey further estimated that even if new developments in AI hardware cheapened compute by 1,000 times beyond current cost-compute trends during this period, this would only add another three to four years before cost growth reached the same limits. Carey concluded that the growth rate in compute costs could not continue beyond 3.5 to 10 years (an estimate derived in 2018).

Cotra (2020) estimated compute growth over time, noting that the current growth rate is "obviously unsustainable in the long run" due to rising costs.[41] Cotra estimated that by 2025 training costs would range from $300 million to approximately $1 billion. Cotra further estimated that by approximately 2040, tech companies could spend hundreds of billions of dollars to train an AI model if there were sufficient economic incentives to doing so. (This equates to a doubling in cost roughly every two years.) In the long run, Cotra estimated that governments' willingness to spend on training runs would cap out at approximately 1 percent of gross domestic product (GDP) for a major country, such as the United States. By assuming cost growth would taper over time, Cotra projected that costs could continue to rise for several decades, albeit at a slower pace than the current rate.

Lohn and Musser (2022) arrived at a similar conclusion to Carey (2018), determining that current growth rates in compute were not sustainable beyond two to three years.[42] Using the 3.4-month compute doubling period from Amodei and Hernandez (2018),[43] and assuming that compute per dollar doubles roughly every two to four years due to hardware improvements, Lohn and Musser estimated that compute costs, on their current trajectory, would eclipse total U.S. GDP by June 2026 or May 2027 at the latest.[44] Lohn and Musser further argued that in addition to cost, hardware availability and the engineering challenges associated with training massive models would limit compute growth. Lohn and Musser concluded that the 3.4-month compute doubling rate is not sustainable and suggested it already may be slowing.

Indeed, subsequent analysis of compute trends has found slower rates of compute growth. Sevilla et al. (2022) arrived at a revised estimate for compute growth based on observations of 98 state-of-art machine learning models from 2010 to 2022.[45] They updated Amodei and Hernandez's estimate and determined that compute doubled approximately every six months from 2010 to 2022. They further found that around late 2015, a new trend of large-scale models emerged, with large models using two to three orders of magnitude more compute than previous state-of-the-art models but growing at a slower rate, doubling approximately every 10 months. Sevilla et al. also estimated a 20-month doubling rate for historical machine learning models in the pre–deep learning era before 2010, similar to the 24-month doubling rate typically associated with Moore's Law.

## Improvements in algorithms can affect how efficiently models use compute.

Besiroglu et al. (2022) projected compute growth (but not cost) using the six-month doubling rate identified in Sevilla et al. (2022). Based on previous work by Carey (2018) and Lohn and Musser (2022), which concluded that the current growth rate in compute was unsustainable, Besiroglu et al. explored three different scenarios based on how long compute continued doubling every six months before reverting back to the pre–deep learning era doubling rate associated with Moore's Law. They estimated that the six-month doubling rate could continue for 8 to 18 years, depending on hardware improvements decreasing the cost of compute.

A major uncertainty in estimating cost growth trends is the rate of hardware improvement in compute per dollar. Hobbhahn and Besiroglu (2022) analyzed 470 GPUs released between 2006 to 2021 and found an approximately 2.5-year doubling period for floating-point operations per second per dollar (FLOP/s per dollar), a measure of performance per dollar for GPUs.[46]

Cottier (2023) developed updated estimates of cost growth based on these new measurements of compute growth and hardware improvements.[47] Relying on an updated version of the dataset on training compute for milestone AI systems in Sevilla et al. (2022), Cottier adjusted for improvements in hardware performance via two methods. First, by simply subtracting the overall GPU price-performance, or compute per dollar, trend identified in Hobbhahn and Besiroglu from the compute growth rate. And second, by using the actual price-performance of the GPUs used for training specific systems. Using the first method, Cottier identified a historical cost growth rate of 0.2 orders of magnitude per year (OOMs/yr), or a doubling of training costs approximately every 18 months. Cottier used this growth rate to project future costs, estimating that the cost of training the largest models could eclipse $200 billion by 2040.

Heim (2023) projected compute costs using a similar methodology, starting with an estimated $9 million to train PaLM in 2022.[48] Heim projected that at current trends, training costs would eclipse U.S. GDP in the mid-2030s and that this growth rate was not sustainable.

Subsequent analysis of hardware performance and compute growth allows further refinement of these estimates. Hobbhahn et al. (2023) assessed hardware performance growth using a dataset of nearly 2,000 GPUs from 2001 to 2021 and 47 machine learning accelerators (GPUs and other AI chips) from 2010 to 2023. They estimated a 2.5-year doubling period for FLOP/s per dollar for general GPUs and a 2.1-year doubling period for machine learning GPUs, revising earlier estimates by Hobbhahn and Besiroglu (2022).[49] Based on an updated dataset of 47 large-scale machine learning models trained from 2015 to 2023, Epoch (2023) estimated a 7.0-month doubling period for compute growth for large models, revising the estimates in Sevilla et al. (2022).[50] These new figures allow for updated cost growth estimates and projections, which this paper presents in Cost and Compute Projections.

Improvements in algorithms can affect how efficiently models use compute. Improved algorithmic efficiency can decrease the amount of compute needed to train a model at the same level of performance over time. Hernandez and Brown (2020) found that the amount of compute needed to train a model to the same level of performance on ImageNet reduced by 44-fold from 2012 to 2019, corresponding to a doubling in algorithmic efficiency for image classifiers on ImageNet every 16 months.[51] Erdil and Besiroglu (2022) similarly estimated algorithmic progress on ImageNet and arrived at a somewhat faster algorithmic efficiency doubling rate of every 8.95 months.[52] Approximately similar rates of improvement in algorithmic efficiency have been found in other machine learning domains. Dorner (2021) estimated algorithmic efficiency in deep reinforcement learning was doubling every 10 to 18 months on Atari games, every 5 to 24 months on state-based continuous control, and every 4 to 9 months on pixel-based continuous control.[53] More recently, Ho et al. (forthcoming) found algorithmic efficiency for large language models was doubling approximately every 8.4 months.[54] Grace (2013) found roughly similar rates of algorithmic progress across six different AI research fields.[55]

In projecting future compute resources, some analysts have included improvements in algorithmic efficiency that increase the amount of *effective* compute available to train models. Cotra (2020) and Hobbhahn (2022) factor this into their calculations of future compute, projecting the effective compute available for future projects. Cotra (2020) estimated algorithmic efficiency doubling every 2 to 3 years, while Hobbhahn (2022) estimated algorithmic efficiency doubling every 1.3 to 1.6 years.[56] Cotra used a more conservative estimate of algorithmic efficiency than the faster rate observed for ImageNet, under the assumption that researchers had "strong feedback loops" for ImageNet but are likely to make slower progress when breaking ground on new models.[57] Other compute growth projections did not include algorithmic progress.[58]

In addition to compute and algorithmic progress, the availability of data is also a significant factor in scaling model performance. Recent assessments by Villalobos and Ho (2022) on the rate of growth in dataset size allow for estimates about how data availability may affect compute growth over time.[59] These are discussed in Appendix A: Additional Limitations on Compute Growth.

A summary of recent observed rates of growth in relevant metrics is shown in Appendix B: Observed Growth Rates.

## Current Best Estimates and Assumptions

This paper projects cost, compute, and effective compute for frontier AI models using the following best estimates:

### Compute Growth

Epoch (2023) assessed that compute used for training state-of-the-art machine learning models has been doubling every 6.3 months (95 percent confidence interval [CI]: 5.5 to 7.2 months) since 2010. It assessed

that for the largest models, training compute has been doubling every 7.0 months (95 percent CI: 5.7 to 8.6 months) since 2015.[60]

Since it is the cost for the largest models that is of interest, this paper uses the 7.0-month doubling period as the baseline assumption for compute growth.

### Hardware Performance

Chips used for training are improving over time, enabling better performance per dollar. Hobbhahn et al. (2023) estimated that price-performance for machine learning GPUs has been doubling every 2.1 years (95 percent CI: 1.6 to 2.91 years).

This paper starts with a baseline assumption that GPU price-performance continues to double every 2.1 years (25.2 months). This assumption will be changed in an excursion scenario that explores limits on hardware improvement.

### Cost Growth

Since few AI papers publish cost figures, cost growth has not been directly measured. However, the rate of cost growth can be estimated by calculating the difference between observed compute growth (doubling every 7.0 months) and improvements in compute per dollar (doubling every 2.1 years). This is the same methodology used by Cottier (2023), updated with the most recent estimates.[61] This yields an estimate of cost growth doubling every 9.7 months.[62]

This paper uses an estimate of training costs doubling every 9.7 months (95 percent CI: 7.3 to 13.5 months) for baseline projections for cost growth.[63] This assumption will be changed in excursions that explore a tapering rate of cost growth as costs approach the limits of private companies.

### Algorithmic Efficiency

Ho et al. (forthcoming) estimated that algorithmic efficiency for large language models is doubling every 8.4 months (95 percent CI: 5.3 to 13 months).[64]

Since large language models currently represent the largest, general-purpose AI models, this paper uses the 8.4-month doubling rate as the baseline assumption for improvements in algorithmic efficiency for any given model. This is used to project the amount of effective compute available for training future frontier models. It also is used to estimate improvements in algorithmic efficiency that decrease the amount of compute needed to train a model to any given level of performance over time, increasing the proliferation of capabilities.

### Current Costs

While many details are known about recent state-of-the-art models, training costs are rarely reported. Epoch (2023) built an extensive dataset of over 500 notable machine learning systems from 1950 to 2023.[65] The dataset includes, when available, parameter count, dataset size, training compute, and other relevant details. Most papers do not include cost. In some cases, costs can be estimated using the reported amount of training compute. Epoch (2023) estimated a cost of $50 million (90 percent CI: $30 million to $90 million) for the final training run for GPT-4.[66] For more discussion of some of the challenges in estimating compute costs, see Appendix C: Estimating Compute Costs.

This paper uses Epoch's estimate of $50 million to train GPT-4 as the baseline for projecting future cost growth. Appendix D, Uncertainty in Cost Projections, shows alternate projections using different starting cost estimates. Cost projections are relatively insensitive to changes in the initial cost. A twofold change in the initial cost in either direction only leads to a twofold change in projected cost at any point in time. By contrast, cost projections are highly sensitive to errors in the rate of cost growth, since they compound over time.

### Current Compute

Since this paper uses GPT-4 as the starting point for projections of training cost, it relies on Epoch's (2023) estimate of $2.1 \times 10^{25}$ FLOP (90 percent CI: $1.1 \times 10^{25}$ to $3.9 \times 10^{25}$ FLOP) used to train GPT-4 as the starting point for compute for a frontier model trained in 2022, the year that GPT-4's training was completed.[67]

GPT-4 is an outlier, using a full order of magnitude more compute than the median trend for large models at the time it was trained. The next publicly announced model to be trained on approximately the same or greater compute was Google's Gemini Ultra, which was announced in December 2023, and which Epoch estimated was trained on $8 \times 10^{25}$ FLOP.[68] Using GPT-4 as a starting point for the cost and compute projections leads to projections for the most expensive and compute-intensive models at any point in time, not the median for large models.

### Assumptions

Projecting cost and compute growth based on current trends makes several assumptions which may not hold up in reality. It assumes that current estimates of compute growth, hardware price-performance, and model cost are at least approximately accurate. It assumes as a baseline

that these trends continue at their current rate, although alternate projections factoring in limitations on cost and hardware improvements are also presented. It assumes no discontinuous progress in spending, hardware performance, and algorithmic efficiency. Many of these assumptions may turn out to be incorrect.

The baseline projections do not make any assumptions about which factor(s) limit the rate of growth of compute and costs. They simply project forward cost and compute based on current trends without an assumption of which factors are driving those trends. Many leading AI labs could afford to spend more on large-scale training runs today if they desired. The amount of compute used in large-scale training runs could be limited by several factors: available chips, the engineering challenges associated with networking together tens of thousands of chips to train a model, and sufficient amounts of high-quality data to train a large model efficiently. For some start-ups, in addition to cost, access to the human capital needed to efficiently orchestrate large-scale training runs may be a significant factor limiting their ability to train the most compute-intensive models.

The baseline projections do not assume that cost is a limiting factor in compute growth. Additional scenarios incorporate cost as a limiting factor. This paper assumes that large corporations likely could marshal on the order of tens of billions of dollars in annual training costs if there was a sufficient payoff for doing so. Beyond that level, spending would likely move into the realm only historically affordable by major governments. For a discussion on private sector spending limits, see Limits on Cost Growth.

The baseline projections also do not account for limits in continued hardware performance improvements. An excursion scenario in Limits on Hardware Improvements accounts for potential limits in hardware performance.

These projections are not intended as predictions of how cost and compute *will* grow over time, but rather as projections of what cost and compute growth would be if they were to continue on their current trajectories. A discussion of additional limitations on compute growth, including from hardware, data, or engineering challenges, is included in Appendix A: Additional Limitations on Compute Growth.

Using these current best estimates and assumptions, this paper answers the research questions posed at the beginning by projecting frontier model training cost and compute over time.

# PART II: ANALYSIS

## Cost and Compute Projections

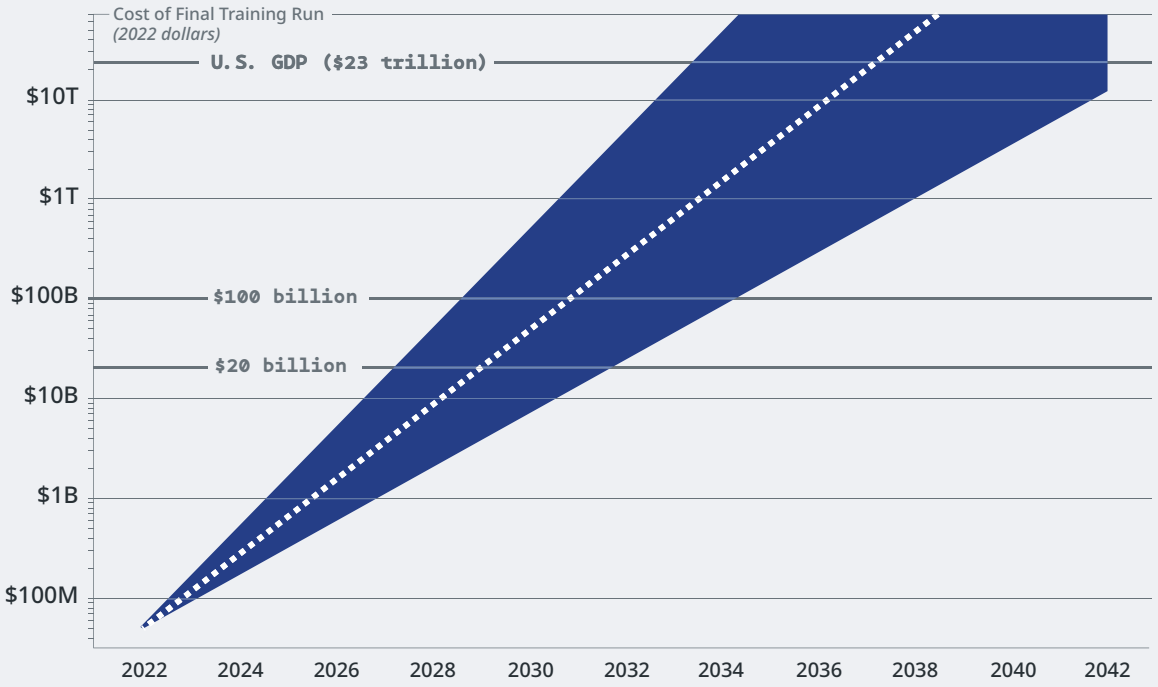
If current trends were to continue, how would the amount of compute used to train frontier AI models and the cost of training rise over time? Accounting for algorithmic progress, how would the amount of effective compute increase over time?

Figures 2.1 to 2.3 show a projection of current trends in cost, compute, and effective compute, respectively, according to the current best estimates for each growth rate. GPT-4 is used as a starting point, leading to a projection for the largest models at AI's frontier of research, rather than the median large model at any given point in time. All cost projections in this paper use 2022 constant dollars.

In Figures 2.1 to 2.3, the y-axes are logarithmic due to the exponential growth of each variable. Each tick mark on the vertical axis represents a tenfold increase in cost, compute, or effective compute. The straight lines indicate exponential growth curves. The uncertainties for each variable are shown in the shaded region, with the high and low estimates representing 95 percent confidence intervals. The dashed line for each variable represents the median estimate.

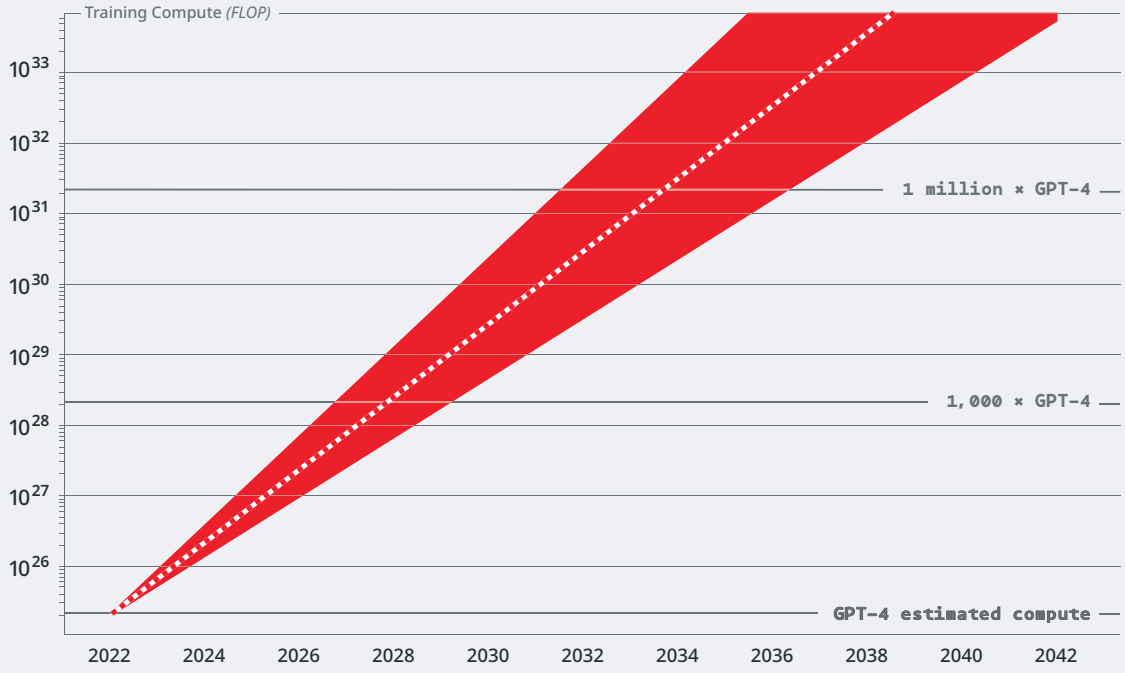
The uncertainties in these projections, shown in the shaded areas, are considerable. For a discussion of uncertainties in cost estimates and projections under different initial training cost estimates, see Appendix D: Uncertainty in Cost Projections.

**FIGURE 2.1 | THE COST TO TRAIN A FRONTIER AI MODEL RISES OVER TIME**
(STRAIGHTFORWARD PROJECTION OF CURRENT TRENDS)



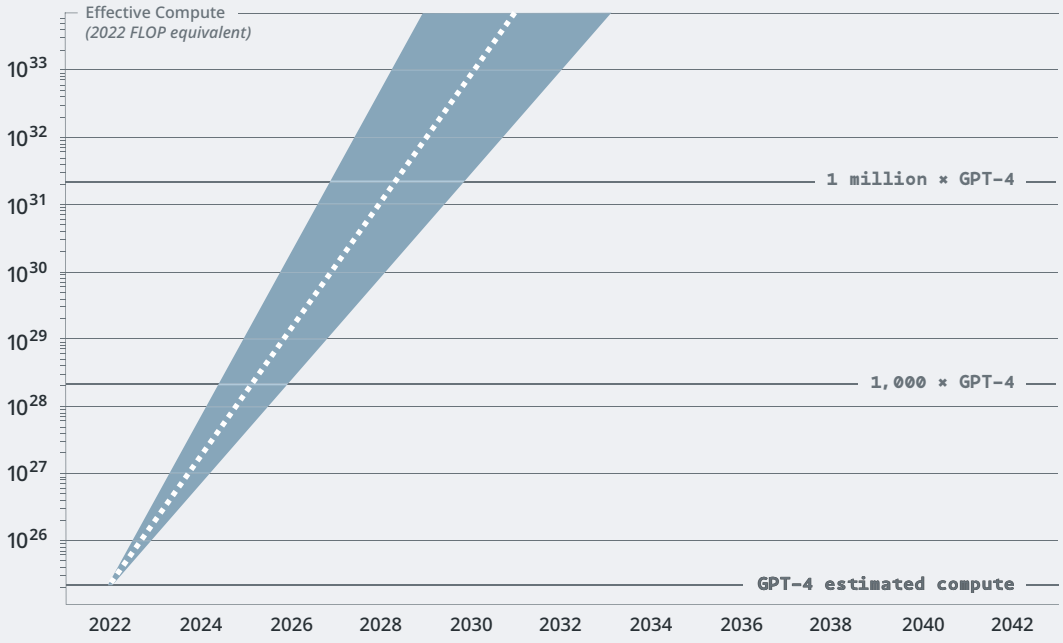

*The cost to train a frontier AI model is projected forward over time, starting from an initial estimate of $50 million to train GPT-4 in 2022, using a 9.7-month cost doubling rate (95 percent CI: 7.3 to 13.5 months).*

**FIGURE 2.2 | COMPUTE USED TO TRAIN A FRONTIER AI MODEL RISES OVER TIME**
(STRAIGHTFORWARD PROJECTION OF CURRENT TRENDS)




*Compute is projected forward over time, starting from an initial estimate of 2.1 × 10²⁵ FLOP to train GPT-4 in 2022, using a 7.0-month compute doubling rate (95 percent CI: 5.7 to 8.6 months).*

**FIGURE 2.3 | EFFECTIVE COMPUTE RISES OVER TIME**
(STRAIGHTFORWARD PROJECTION OF CURRENT TRENDS)




*The effective compute used to train frontier models is projected forward over time, starting from an initial estimate of 2.1 × 10²⁵ FLOP to train GPT-4 in 2022 and using a 7.0-month doubling rate for compute and an 8.4-month doubling rate for algorithmic efficiency (95 percent CI: 5.3 to 13 months).*[69]

**TABLE 1.1 | TRAINING COST AND COMPUTE OVER TIME**
(STRAIGHTFORWARD PROJECTION OF CURRENT TRENDS)

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **Cost of final training run in 2022 dollars (9.7-month cost doubling rate)** | $280M | $3.6B | $50B | $600B | $8T |
| **Training compute in FLOP (7.0-month compute doubling)** | $2.3 \times 10^{26}$ | $8.0 \times 10^{27}$ | $2.8 \times 10^{29}$ | $1.0 \times 10^{31}$ | $3.5 \times 10^{32}$ |
| **Compute relative to GPT-4** | 10 × | 380 × | 13,000 × | 500,000 × | 17 million × |

**Finding**
Training compute could increase approximately 1,000 times above GPT-4, to around $10^{28}$ FLOP, before reaching the current spending capacity of private companies (assumed to be in the tens of billions of dollars) in the late 2020s.

**TABLE 1.2 | TRAINING COST AND EFFECTIVE COMPUTE OVER TIME**
(STRAIGHTFORWARD PROJECTION OF CURRENT TRENDS)

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **Cost of final training run in 2022 dollars (9.7-month cost doubling rate)** | $280M | $3.6B | $50B | $600B | $8T |
| **Effective compute in 2022 FLOP equivalent (8.4-month algorithmic efficiency doubling)** | $1.6 \times 10^{27}$ | $1.1 \times 10^{30}$ | $7.8 \times 10^{32}$ | $5.4 \times 10^{35}$ | $3.7 \times 10^{38}$ |
| **Effective compute relative to GPT-4** | 80 × | 50,000 × | 40 million × | 25 billion × | 18 trillion × |

**Finding**
Accounting for algorithmic progress, the amount of effective compute used to train frontier models could increase to around one million times GPT-4, or approximately the equivalent of $10^{31}$ FLOP in 2022, before reaching the spending capacity of private companies in the late 2020s.

## Discussion

Compute could continue to grow several more orders of magnitude before reaching the current limit of training expenditures for major corporations, likely tens of billions of dollars.

Table 1.1 shows cost and compute over time, with compute represented both in absolute numbers (FLOP) and relative to GPT-4.

Accounting for algorithmic progress, the amount of effective compute used for training is even higher. Table 1.2 shows cost and effective compute, with effective compute represented both in 2022 equivalent FLOP and relative to GPT-4.

Current cost and compute trends are not sustainable. On their current trajectory, training costs will reach the current limits of private sector actors and move into what has been historically the realm of government-level expenditures in the late 2020s. However, compute could increase considerably before reaching that limit. The amount of compute used to train frontier models could increase on the order of 1,000 times GPT-4, or one million times in effective compute.

These straightforward projections do not account for how rising costs might affect the rate of cost growth. As costs begin to approach the historical spending limits of corporations, cost growth might reasonably slow. The next set of projections account for the spending limits of private companies in projecting cost and compute growth.

## Limits on Cost Growth

How much could compute increase before reaching the spending limits of private companies, and when would that occur? If the rate of cost growth slows as costs rise, how might that affect the amount of compute used for training frontier models?

Even the wealthiest actors have limits in their spending capacity. As costs continue to rise, it is reasonable to expect that these limits will affect compute growth.

Large corporations currently have a spending capacity in the tens of billions of dollars annually for research and/or capital expenditures. TSMC's capital expenditures were $36 billion in 2022 and estimated to be $32 billion in 2023.[70] Meta's capital expenditures were $32 billion in 2022, estimated to be $27 billion to $29 billion in 2023, and projected to be $30 billion to $35 billion in 2024. Meta's main driver for increased capital expenditures was an increase in AI capacity.[71] Amazon's capital expenditures were $59 billion in 2022 and estimated to be "slightly more than $50 billion" in 2023, although Amazon's figures include fulfillment and transportation costs.[72] Tech corporations have spent tens of billions of dollars on speculative research projects with no immediate return and no guaranteed long-term return. Meta spent $36 billion on metaverse research from 2019 to 2022.[73] Large corporations likely could marshal on the order of tens of billions of dollars in annual training costs if there was a sufficient payoff for doing so. Beyond that level, spending would move into the realm historically only affordable by major governments.

Governments historically have had significantly greater spending capacity than corporations and have pursued expensive technology projects when sufficiently motivated. The Manhattan Project to build the first atomic bomb cost the U.S. government nearly $2 billion at the time, the equivalent of more than $30 billion in 2022.[74] Moreover, it was willing to undertake such a large project alongside other major technology development efforts during World War II. The B-29 bomber program cost $3 billion in total, or the equivalent of approximately $45 billion in 2022, making it the largest military expenditure of the war.[75] The Apollo Program was even more expensive, costing nearly $3 billion per year during peak spending in the mid-1960s, the equivalent of $25 billion annually in 2022.[76] Costs in the range of tens of billions of dollars per year for strategically relevant technology projects are feasible for the largest governments.

Governments can marshal massive financial resources when necessary. The U.S. Defense Department's budget is over $800 billion today, in peacetime. At the start of the Cold War in the early 1950s, the United States spent over 12 percent of its GDP annually on national defense.[77] During World War II, defense spending rose to 36 percent of GDP. (For comparison, 36 percent of current U.S. GDP would be $9 trillion.) Governments also can dramatically surge spending in response to strategic needs. The U.S. government sent $48 billion in aid to Ukraine during the first ten months of the war in 2022.[78] The U.S. Defense Department spent over $2 trillion total in direct spending on the wars in Iraq and Afghanistan and other post-9/11 military operations.[79] In response to the COVID-19 pandemic, the U.S. government engaged in a swift and massive financial response, spending around $4.6 trillion in 2020 and 2021.[80] This paper does not explicitly model an upper limit on government expenditures, although costs in the hundreds of billions of dollars annually are possible for the largest governments. In extremis, a trillion dollars of annual spending is, in principle, possible.

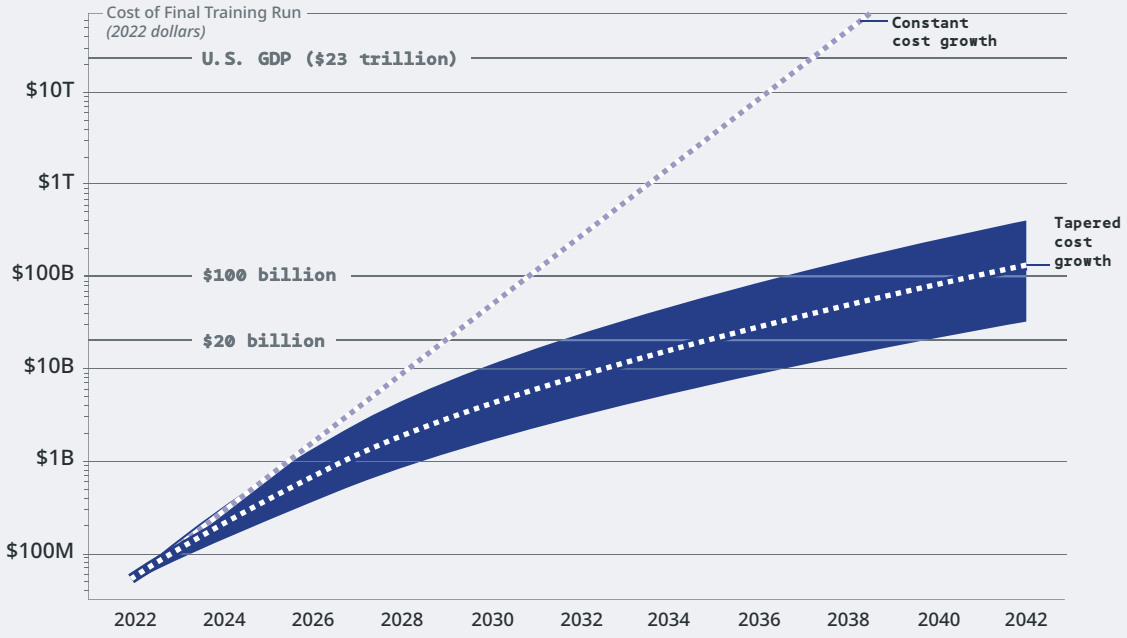
**Large corporations likely could marshal on the order of tens of billions of dollars in annual training costs.**

### Tapered Cost Growth Projection

Under a straightforward projection of cost growth, training costs will reach the historical limits of private companies in the late 2020s. A more sophisticated projection would account for cost growth tapering over time as costs rise. Rather than costs doubling at a rapid rate before abruptly stopping at some limit of maximum expenditures, a more reasonable assumption is that the rate of cost growth slows as training costs become increasingly large and begin to push the limits of affordability.

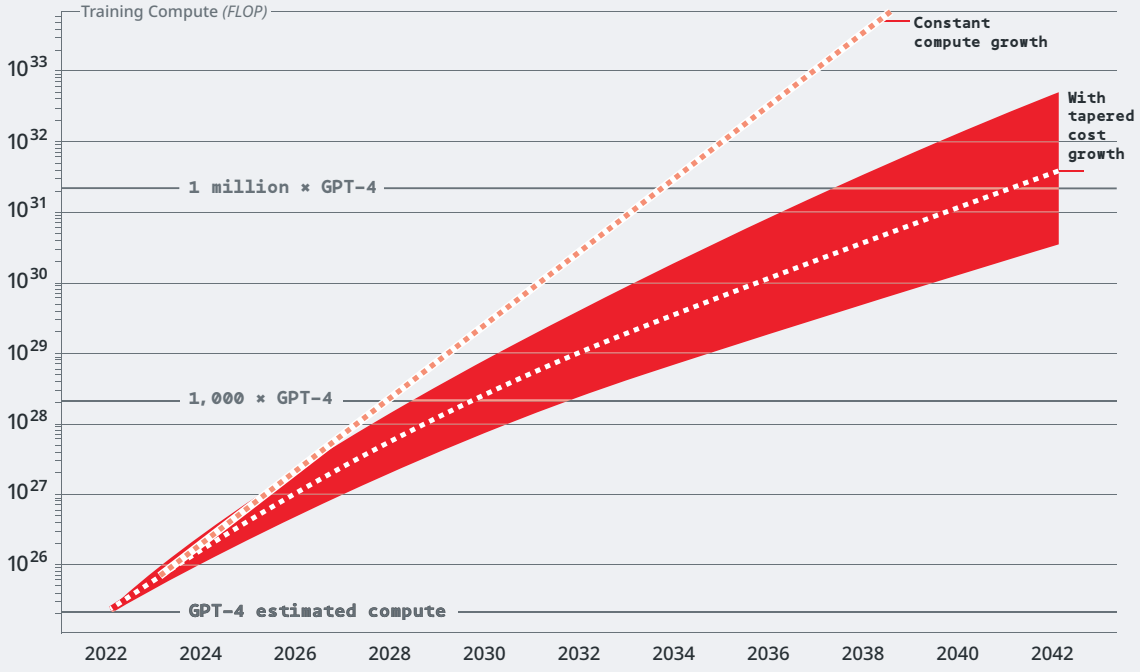
To illustrate the effect that this slowing rate of growth could have on training costs, a tapered cost growth projection is presented in which the cost doubling period is arbitrarily assumed to increase by 1.5 months each year. Under this projection, as costs reach $1 billion (by 2027), the rate of cost growth has slowed to a doubling every 17 months. As costs reach $10 billion (by 2033), the rate of cost growth has slowed to a doubling every 26 months (approximately 2 years). And as costs reach $100 billion (by 2041), costs are doubling every 38 months (3.2 years). Figure 3.1 shows cost growth under such a tapered cost growth projection, with the constant 9.7-month cost doubling rate shown for comparison.

**FIGURE 3.1 | FRONTIER MODEL TRAINING COSTS**
(ARBITRARY TAPERED COST GROWTH PROJECTION)




*Training costs are projected using a tapered cost growth model. Training costs initially double every 9.7 months, but the doubling rate is arbitrarily assumed to increase by 1.5 months per year, slowing the rate of cost growth. A constant 9.7-month cost doubling rate is shown for comparison.*

**FIGURE 3.2 | TRAINING COMPUTE**
(TAPERED COST GROWTH PROJECTION)




*The amount of compute used to train frontier AI models is projected using a tapered cost growth model. The doubling period for training costs is arbitrarily assumed to increase by 1.5 months per year, slowing the rate of cost and compute growth. A constant 7.0-month compute doubling rate is shown for comparison.*

Under this illustrative projection of a modest tapering of the rate of cost growth over time, costs still rise initially at a very rapid clip. The effect of this modest tapering of cost growth is that training costs do not exceed the available expenditures of large corporations until the mid-2030s. Under such a projection, training costs (and compute) continue to rise and remain within the realm of private sector actors for the next 10 to 15 years. Only until the mid- to late 2030s do costs begin to exceed the level currently affordable by large corporations and shift into the realm historically only possible by major governments. (For tapered cost growth projections under alternate scenarios of slowing cost growth, see Appendix E: Tapered Cost Growth Projections.)

This tapered cost growth projection is not presented as a prediction. The uncertainties in the rate of growth in compute are massive, and many future paths are possible. This projection is presented merely to illustrate, in a general sense, how a slowing rate of cost growth could affect compute projections. It is possible to envision reasonable trajectories in which training costs and compute continue to grow, albeit at a slower rate than today, for another 10 to 15 years.

Alternatively, costs could continue rising beyond the current estimate for the spending limit of private companies (tens of billions of dollars). This could occur in one of two ways:

First, governments could finance large-scale training runs. Governments could do this by establishing a government project to train large models, akin to the Manhattan Project or Apollo Program. This would be a significant shift from the role of governments in AI research today, but it could be possible if governments saw sufficient strategic value in pursuing next-generation models to warrant the expenditures. Alternatively, governments could provide financial support to a select group of companies training large models to help offset the cost, similar to how some governments support the construction of capital-intensive fabs today.

Second, private companies conceivably could fund training runs at higher levels beyond tens of billions of dollars if the revenue they generate from AI allows greater spending. Some researchers have taken into consideration the possibility of AI accelerating growth through increased productivity. Davidson (2023) has estimated how feedback loops on increasing investment and automation in AI research and development (R&D) could accelerate AI progress.[81] The projections presented in this paper do not take into account increased revenue for AI companies that could raise their spending capacity. Similarly, this paper does not account for the ability of AI itself to accelerate AI R&D, such as through improved chip design or algorithmic progress.

### Compute under a Tapered Cost Growth Projection

Under a tapered cost growth projection, in which the cost doubling period is arbitrarily assumed to increase by 1.5 months per year, compute grows more slowly. Figure 3.2 shows a projection of the amount of compute used to train frontier AI models under a tapered cost growth model. A straightforward projection of compute under constant cost growth is included for comparison.

### Discussion

The effect of slowing cost growth is to reduce the amount of compute used to train a frontier model at any given point in time. However, this merely delays by a few years the time to reach various compute thresholds.

Under a straightforward projection, training compute reaches $2 \times 10^{28}$ FLOP, or approximately 1,000 times above GPT-4, in 2028 at a cost of around $10 billion. Under a tapered cost projection, this milestone is delayed by only two years, arriving in 2030. In this projection, however, hardware performance continues to improve, somewhat making up for the lack of spending. This assumption may not be realistic and is removed in a subsequent scenario.

## Limits on Hardware Improvements

How might limits on continued hardware improvements affect future compute growth?

Chips are continuing to improve, which enables greater compute per dollar over time. Machine learning GPU price-performance, or performance per dollar, has been doubling about every 2.1 years. Hardware improvements may not continue indefinitely, however.

Physics-based analysis from Hobbhahn and Besiroglu (2022) suggested that GPU performance will stop improving between 2027 and 2035 as transistors approach the size of roughly a single silicon atom.[82] It is possible that chips continue improving through new techniques. Advanced packaging techniques, more specialized AI-specific chips, or entirely new computing paradigms could enable continued growth beyond the mid-2030s. However, one plausible scenario is that hardware improvements slow dramatically or stop

completely as chips reach fundamental physical limits. Under one scenario, these limits could be reached relatively quickly, in roughly the next three to ten years. Alternatively, even if these near-term anticipated limits do not materialize, other limiting factors could emerge as chips continue to improve. Ho et al. (2023) estimated fundamental limits in microprocessor energy efficiency could be reached in around 20 years, in the mid-2040s.[83] (For further discussion of hardware limitations on compute growth, see Appendix A: Additional Limitations on Compute Growth.)

There is a wide range of plausible scenarios for continued hardware improvements, ranging from hard limits in the relatively near term to continued improvements for another 20 years or more. As an illustrative example of how hardware performance limits could affect future training compute, this paper explores a scenario in which GPU performance stops improving between 2027 and 2035, per Hobbhahn and Besiroglu (2022). This is not presented as a prediction, but rather as an example of one plausible scenario for how future limits on continued hardware improvements could affect compute growth.

Figure 4.1 shows compute growth due to hardware improvements alone under two scenarios: (1) a straightforward projection of the current 2.1-year doubling rate for machine learning GPU price-performance; and (2) hardware improvements slowing beginning in 2027 and stopping completely by 2035. (The slowing in hardware improvements used in this projection results in a final FLOP/s per dollar level roughly equivalent to an abrupt stop in approximately 2031.)

This new projection of hardware improvements, slowing beginning in 2027 and stopping completely by 2035, can be used to update projections of training compute and effective compute, accounting for potential hardware limitations. Figures 4.2 to 4.4 project frontier model training cost, compute, and effective compute, respectively, accounting for hardware limits and tapered cost growth. The median estimates for cost, compute, and effective compute without any cost or hardware limits are shown for comparison.

### Discussion

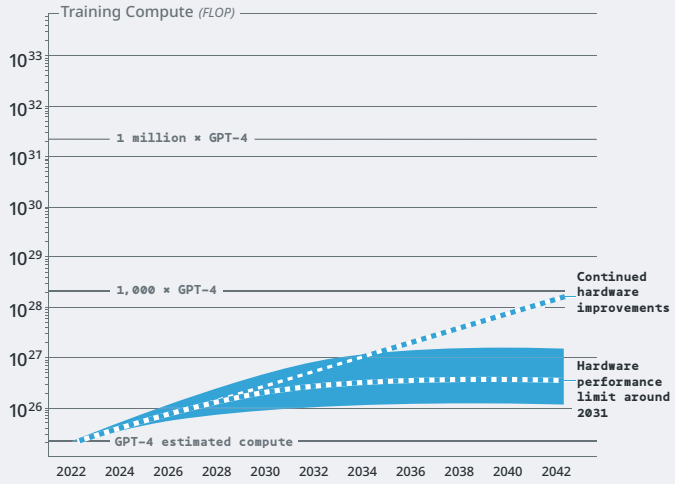
Under this projection, costs keep rising but compute gains are slower after around 2030 as GPU price-performance stalls. Effective compute continues to rise, however, since algorithmic improvements are assumed to continue. In this scenario in which cost growth slows and hardware performance gains stall, algorithmic improvements become the dominant driver of continued performance gains in the 2030s. This does not mean that compute ceases to be important or that the most capable models are no longer expensive. Huge costs and massive amounts of compute remain the price of entry for training the most compute-intensive frontier models. Merely, as compute growth from hardware improvements and increased expenditures slow, algorithmic improvements continue to increase the effectiveness of the available compute, enabling better performance. Tables 2.1 and 2.2 show compute and effective compute, respectively, over time assuming limits on hardware performance and a tapered cost growth projection.

These projections of effective compute assume that the rate of algorithmic efficiency improvements is constant, doubling every 8.4 months, and independent of compute growth. This assumption may not be valid. If growth in compute slows due to rising costs and/or diminishing hardware performance gains, researchers could focus more attention on algorithms, increasing the rate of algorithmic improvements. Alternatively, current rapid improvements in algorithmic efficiency could be partly a function of rising compute. Researchers

**Huge costs and massive amounts of compute remain the price of entry for training the most compute-intensive frontier models.**

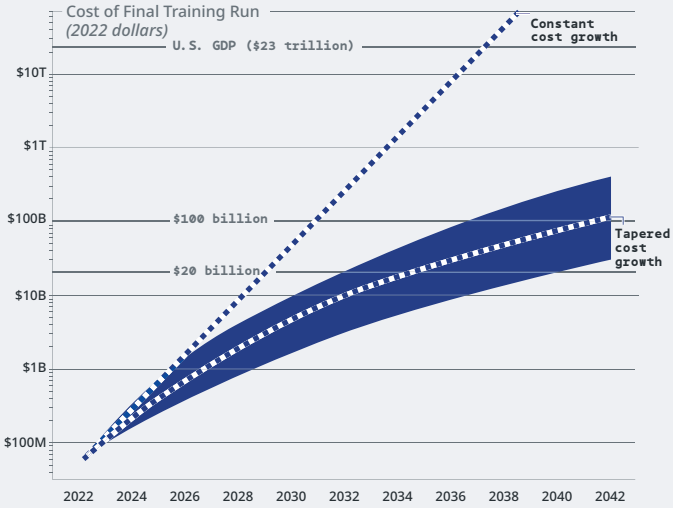
are able to achieve performance gains quickly by scaling compute, allowing fast feedback loops on model performance. Researchers then can build on these initial compute-driven gains by improving model efficiency through algorithmic improvements. If this is the case, then a slowdown in compute could lead to an accompanying slowdown in algorithmic progress. Additionally, some leading AI companies such as OpenAI, Anthropic, and Google have begun withholding details of their most advanced models. If the AI research ecosystem becomes more closed than it has been historically, and leading AI labs refrain from releasing model weights or even publishing details of their most advanced models, improvements in algorithmic efficiency for frontier models could slow or could be confined to within leading companies. This shift may be under way already. Government regulation also could slow the rate of algorithmic efficiency gains, at least for frontier models, if government regulations on the most compute-intensive models prohibit or delay their release.[84]

**FIGURE 4.1 | COMPUTE GROWTH DUE TO HARDWARE IMPROVEMENTS ALONE** (HARDWARE PERFORMANCE LIMIT AROUND 2031)



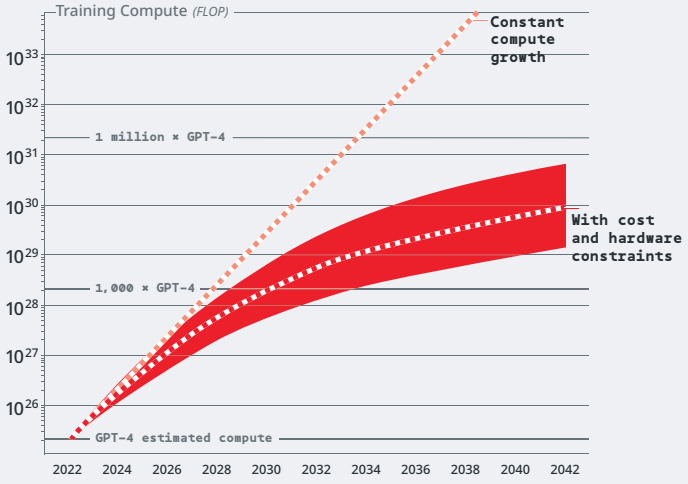

*Compute growth due to hardware improvements alone is shown under two scenarios: (1) a constant 2.1-year doubling rate for machine learning GPU price-performance; and (2) hardware performance gains slowing beginning in 2027 and stopping completely by 2035, resulting in a final FLOP/s per dollar level roughly equivalent to an abrupt stop in hardware performance improvements around 2031.*

**FIGURE 4.2 | FRONTIER MODEL TRAINING COSTS** (ARBITRARY TAPERED COST GROWTH PROJECTION)



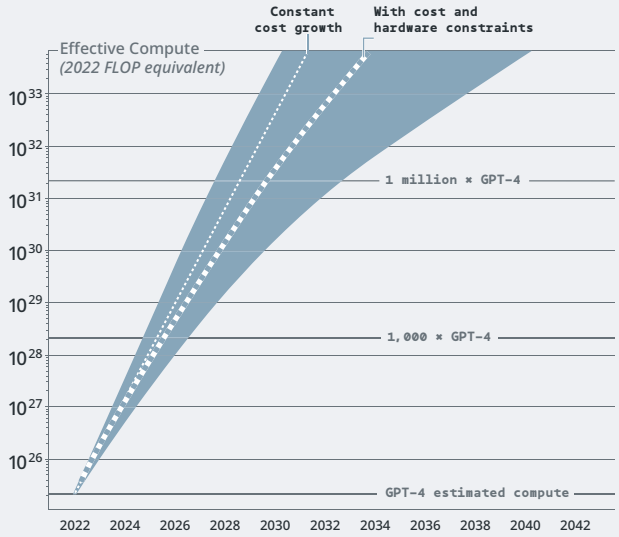

*Training costs are projected using a tapered cost growth model. Training costs initially double every 9.7 months, but the doubling rate is arbitrarily assumed to increase by 1.5 months per year, slowing the rate of cost growth. A constant 9.7-month cost doubling rate is shown for comparison.*

**FIGURE 4.3 | TRAINING COMPUTE** (TAPERED COST PROJECTION AND HARDWARE LIMITS AROUND 2031)




*The amount of compute used to train frontier AI models is projected using a tapered cost growth model and accounting for limits in hardware performance. The doubling period for training costs is arbitrarily assumed to increase by 1.5 months per year, slowing the rate of cost and compute growth. Hardware performance gains are assumed to slow beginning in 2027 and stop completely by 2035, resulting in a final FLOP/s per dollar level roughly equivalent to an abrupt stop in hardware performance improvements around 2031. A constant 7.0-month compute doubling rate is shown for comparison.*

**FIGURE 4.4 | EFFECTIVE COMPUTE** (TAPERED COST PROJECTION AND HARDWARE LIMITS AROUND 2031)




*The effective compute used to train frontier AI models is projected using a tapered cost growth model and accounting for limits in hardware performance. A constant growth rate using a 7.0-month doubling rate for compute and an 8.4-month doubling rate for algorithmic efficiency is shown for comparison.*[85]

**TABLE 2.1 | FRONTIER MODEL TRAINING COST AND COMPUTE OVER TIME**
(ASSUMING HARDWARE LIMITS AROUND 2031 AND TAPERED COST GROWTH PROJECTION, +1.5 MONTH INCREASE IN COST DOUBLING TIME PER YEAR)

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **Cost of final training run in 2022 dollars (tapered cost growth)** | $220M | $1.2B | $4.3B | $12B | $30B |
| **Training compute in FLOP (hardware limits around 2031, tapered cost growth)** | $1.8 \times 10^{26}$ | $2.6 \times 10^{27}$ | $1.9 \times 10^{28}$ | $8.1 \times 10^{28}$ | $2.1 \times 10^{29}$ |
| **Compute relative to GPT-4** | 9 × | 100 × | 900 × | 4,000 × | 10,000 × |

**Finding**
The amount of compute used to train frontier models continues to rise quickly for the next approximately four to five years, reaching around 100 times more than GPT-4 by 2027. By 2030, however, compute growth slows significantly due to cost and hardware constraints. By 2034, compute has reached $10^{29}$ FLOP, or around 5,000 times more than GPT-4, at a cost of around $15 billion.

**TABLE 2.2 | FRONTIER MODEL TRAINING COST AND EFFECTIVE COMPUTE OVER TIME**
(ASSUMING HARDWARE LIMITS AROUND 2031 AND TAPERED COST GROWTH PROJECTION, +1.5 MONTH INCREASE IN COST DOUBLING TIME PER YEAR)

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **Cost of final training run in 2022 dollars (tapered cost growth)** | $220M | $1.2B | $4.3B | $12B | $30B |
| **Effective compute in 2022 FLOP equivalent (hardware limits around 2031, tapered cost growth)** | $1.3 \times 10^{27}$ | $3.7 \times 10^{29}$ | $5.3 \times 10^{31}$ | $4.4 \times 10^{33}$ | $2.2 \times 10^{35}$ |
| **Effective compute relative to GPT-4** | 60 × | 20,000 × | 3 million × | 200 million × | 10 billion × |

**Finding**
Accounting for algorithmic progress, effective compute could increase by 2030 to approximately 1 millionfold above GPT-4, to around the equivalent of $10^{31}$ FLOP in 2022. If algorithms continue to improve, effective compute could increase by the mid-2030s to approximately 1 billionfold above GPT-4, to around the equivalent of $10^{34}$ FLOP in 2022.

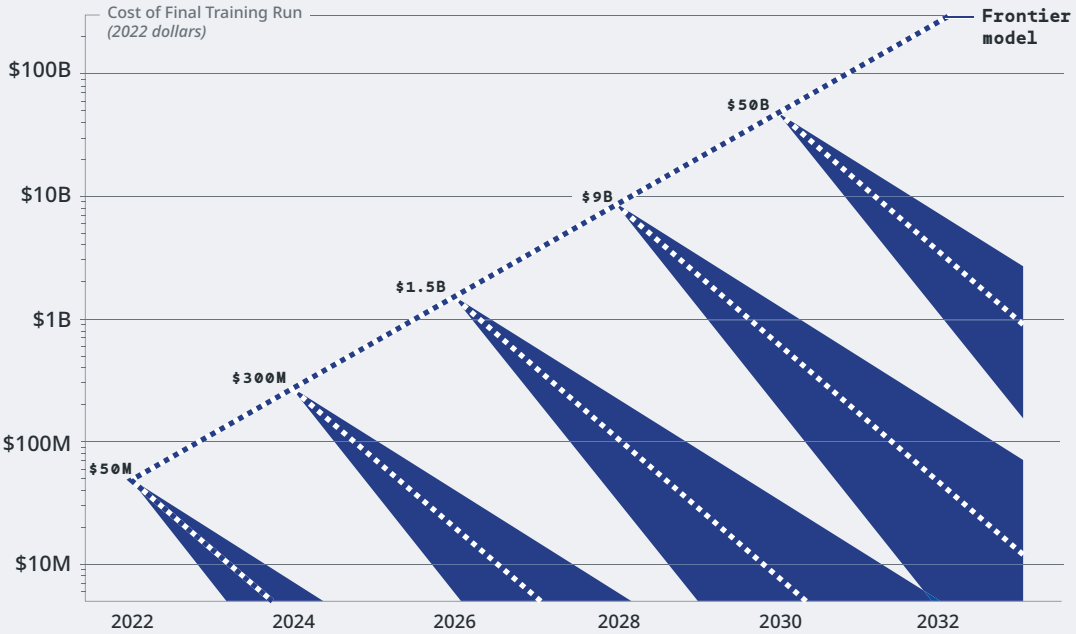
## Proliferation

How might improvements in hardware and algorithmic efficiency affect the availability of AI capabilities over time?

Cost is already an obstacle to many AI researchers training the largest models. If costs continue to rise, the number of actors that can conduct the most compute-intensive research will shrink further, leading to an oligopoly in frontier AI research. In the most extreme case, if costs move into the realm of major governments, the number of global actors that could afford building the most capable AI models could be very small (for example, the United States, China, and the European Union). However, the cost to train a model at any given level of capability rapidly decreases over time due to both hardware and algorithmic improvements. At present, AI breakthroughs proliferate rapidly. It took 35 months for an open-source equivalent of AlphaFold to be released, 14 months for an open-source equivalent of GPT-3, and seven months for an open-source equivalent of GPT-3.5.[86] Some AI labs have switched to a limited release approach, only allowing access to their latest models through an API, slowing proliferation. As costs drop over time, however, eventually training costs become low enough that they become affordable to an actor willing to open-source the model. For example, the first version of Stable Diffusion, which was released open source, cost $600,000 to train.[87] As model costs become low enough that they are affordable to a wider array of actors, it is increasingly likely that someone releases an open-source version.

Figure 5 shows how training costs for any given once-frontier model decrease over time due to improvements in hardware and algorithmic efficiency. Hardware improvements are assumed to lead to a doubling in GPU price-performance approximately every 2.1 years.[88] Improvements in algorithmic efficiency are assumed according to the approximately 8.4-month observed doubling rate in large language models.

**FIGURE 5 | CAPABILITIES OF ONCE-FRONTIER MODELS BECOME MORE ACCESSIBLE OVER TIME AS COSTS DECREASE**




*The cost to train a model with the equivalent capabilities of a once-frontier model decreases over time due to improvements in hardware and algorithmic efficiency. Starting at the cost of training a frontier model in 2022 ($50 million), 2024 ($300 million), 2026 ($1.5 billion), 2028 ($9 billion), and 2030 ($50 billion), the cost to train a model with equivalent capabilities is shown decreasing over time due to hardware and algorithmic improvements. Machine learning GPU price-performance is assumed to double every 2.1 years (95 percent CI: 1.6 to 2.91 years) and algorithmic efficiency is assumed to double every 8.4 months (95 percent CI: 5.3 to 13 months). The uncertainty in the shaded blue regions combines uncertainty for hardware and algorithmic improvements.*[89]

### Discussion

Capabilities that once were limited to state-of-the-art models quickly become more affordable over time. Because compute requirements halve every 8.4 months due to algorithmic improvements and hardware performance per dollar doubles every 2.1 years, the cost to train a model with equivalent performance rapidly decreases. Training a frontier model would cost $50 million in 2022 (the estimate for GPT-4); a straightforward projection of hardware and algorithmic efficiency trends predicts that training a model with the equivalent performance of GPT-4 would cost $4 million in 2024 and only $250,000 by 2026. Under this projection, training a frontier model would cost $1.5 billion in 2026; training a model with the equivalent performance would cost $30 million by 2029 and $2 million by 2031. Even if models reach very high costs, exponential improvements in hardware and algorithmic efficiency make such models affordable in only a few years. A frontier model in 2028 would cost $10 billion to train, yet the equivalent capabilities could be reached with a $160 million model in three years (2031) and a $10 million model in five years (2033).

**FINDING**

Even if costs rise to the point where they severely limit access to only the wealthiest global actors, any given level of capability is still likely to proliferate rapidly within a few years as training costs decline if hardware and algorithmic advancements remain widely available. Wealthy actors training models at AI's frontier are likely to retain a relative advantage over competitors, however, if they continue to train ever-larger and more compute-intensive models.

Under these projections, the capabilities of once-frontier models become widely accessible in only a few years as training costs for any given level of compute rapidly fall due to improvements in hardware and algorithmic efficiency. One or both of these assumptions may not be correct. Hardware improvements could slow or stall completely, a scenario this projection does not include. Some actors may not have access to the most advanced chips, for example, due to export controls. Algorithmic efficiency gains may not be widely available if frontier labs restrict information about their models, as some are doing already. On the other hand, even if frontier labs restrict information, the open-source community can still improve algorithms using smaller models.[90] And some information about frontier models may leak out.[91] As in other areas of projecting AI trends, there is tremendous uncertainty about how rapidly training costs might fall in the future, enabling proliferation.

## Costs for Hardware-Restricted Actors

How might constraints on hardware availability (for example, due to export controls) affect cost and compute growth for actors denied access to continued improvements in AI hardware?

Access to advanced computing hardware is likely to be a limiting factor for some actors in training large models. On October 7, 2022, the U.S. Commerce Department issued sweeping export controls on semiconductor technology destined for China, which the Commerce Department updated and expanded on October 17, 2023. The controls limit the export to China of semiconductor manufacturing equipment and advanced GPUs, even when the chips themselves or their components are manufactured overseas, such as in Taiwan. Nvidia's A100 chip, which has been used to train several milestone machine learning models, is now banned for export to China, as is Nvidia's most recent H100 chip. The 2023 update additionally covers Nvidia's A800 and H800 chips. If these export controls remain in place and are maximally effective, they could have a significant effect on slowing Chinese frontier AI development over time.

Chinese AI labs could attempt to circumvent these export controls through several means. Some Chinese suppliers reportedly are smuggling banned chips into China, although the scale of such efforts is unclear.[92] Additionally, Chinese AI researchers can still access compute through cloud providers, which are not captured under current export controls, although recent moves by the U.S. government suggest it may be looking to address this gap.[93] Additionally, while Chinese-headquartered companies and their overseas subsidiaries can no longer purchase export-controlled chips, creative corporate restructuring may be able to circumvent these rules.[94] For U.S. export controls to be effective in restraining Chinese labs' ability to train frontier AI models, the U.S. government would need to put in place additional measures to cut down on large-scale smuggling of advanced chips and deny Chinese AI labs access to large-scale compute through cloud providers or data centers outside China.

The immediate effect of U.S. export controls on Chinese AI development is likely marginal. Some now-banned A100 chips already were sold to Chinese firms before the export controls took effect.[95] Additionally, Nvidia responded to U.S. export controls by releasing the A800 and H800 chips, export-compliant versions of the banned A100 and H100, respectively.[96] While the A800 and H800 now are banned based on updated rules issued by the U.S. government in October 2023, approximately 100,000 have been sold to Chinese firms.[97] Using 800-series chips, which have lower interconnect bandwidth than the A100 and H100, would make large-scale training runs more inefficient and costly.

Even if the immediate effect is small, if U.S. export controls remain in place and can be effectively enforced, they are likely to severely restrict China's access to leading-edge GPUs in the coming years. U.S. Commerce Department officials have said that they anticipate the current threshold for restricted GPUs will remain in place, even as chips continue to advance, preventing Chinese labs from accessing future, more efficient GPUs.[98] Moreover, because U.S. export controls also target the manufacturing equipment needed for advanced Chinese fabs, China also will be hindered in its ability to manufacture its own advanced chips. U.S. export controls currently prohibit the sale of U.S.-origin manufacturing equipment to leading manufacturing nodes in China. Japan and the Netherlands have adopted similar restrictions, effectively shutting China out of manufacturing its own advanced chips in the near term.[99] (The United States, the Netherlands, and Japan collectively control 90 percent of the global semiconductor manufacturing equipment market.)

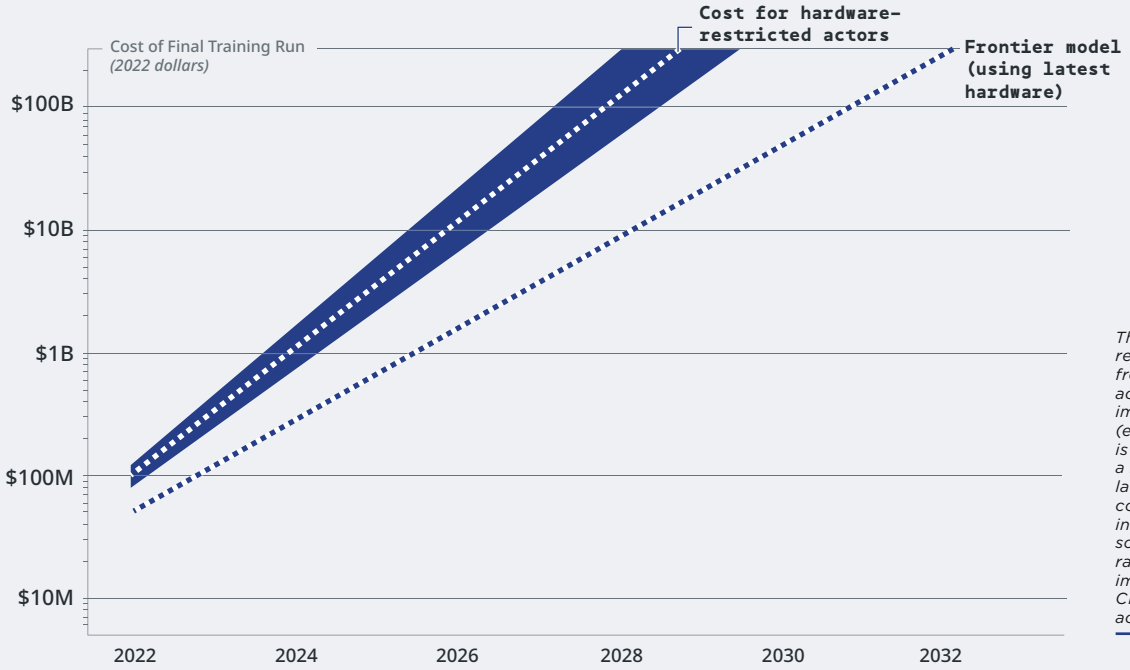
In the long run, U.S. chip export controls will not remain effective forever. China is working to develop its own indigenous tooling capacity for manufacturing advanced chips, although it faces significant technological hurdles to doing so. China has seen some success with chips produced at SMIC's 7 nm node, which are in Huawei's Mate 60 Pro phone. These chips represent a significant improvement over previous Chinese attempts to indigenously produce advanced chips.[100] However, there remain challenges to producing these chips in large quantities using China's currently available equipment (deep ultraviolet lithography), and these chips remain substantially behind TSMC's in quality.[101] In addition to Chinese efforts to improve indigenous chip production, U.S. export controls create incentives for global semiconductor manufacturers to design-out U.S. components from their fabs to circumvent U.S. restrictions to sell to the Chinese market. However, U.S. export controls could hinder China's access to large-scale compute for at least several years.

If the United States is successful in denying China access to leading-edge chips, Chinese AI research labs would be forced to use older, less advanced GPUs for training, effectively increasing the cost of large-scale training runs. Khan and Mann (2020) estimated that denying an actor access to leading-edge chips could quickly make training large models "economically prohibitive."[102] In addition to increased costs, using older chips also is likely to increase the engineering challenges in training larger numbers of chips in parallel. It also may slow AI research by forcing longer training runs and slowing iteration on model development. The effects of denying actors access to advanced chips is likely to be complex and multifaceted. Hardware-restricted actors certainly will not be able to overcome all these hurdles simply by paying more in compute. However, one effect is likely to increase compute costs for accessing state-of-the-art models.

**Advances in algorithmic efficiency will make training models with the equivalent capabilities of once-frontier models affordable within only a few years.**

Figure 6.1 projects costs for "hardware-restricted actors" who are denied access to continued hardware improvements, such as due to export controls. This projection assumes that export controls are maximally effective, denying hardware-restricted actors from any further advances in chips. The projection assumes hardware-restricted actors are unable to access hardware improvements beyond 2020, since the now-banned Nvidia A100 was released in 2020. As such, this projection estimates an upper bound on cost for hardware-restricted actors. In practice, improvements in Chinese indigenous chip fabrication could enable China to close the cost gap somewhat. These projections estimate the cost for hardware-restricted actors to attempt to keep pace with frontier models, using the naive assumption that hardware-restricted actors can access larger amounts of compute by simply paying more, using more older model chips for longer training runs. This simplification assumes away significant engineering challenges and limitations in scaling compute using older model chips and likely underestimates the engineering challenges for hardware-restricted actors. Under a scenario of maximally effective export controls, the cost for hardware-restricted actors to keep pace with frontier AI model development quickly becomes unfeasible. The uncertainty in the projection, indicated in the shaded blue region, is due to uncertainty in the rate of hardware improvements, which hardware-restricted actors are denied.

**FIGURE 6.1 | RESTRICTIONS ON HARDWARE MAKE IT UNAFFORDABLE TO KEEP PACE WITH FRONTIER MODELS**




*The cost for hardware-restricted actors to train a frontier AI model without access to any hardware improvements after 2022 (e.g., due to export controls) is shown, with the cost to train a frontier AI model using the latest hardware included for comparison. The uncertainty in the shaded blue region is solely from uncertainty in the rate of growth in hardware improvements (95 percent CI), which hardware-restricted actors are denied.*

## Discussion

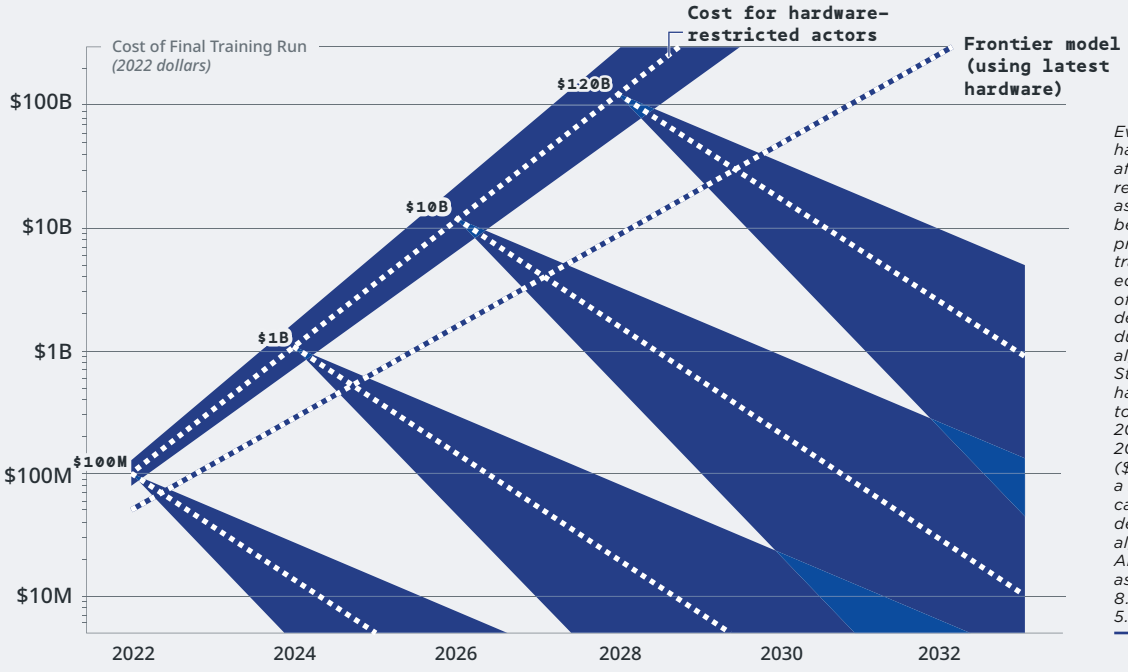
Using the simplistic assumption that hardware-restricted actors cannot access hardware improvements after 2020, this projection estimates a modest present-day twofold cost penalty for hardware-restricted actors to attempt to keep pace with present-day frontier AI models. This is *not* presented as an accurate assessment of the current cost to train a GPT-4 level model with export-compliant chips and does not reflect the practical engineering challenges in doing so. Rather, this projection shows that, even if hardware-restricted actors attempted to keep pace with frontier AI models, and even if the engineering challenges in doing so were assumed away, costs quickly would become unaffordable.

By 2024, the effect of denying an actor access to all hardware improvements after 2020 is a nearly fourfold increase in cost. A frontier model that normally would cost approximately $280 million to train in 2024 would cost a hardware-restricted actor approximately $1 billion. By 2025, the difference is a fivefold increase in cost. An approximately $650 million frontier model in 2025 would cost a hardware-restricted actor more than $3 billion.

As costs rise, a more likely effect of hardware restrictions is only to delay the time when hardware-restricted actors are able to access capabilities until improvements in algorithmic efficiency make models more affordable. Even though hardware-restricted actors cannot afford to keep pace with frontier models (assuming export controls are maximally effective), capabilities nevertheless proliferate rapidly due to algorithmic improvements. Figure 6.2 estimates how costs for any given level of capabilities fall over time due to algorithmic improvements, enabling proliferation of capabilities to hardware-restricted actors.

The cost to train a model with equivalent performance to a once-state-of-the-art $50 million model in 2022 (i.e., GPT-4 level model) rapidly declines to $7 million by 2024 due to algorithmic improvements alone. To train the equivalent of a $1.5 billion frontier model in 2026 would cost a hardware-restricted actor initially over $10 billion, even assuming they could solve the associated engineering challenges. Yet the cost to train a model with equivalent capabilities declines to around $50 million in only four years, by 2030. Even if rising costs force hardware-restricted actors to fall behind the frontier of AI research for the largest models, the time delay is likely to be only a few years. Even costs for very expensive models decline rapidly. Training the equivalent of an approximately $10 billion frontier model in 2028 initially would cost a hardware-restricted actor a whopping $120 billion. However,

**FIGURE 6.2 | CAPABILITIES PROLIFERATE TO HARDWARE-RESTRICTED ACTORS DUE TO ALGORITHMIC IMPROVEMENTS**




*Even without access to hardware improvements after 2022, hardware-restricted actors are assumed to be able to benefit from algorithmic progress. The cost to train a model with the equivalent capabilities of a once-frontier model decreases over time due to improvements in algorithmic efficiency. Starting at the cost for hardware-restricted actors to train a frontier model in 2022 ($100M), 2024 ($1B), 2026 ($10B), and 2028 ($120B), the cost to train a model with equivalent capabilities is shown decreasing over time due to algorithmic improvements. Algorithmic efficiency is assumed to double every 8.4 months (95 percent CI: 5.3 to 13 months).*

algorithmic improvements would drive training costs down to $600 million in only four years, by 2032. These rapid cost decreases are driven entirely by algorithmic progress, which this projection estimates doubling every 8.4 months. A slower rate of publicly available algorithmic progress, whether due to decreased information-sharing by leading AI labs or government regulation of large-scale models, would lead to longer timelines.

**FINDING**

Export controls that restrict actors' ability to access the most advanced chips, if effectively enforced, will rapidly price hardware-restricted actors out of competing in frontier AI models. Using larger numbers of older model chips quickly will become unaffordable, even without accounting for engineering challenges. However, advances in algorithmic efficiency will make training models with the equivalent capabilities of once-frontier models affordable within only a few years. The most likely effect of hardware restrictions will be to delay those actors' access to any given level of capability by only a few years, if algorithmic improvements remain widely available.

## Compute Regulatory Threshold

How might improvements in hardware and algorithmic efficiency impact the effectiveness of training compute as a regulatory threshold for frontier models over time?

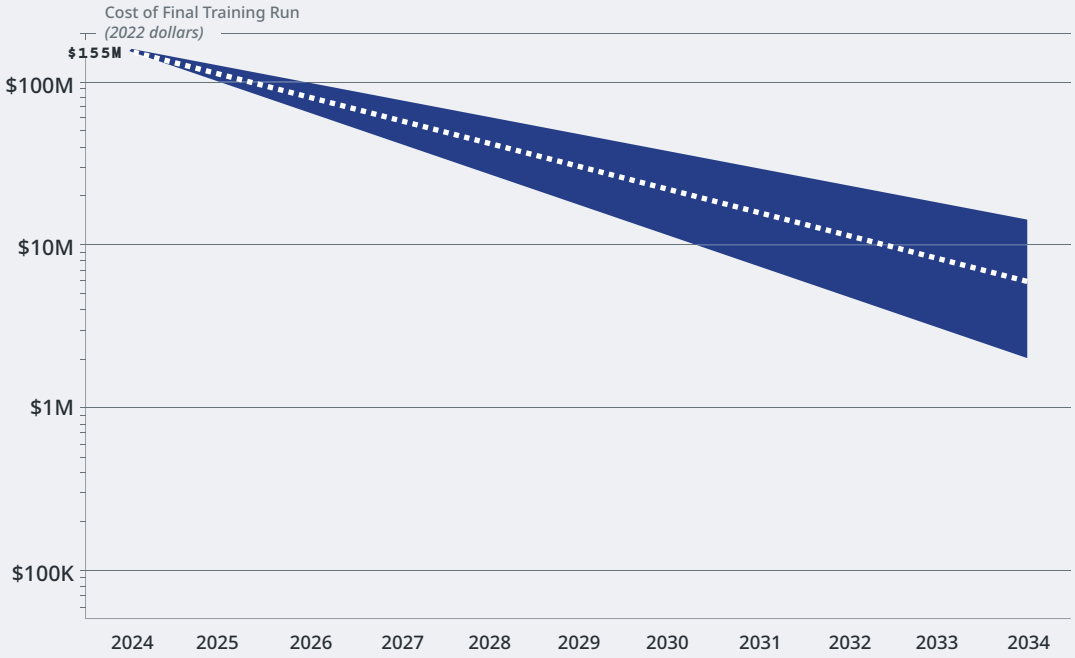
As frontier AI models have increased in capability, including dual-use capabilities such as the ability to generate disinformation or enable cyber, chemical, or biological attacks, a growing number of voices have called for regulating their development and/or use. Some AI policy researchers, including some at frontier AI labs, have proposed using the amount of compute used to train a model as a regulatory threshold triggering greater oversight, safeguards, or government licensing and approval before release.[103] On October 30, 2023, the Biden administration issued an executive order on AI that, among other actions, established a $10^{26}$ operations threshold for reporting to the government.[104] This threshold, which is roughly five times higher than the estimated compute used to train GPT-4, aims to capture future models trained with more compute than preexisting frontier models (as of October 2023).

If current trends continue, using GPT-4 as a starting point, the $10^{26}$ FLOP threshold is projected to be reached in early 2024 with a cost of roughly $155 million (95 percent CI: $135 million to $170 million) for the final training run. (In practice, the actual timing for when the first model will cross the notification threshold will depend on the compute availability, experiment pipeline, and strategies of the handful of labs that are able to train a $10^{26}$ FLOP model today.)

Initially, only a small number of leading AI labs would be able to train such a model. Training costs would fall over time, however, due to hardware improvements. Machine learning GPU price-performance is doubling every 2.1 years. Figure 7.1 projects the cost to train a $10^{26}$ FLOP model declining over time due to hardware improvements, starting with a $155 million cost in 2024.

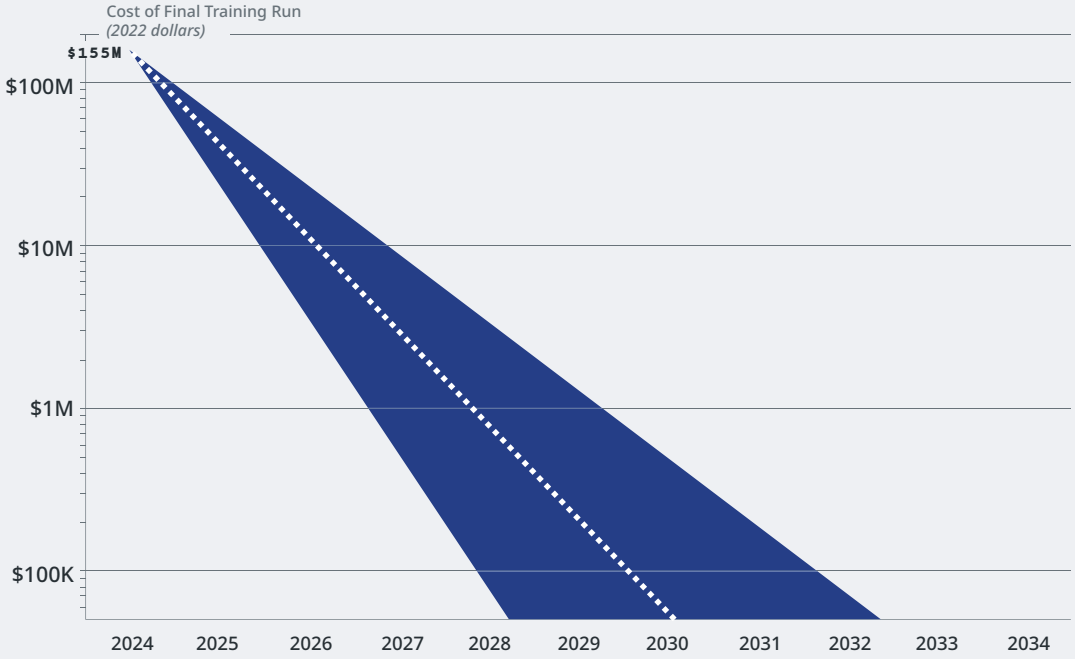
Under this projection, the cost to train a $10^{26}$ FLOP model would decrease relatively slowly from $155 million in 2024 to $30 million five years later in 2029. This would limit the number of actors that could train a model at the regulatory threshold. If AI labs were restricted from releasing the model weights open source for a trained model at the regulatory threshold, then only a relatively small number of actors would be able to train a $10^{26}$ FLOP regulatory threshold model throughout the 2020s.

**FIGURE 7.1 | THE COST TO TRAIN A $10^{26}$ FLOP REGULATORY THRESHOLD MODEL DECREASES OVER TIME**




*The cost to train a regulatory threshold model ($10^{26}$ FLOP), estimated to be $155 million in 2024, decreases due to machine learning GPU price-performance doubling every 2.1 years (95 percent CI: 1.6 to 2.91 years).*

**FIGURE 7.2 | THE COST TO TRAIN A MODEL WITH CAPABILITIES EQUIVALENT TO A $10^{26}$ FLOP MODEL IN 2024 DECREASES OVER TIME**




*The cost to train a model with capabilities equivalent to a $10^{26}$ FLOP model in 2024, assumed to cost $155 million in 2024, rapidly drops due to improvements in hardware and algorithmic efficiency. Machine learning GPU price-performance is assumed to double every 2.1 years (95 percent CI: 1.6 to 2.91 years), and algorithmic efficiency is assumed to double every 8.4 months (95 percent CI: 5.3 to 13 months). The uncertainty in the shaded blue region combines uncertainty for hardware and algorithmic improvements.[105]*

This projection does not account for algorithmic efficiency gains, however. Algorithmic efficiency for large language models is currently doubling every 8.4 months. Improvements in algorithmic efficiency will allow models with equivalent performance to a $10^{26}$ FLOP model in 2024 to be trained in the future by using compute more efficiently. For example, due to algorithmic improvements, a $10^{24}$ FLOP model in 2028 will have the same performance as a $10^{26}$ FLOP model in 2024. Correspondingly, this will drop the cost to train a model with performance *equivalent* to a $10^{26}$ FLOP model in 2024, and this drop in cost will happen much faster than with hardware improvements alone. Figure 7.2 shows how the cost to train a model with performance equivalent to a $10^{26}$ FLOP model in 2024 will drop over time due to both hardware and algorithmic improvements.

Under this projection, training a model equivalent in performance to a $10^{26}$ FLOP model in 2024 would cost around $10 million by 2026 and less than $1 million by 2028. This assumes that hardware and algorithmic improvements continue at their current rate. If improvements in algorithmic efficiency slow down, due to a shift to a more closed research ecosystem or government regulations, that could slow the rate of decline of training costs for models with performance equivalent to a $10^{26}$ FLOP model in 2024.

## Discussion

The capabilities of $10^{26}$ FLOP models and their risks are, at present, entirely unknown. The Biden administration established the compute threshold at $10^{26}$ operations because it was higher than current publicly known models. GPT-4 is estimated to have used $2.1 \times 10^{25}$ FLOP, so the compute threshold is roughly five times the amount of compute used to train GPT-4. The compute threshold was not chosen to reflect known hazards, but rather the absence of information and the potential for hazards at the frontier of AI progress. The notification requirement will require companies to give information to the U.S. government about the model, potential hazards, and safety testing. As models of this scale are trained and information is reported to the government, increased information about the capabilities, any potential risks, and safeguards will grow over time. Future regulators should have more information about the performance and any potential hazards from models trained with $10^{26}$ or greater FLOP.

The fact that an increasing array of actors will be capable of training a model with capabilities equivalent to a $10^{26}$ FLOP model in 2024 is not necessarily a problem. Nor does it necessarily obviate the need for the compute threshold. If models at the equivalent of $10^{26}$ FLOP turn out to be dangerous—for example, if they enabled the development of biological weapons—then restricting their release could be warranted. In that case, the compute threshold may need to decrease over time to account for algorithmic efficiency. On the other hand, if $10^{26}$ FLOP models turn out not to be dangerous—or if society is able to take steps to harden itself against such dangers over time—then the increasing accessibility of such models may not be problematic.

If government regulators wanted to retain the compute threshold at the frontier of AI progress, only capturing the most advanced models at any given point in time, then the compute threshold may need to increase over time to keep pace with compute growth.

Alternatively, government regulators may wish to keep the compute threshold at $10^{26}$ FLOP specifically to future-proof against algorithmic gains that will increase the accessibility of future capabilities. By the late 2020s, when frontier models are projected to reach $10^{28}$ FLOP, it will take another approximately five years for algorithmic gains to make it possible to train a model using only $10^{26}$ FLOP that has performance equivalent to a 2028 frontier model. Maintaining the regulatory threshold at $10^{26}$ FLOP may slow the proliferation of future capabilities that are even more powerful—and potentially dangerous—than the capabilities of models at the regulatory threshold when it is first crossed. The cost to train a $10^{26}$ FLOP model is likely to remain in the tens of millions of dollars through the 2020s, so the number of actors who would be affected by a compute threshold that remained at $10^{26}$ FLOP likely would be relatively small.

Using a training compute threshold for triggering a requirement to notify the government about model development and deployment is a novel approach. While training compute is an input rather than an output, and it is model capabilities that are actually relevant for causing potential harm, capabilities cannot be reliably anticipated in advance of training, especially at AI's frontier. In some cases, further capability improvements can be gained from a model even after training simply by improved prompting, chain-of-thought reasoning, or linking multiple models together. The amount of compute used for training can be a useful trigger for greater scrutiny of potential risks from new models whose capabilities may be unknown. Government regulators will need to be flexible and responsive to future hazards, known and unknown, as AI researchers continue to scale to larger models and probe the unknown frontiers of AI development.

## Conclusion

Policymakers should prepare for a world of significantly more powerful AI systems over the next 10 to 15 years. These developments could occur without fundamental breakthroughs in AI science, but simply by scaling up today's techniques to train larger models on more data and computation.

Despite constraints on spending and hardware performance, the amount of compute used to train frontier AI models could increase significantly in the next decade. By the late 2020s or early 2030s, training compute could be approximately 1,000 times that used for GPT-4, and effective compute could be approximately 1 million times GPT-4. There is some uncertainty about when these thresholds could be reached, but this level of growth appears possible within cost and hardware constraints. Improvements of this magnitude are possible without government intervention, entirely funded by private corporations on the scale of large tech companies today. Nor do they require fundamental breakthroughs in chip manufacturing
or design.

Even further gains could be possible if any of the following turn out to be true:

- Companies are able to spend larger sums of money due to increased revenues from AI
- Governments fund even larger and more costly training runs
- Hardware performance continues to improve into the 2030s
- Algorithmic efficiency continues to improve into the 2030s.

This paper makes no predictions about the specific capabilities of a system with one million times the performance of GPT-4. However, historical experience with scaling shows that the result is not just higher performing models but unlocking entirely new capabilities. The baseline trajectory of frontier AI systems in the near future is not systems slightly more powerful than the current state of the art; it is systems vastly more powerful than today.

The challenge that policymakers face is not regulating GPT-4 level systems; it is regulating the vastly more powerful systems that could be built in the next 10 to 15 years. Policymakers should begin to put in place today a regulatory framework to prepare for such systems. Just as some AI labs are developing "responsible scaling policies," governments must future-proof their regulations to be prepared for future models that are significantly more compute-intensive and capable than those today.[106] Building a regulatory framework in anticipation of more powerful AI systems is essential because of the disconnect in speeds between AI progress and the policymaking process, the difficulty in predicting the capabilities of new AI systems on specific tasks, and the speed with which AI models proliferate today, absent regulation. Waiting to regulate future, more powerful AI systems until concrete harms materialize almost certainly will result in regulation being too late. In the worst case, dangerous models could have been already released open source, making efforts to constrain their proliferation effectively impossible.

Computing hardware is likely to be a fruitful avenue for regulation if current trends continue. Massive amounts of compute are the cost of entry to training frontier AI models. If current trends continue, compute is likely to increase in importance over the next 10 to 15 years as an essential input to training the most capable AI systems. Regulations on access to compute (e.g., export controls on chips, know-your-customer cloud computing requirements), if effective, could restrict China's abilities to keep pace with frontier AI development. However, restrictions on access to compute are likely to slow, but not halt, proliferation, given the ability of algorithmic advances to enable training equivalent systems on less compute over time.

Regulations on compute will be more effective if paired with regulations on models themselves. Algorithmic progress has a significant impact on the rate of proliferation of frontier models. Changes in industry behavior or government regulations with respect to trained models conceivably could slow the rate of proliferation. Widely available algorithmic improvements could slow, for example, if leading AI labs withheld further details about their frontier models, or if the U.S. government applied export controls to trained models above a certain compute threshold, as it already has to chips.[107]

Policymakers face difficult choices about regulating AI in the face of significant uncertainty. Understanding historical trends in cost, compute, and algorithmic efficiency can help policymakers make reasonable projections about what kinds of futures are possible over the next decade-plus, accounting for potential limitations in further growth. Better informed policymaking can enable better policies, not just to respond to the concerns of AI systems today, but to address the possibilities of future systems as well.

# APPENDICES

## Appendix A: Additional Limitations on Compute Growth

There are several factors that may limit compute growth over time. This paper explicitly models scenarios in which cost, hardware performance, and access to hardware are all limitations on training compute continuing to grow at the current rate. The paper does not model other limiting factors, such as upper limits on government spending, the total amount of available data for training, and engineering challenges associated with parallelizing training for very large models. These factors could limit future compute growth under certain scenarios. A brief discussion of some of these limitations is included below.

### Hardware Limitations

Improvements in hardware performance are a major factor in the future cost of compute since they increase the amount of compute available per dollar. Hobbhahn et al. (2023) estimated that machine learning GPU price-performance has been doubling every 2.1 years based on an analysis of 47 machine learning accelerators (GPUs and other AI chips) from 2010 to 2023. The baseline assumption this paper uses projects compute per dollar continuing to improve at this rate. This may not be realistic, however. There are good reasons, both physics-based and economics-based, for believing that current hardware improvement trends are not sustainable. As transistors continue to shrink, they are approaching the size of a single silicon atom. The costs for building new leading-edge fabs also is increasing, an economic reflection of the increasingly herculean engineering efforts required to build smaller chips.[108] New leading-edge fabs can cost $10 billion to $20 billion.[109] A more sophisticated projection would take into account limits that cause hardware improvements to slow or stop entirely in the future.

Other compute projections have included limits on hardware improvements. Cotra (2020) projected compute per dollar doubling every 2.5 years until the late 2070s before leveling off at $10^{24}$ FLOP/s per dollar. Hobbhahn (2022) used a more conservative projection of GPU price-performance doubling every 2.8 years until around 2030, after which it slows to doubling approximately every 3.5 years.

Recent physics-based analysis by Hobbhahn and Besiroglu (2022) of GPU improvements estimated that GPU price-performance would stop improving between 2027 and 2035 as transistors approach the size of roughly a single silicon atom. Hobbhahn and Besiroglu estimated that GPU price-performance would cap at around $10^{22}$ to $10^{23}$ FLOP/s per dollar.[110]

Taking a different approach, Ho et al. (2023) estimated fundamental limits in microprocessor performance due to limits in energy efficiency. GPU energy efficiency is currently doubling approximately every 2.7 years. Ho et al. estimated that the current paradigm for microprocessors may reach a fundamental limit in improved energy efficiency at around $4.7 \times 10^{15}$ 4-bit floating-point operations per Joule, which is around 200 times more energy efficient than today's microprocessors.[111] At the current rate of improvement, this limit would be reached in around 20 years, or in the mid-2040s.

The section Limits on Hardware Improvements projects cost and compute growth under a scenario in which GPU price-performance levels off between 2027 and 2035, per Hobbhahn and Besiroglu (2022).

### Data Limitations

Data also could become a limiting factor in compute growth. Hoffman et al. (2022) demonstrated significantly improved performance with the language model Chinchilla (70B) over Gopher (280B) with the same compute budget by leveraging four times as much training data.[112] They concluded that training data should scale approximately equal to model size as the compute budget increases, an update to the scaling laws presented by Kaplan et al. (2020).[113] These findings suggest that some large language models (e.g., GPT-3, MT-NLG) are overly large and have undersized datasets relative to their compute budget. A more optimal use of compute resources would focus greater attention on scaling data, such that data scales roughly proportional to model size.

If researchers adjust the design of future large models accordingly, data could, in principle, become a limiting factor in compute growth. Villalobos and Ho (2022) found that language dataset size is growing 0.22 orders of magnitude per year (OOMs/yr), equivalent to doubling every 16.4 months.[114] This is significantly slower than the 7.0-month doubling rate for compute for large language models. If Chinchilla's performance leads researchers to increase attention on data, it is possible that training datasets will grow at a faster rate than they have been to date. Alternatively, it could be the case that it is difficult to increase dataset growth faster than the present rate—for example, if building ever-larger datasets becomes increasingly time-intensive, if data becomes increasingly difficult to find, or if high-quality data turns out to matter the most and is in short supply. Even if dataset size grows at an increased rate, it may not match the 7.0-month historical compute doubling rate for large models.

It also is possible that the total quantity of useful data is limited, and researchers simply run out of new data on which to train larger models. Villalabos et al. (2022) found that dataset size is growing faster than available stocks for language and vision data.[115] They estimated that the available stock of high-quality language data is likely to be exhausted before 2026, low-quality language data by 2030 to 2050, and vision data by 2030 to 2060.

However, there are additional data sources that researchers could turn to once these stocks are exhausted, including video, audio, geolocation data, Internet of Things device data, and proprietary text and image datasets. If data becomes a limiting factor in increasing AI capabilities, companies also may put additional resources into generating new data, such as by capturing data from the environment or human behavior. In the past, companies have engaged in major data creation efforts, from geolocation data for mapping applications to detailed environmental maps for self-driving cars. In theory, AI-generated data may even be able to supplement datasets.

The total amount of data is not fixed, and a long-term trend in the information revolution has been toward cheaper, ubiquitous sensors that can collect more data in a greater number of modalities. Data may become more difficult to find, and a shift toward proprietary datasets could further advantage large corporations and disadvantage academic researchers, but a hard limit on data seems unlikely. For this reason, this paper does not model a limit on data available for training.

### Engineering Challenges

As researchers train larger models, parallelizing training across thousands of chips presents significant engineering hurdles.[116] As a practical matter, the engineering challenges associated with large-scale training runs may be a more significant factor limiting the rate of growth in large-scale training runs for leading AI labs than cost, per se. Marshaling the funding needed to purchase 100,000 chips may be simpler than actually networking them together into a usable cluster for training a large model.

Additionally, the human capital required to overcome these engineering obstacles, rather than cost alone, is likely a limiting factor for many actors in competing with top AI labs. Actors who are not able to access the human capital required to manage the engineering challenges associated with large training runs may be shut out of competing in frontier models even if other factors, such as cost, data, and hardware, are not limiting factors. (However, cost may be a crude proxy for the engineering challenges associated with large-scale training runs.)

This paper makes no assumptions about which limiting factor(s) drive the current compute doubling rate. It merely projects the current compute growth rate forward until cost becomes a constraint. It may be the case that the engineering challenges associated with large-scale training runs are the most significant constraint on the rate of compute growth. If so, and if addressing these technical issues becomes increasingly challenging, this could slow the rate of compute growth in the future.[117]

# Appendix B: Observed Growth Rates

**TABLE A.1 | OBSERVED RATES OF GROWTH IN RELEVANT METRICS**

| Metric | Doubling period | Orders of magnitude per year (OOMs/yr) | Years to reach 10 × improvement | Source |
|---|---|---|---|---|
| **Compute for training (FLOP)** | 6.3 months (0.53 years) | 0.57 OOMs/yr | 1.74 years | Epoch (2023)[118] |
| **Compute for training large models (FLOP)** | 7.0 months (0.58 years) | 0.52 OOMs/yr | 1.94 years | Epoch (2023)[119] |
| **Machine learning GPU price-performance (FLOP/s per dollar)** | 25.2 months (2.1 years) | 0.14 OOMs/yr | 6.98 years | Hobbhahn et al. (2023)[120] |
| **Large language model algorithmic efficiency** | 8.4 months (0.70 years) | 0.43 OOMs/yr | 2.33 years | Ho et al. (forthcoming)[121] |
| **Language dataset size** | 16.4 months (1.37 years) | 0.22 OOMs/yr | 4.55 years | Villalobos and Ho (2022)[122] |
| **Vision dataset size** | 40.1 months (3.34 years) | 0.09 OOMs/yr | 11.11 years | Villalobos and Ho (2022)[123] |

## Appendix C: Estimating Compute Costs

Because training costs are rarely reported, some researchers have attempted to estimate the costs of recent state-of-the-art AI models. H. (2018) estimated the costs to replicate AlphaGo Zero at approximately $35 million, based on the cloud computing rates Google offers to the public. DeepMind's actual costs to conduct the experiments would have been lower, since Google owns the cloud infrastructure.[124] Carey (2018) estimated the actual costs to Google of AlphaGo Zero at around $3 million for the final training run and $3 million to $10 million for the entire experiment.[125] Wang (2020) estimated that it would cost an independent researcher approximately $13 million to replicate the final training run for AlphaStar Final.[126] Sharir et al. (2020) estimated the "list-price" of training T5-11B to be above $1.3 million for a single training run and may have been $10 million for the entire experiment.[127] Li (2020) estimated $4.6 million to train GPT-3.[128] Lohn and Musser (2022) estimated $1.65 million to $4.6 million to train GPT-3.[129] Heim (2022) estimated the cost to an independent researcher to replicate PaLM at between $9 million to $23 million.[130] Epoch (2023) estimated a cost of $50 million (90 percent CI: $30 million to $90 million) for the final training run for GPT-4.[131] OpenAI's CEO Sam Altman stated that the company spent $100 million to train GPT-4, although he did not give further details, and this may include earlier experiments in addition to the final training run.[132] When estimates use prices available to the general public, the actual price to the research labs training these models would have been lower, because in most cases they own, or their parent company owns, the cloud infrastructure.

Establishing a baseline for current model costs from which to project forward is challenging for several reasons. Costs generally are not reported and must be estimated by outside researchers based on reported compute usage. Estimated costs vary depending on whether one is considering the final training run versus the entire cost of an experiment. And the cost for an independent research team to replicate a model is likely to be significantly higher than the actual costs to the researchers conducting the experiment, who often have access to the internal cloud resources of major tech corporations. (Google DeepMind has access to Google's cloud resources and OpenAI uses Microsoft's). Most cost estimates calculate the cost of a final training run if conducted by an independent research group that does not have access to internal cloud resources. This allows for a more precise estimate, using the reported compute for a final training run and commercially available computing prices. The actual cost to conduct a complete experiment is likely to be higher than the final training run, and costs can vary considerably depending on whether a research team has access to a major tech company's cloud infrastructure. The question of cost-to-whom is particularly relevant for this analysis, which aims to better understand how rising costs might affect accessibility and proliferation of state-of-the-art models.

# Appendix D: Uncertainty in Cost Projections

Uncertainties in cost projections come from several factors:

1. Uncertainty in the cost estimates for current models (the starting point for future projections)
2. Uncertainty in the historical rate of cost growth
3. Uncertainty in the extent to which future cost growth will mirror historical trends.

Current best estimates for historical cost growth and current costs are:

- **Cost growth:** Training costs doubling every 9.7 months (95 percent CI: 7.3 to 13.5 months)
- **Training costs:** Cost of final training run for GPT-4: $50 million (90 percent CI: $30 million to $90 million).[133]

## Sensitivity to Initial Starting Cost

Figure A.1 shows straightforward cost projections under three starting costs: $30 million, $50 million, and $90 million, representing the range of plausible costs for GPT-4. The high and low growth estimates for a $50 million starting cost are included for comparison.

Uncertainty in the starting cost is quickly overwhelmed by uncertainty in the rate of growth, which could lead to a wide range of plausible future cost projections. As long as the $50 million cost estimate for GPT-4 is within a factor of two in either direction, any errors in estimating the starting cost are not likely to significantly affect future cost projections.

## Sensitivity to the Rate of Growth

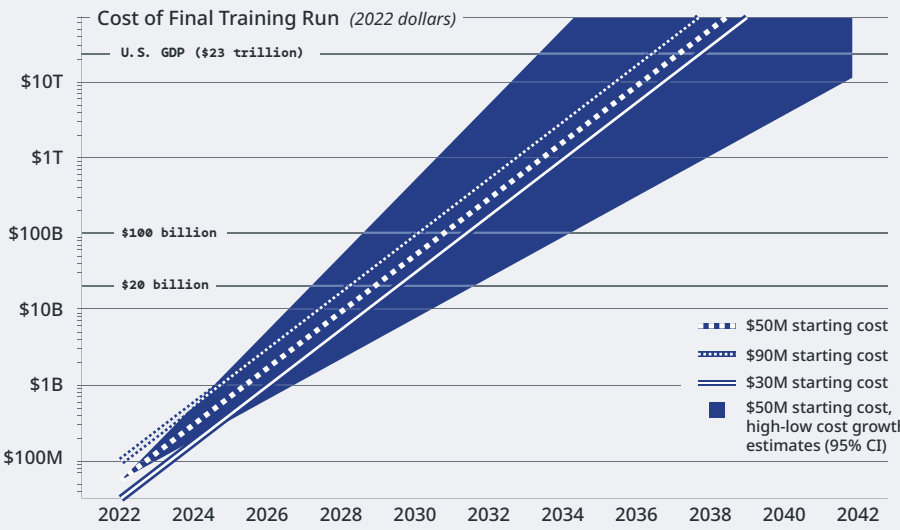
Cost and compute projections are highly sensitive to the rate of growth. Tables A.2 and A.3 show training cost and compute estimates, respectively, under the median, high, and low growth rates.

Initially, the differences between the median, high, and low estimates are relatively small, within a factor of two. Over time, however, the gaps widen due to the compounding effect of the differences in growth rates. By 2030, the high and low estimates are an order of magnitude difference from the median estimates.

When viewed from the perspective of when various cost and compute milestones are reached, however, the variability in growth rates is not as severe. Tables A.4 and A.5 show the year at which various cost and compute milestones, respectively, are reached under median, high and low growth rates.

Because the growth rates under any scenario are so rapid, the effect of the uncertainty in growth rates is simply to shift the date at which various cost and compute milestones are reached by only around one to three years in either direction. The estimated cost limit for private companies, $20 billion, is reached in the late 2020s or early 2030s under all growth rates. Similarly, $10^{28}$ FLOP, or approximately 1,000 times the compute used to train GPT-4, is reached in the late 2020s under all compute growth rate projections.

**FIGURE A.1 | FRONTIER MODEL TRAINING COST OVER TIME (SENSITIVITY ANALYSIS)**




*A straightforward projection of frontier model training costs under three different initial starting costs in 2022: $30M, $50M, and $90M. Cost growth is assumed to double every 9.7 months (95 percent CI: 7.3 to 13.5 months).*

**TABLE A.2 | MEDIAN, HIGH, AND LOW ESTIMATES FOR TRAINING COST OVER TIME IN 2022 DOLLARS** (STRAIGHTFORWARD PROJECTION)

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **9.7-month cost doubling** (median estimate) | $280M | $3.6B | $50B | $600B | $8T |
| **7.3-month cost doubling** (high growth estimate) | $490M | $15B | $450B | $14T | $400T |
| **13.5-month cost doubling** (low growth estimate) | $170M | $1.1B | $7B | $40B | $280B |

**TABLE A.3 | MEDIAN, HIGH, AND LOW ESTIMATES FOR TRAINING COMPUTE OVER TIME IN FLOP** (STRAIGHTFORWARD PROJECTION)

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **7.0-month compute doubling** (median estimate) | $2.3 \times 10^{26}$ | $8.0 \times 10^{27}$ | $2.8 \times 10^{29}$ | $1.0 \times 10^{31}$ | $3.5 \times 10^{32}$ |
| **5.7-month compute doubling** (high growth estimate) | $3.9 \times 10^{26}$ | $3.1 \times 10^{28}$ | $2.5 \times 10^{30}$ | $2 \times 10^{32}$ | $2 \times 10^{34}$ |
| **8.6-month compute doubling** (low growth estimate) | $1.5 \times 10^{26}$ | $2.6 \times 10^{27}$ | $4.8 \times 10^{28}$ | $9 \times 10^{29}$ | $2 \times 10^{31}$ |

**TABLE A.4 | YEAR TO REACH TRAINING COST MILESTONE FOR MEDIAN, HIGH, AND LOW ESTIMATES FOR TRAINING COST GROWTH** (STRAIGHTFORWARD PROJECTION)

| Training cost in 2022 dollars | $500M | $1B | $5B | $10B | $20B |
|---|---|---|---|---|---|
| **9.7-month cost doubling** (median estimate) | 2025 | 2026 | 2027 | 2028 | 2029 |
| **7.3-month cost doubling** (high growth estimate) | 2024 | 2025 | 2026 | 2027 | 2027 |
| **13.5-month cost doubling** (low growth estimate) | 2026 | 2027 | 2030 | 2031 | 2032 |

**TABLE A.5 | YEAR TO REACH TRAINING COMPUTE MILESTONE FOR MEDIAN, HIGH, AND LOW ESTIMATES FOR TRAINING COMPUTE GROWTH** (STRAIGHTFORWARD PROJECTION)

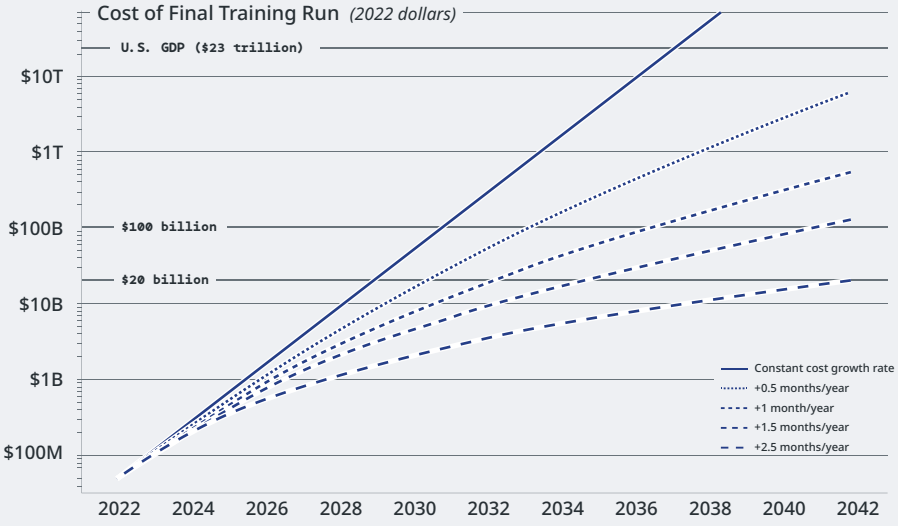
| Training compute in FLOP | $10^{26}$ | $10^{27}$ | $10^{28}$ | $10^{29}$ | $10^{30}$ |
|---|---|---|---|---|---|
| **7.0-month compute doubling** (median estimate) | 2024 | 2026 | 2028 | 2030 | 2032 |
| **5.7-month compute doubling** (high growth estimate) | 2023 | 2025 | 2027 | 2028 | 2030 |
| **8.6-month compute doubling** (low growth estimate) | 2024 | 2026 | 2029 | 2031 | 2034 |

## Appendix E: Tapered Cost Growth Projections

As costs continue to rise, the rate of growth in training costs (currently doubling every 9.7 months) may slow. Figure A.2 and Table A.6 show several possible projections for training costs under different scenarios for cost growth slowing over time. Tapered cost growth projections are shown under four scenarios in which the cost doubling period increases by 0.5, 1, 1.5, and 2.5 months per year. A straightforward projection in which cost growth increases at the current doubling rate of 9.7 months is shown in Figure A.2 for comparison.

The scenario in which the cost doubling rate changes by +1.5 months per year is used for the tapered cost growth projection shown in the body of this paper as an illustrative example of how cost growth could slow as costs approach the historical spending limits of private companies.

**FIGURE A.2 | FRONTIER MODEL TRAINING COST OVER TIME** (TAPERED COST GROWTH PROJECTIONS)




*The cost to train a frontier model is projected under four different scenarios of tapered cost growth, with the cost doubling rate increasing by 0.5, 1, 1.5, and 2.5 months per year. A straightforward projection of training cost with a constant 9.7-month doubling rate is shown for comparison.*

**TABLE A.6 | MODEL TRAINING COST AND COST DOUBLING RATE UNDER VARIOUS SCENARIOS OF TAPERED COST GROWTH**

| | 2024 | 2027 | 2030 | 2033 | 2036 |
|---|---|---|---|---|---|
| **Constant cost doubling rate of 9.7 months** | $280M (9.7 mos.) | $3.6B (9.7 mos.) | $50B (9.7 mos.) | $600B (9.7 mos.) | $8T (9.7 mos.) |
| **Doubling rate increases +0.5 months/year** | $260M (11 mos.) | $2.3B (12 mos.) | $16B (14 mos.) | $90B (15 mos.) | $400B (17 mos.) |
| **Doubling rate increases +1.0 months/year** | $240M (12 mos.) | $1.6B (15 mos.) | $7B (18 mos.) | $30B (21 mos.) | $80B (24 mos.) |
| **Doubling rate increases +1.5 months/year** | $220M (13 mos.) | $1.2B (17 mos.) | $4B (22 mos.) | $12B (26 mos.) | $30B (31 mos.) |
| **Doubling rate increases +2.5 months/year** | $200M (15 mos.) | $800M (22 mos.) | $2B (30 mos.) | $4B (37 mos.) | $8B (45 mos.) |

*Model training cost in 2022 dollars. Cost doubling rate in months.*

## Selected Bibliography

### AI Trends


- Alvi, Ali, and Paresh Kharya. "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model." Microsoft Research Blog, October 11, 2021. https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.
- *Artificial Intelligence Challenges and Opportunities for the Department of Defense: Hearing Before U.S. Senate Committee on Armed Services, Subcommittee on Cybersecurity*, 118th Cong. 2023. Statement of Jason Matheny, President and Chief Executive Officer, RAND Corporation. https://www.rand.org/pubs/testimonies/CTA2723-1.html.
- Amodei, Dario, and Danny Hernandez. "AI and Compute." OpenAI, May 16, 2018. https://openai.com/blog/ai-and-compute/.
- Anderljung, Markus, et al. *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. arXiv.org, September 4, 2023. https://arxiv.org/pdf/2307.03718.pdf.
- "Announcing Epoch's Updated Parameter, Compute and Data Trends Database." Epoch, October 23, 2023. https://epochai.org/blog/announcing-updated-pcd-database.
- Arya, Nisha. "GPT-4 Details Have Been Leaked!" KDNuggets, July 19, 2023. https://www.kdnuggets.com/2023/07/gpt4-details-leaked.html.
- Benaich, Nathan, and Ian Hogarth. "State of AI Report 2022." Air Street Capital, October 11, 2022. Presentation. https://www.stateof.ai/2022.
- Besiroglu, Tamay, et al. "Projecting Compute Trends in Machine Learning." Epoch, March 7, 2022. https://epochai.org/blog/projecting-compute-trends.
- Carey, Ryan. "Interpreting AI Compute Trends." AI Impacts, July 10, 2018, https://aiimpacts.org/interpreting-ai-compute-trends/.
- Cotra, Ajeya. "Forecasting TAI with Biological Anchors." July 2020, https://docs.google.com/document/d/1I-J6Sr-gPeXdSJugFulwIpvavc0atjHGM82QjIfUSBGQ.
- Cotra, Ajeya. "Appendices to Biological Anchors Report." July 2020. https://docs.google.com/document/d/1qjgB-koHO_kDuUYqv_Vws0fpf-dG5pTU4b8Uej6ff2Fg.
- Cottier, Ben. "Trends in the Dollar Training Cost of Machine Learning Systems." Epoch, January 31, 2023. https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.
- Davidson, Tom. "What a Compute-Centric Framework Says about Takeoff Speeds." Open Philanthropy, June 27, 2023. https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-take-off-speeds/.
- Dorner, Florian E. *Measuring Progress in Deep Reinforcement Learning Sample Efficiency*. arXiv.org, February 9, 2021. https://arxiv.org/abs/2102.04881.
- Erdil, Ege, and Tamay Besiroglu. *Algorithmic Progress in Computer Vision*. arXiv.org, December 16, 2022. https://arxiv.org/abs/2212.05153.
- Erdil, Ege, and Tamay Besiroglu. "Revisiting algorithmic progress. Epoch, December 12, 2022. https://epochai.org/blog/revisiting-algorithmic-progress.
- Epoch. Cost Estimates for GPT-4. 2023. https://colab.research.google.com/drive/1O99z9b1I5O66bT78r9ScslE_nOj5irN9?usp=sharing#scrollTo=Pqkx-E3NQocI.
- Epoch. ML Inputs Visualization. Accessed November 11, 2023. https://epochai.org/mlinputs/visualization?startDate=2015-9-1&startLargeScaleEra=2015-9-1&largeScaleAction=isolate&preset=Large%20scale%20models%20-%20compute&labelEras=false&labelPoints=true.
- Epoch. "Parameter, Compute and Data Trends in Machine Learning." Accessed December 20, 2023. https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/.
- Gholami, Amir. "AI and Memory Wall." Medium, March 29, 2021. https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8.
- Grace, Katja. *Algorithmic Progress in Six Domains*. Berkeley, CA: Machine Intelligence Research Institute, 2013. https://intelligence.org/files/AlgorithmicProgress.pdf.

- H. Dan. "How Much Did AlphaGo Zero Cost?" Dansplaining. Updated June 2020. https://www.yuzeh.com/data/agz-cost.html.
- Heim, Lennart. "Estimating PaLM's Training Cost." blog.heim.xyz, April 5, 2022. https://blog.heim.xyz/palm-training-cost/.
- Heim, Lennart. "This Can't Go On(?) - AI Training Compute Costs." blog.heim.xyz, June 1, 2023. https://blog.heim.xyz/this-cant-go-on-compute-training-costs/.
- Hernandez, Danny, and Tom B. Brown. *Measuring the Algorithmic Efficiency of Neural Networks*. arXiv.org, May 8, 2020. https://arxiv.org/abs/2005.04305.
- Ho, Anson, et al. *Algorithmic Progress in Language Modeling*. Forthcoming.
- Ho, Anson et al. *Limits to the Energy Efficiency of CMOS Microprocessors*. arXiv.org, December 14, 2023. https://arxiv.org/abs/2312.08595.
- Hobbhahn, Marius. "Marius' Disagreements with Bio Anchors." 2022. https://docs.google.com/document/d/1EKI7nU1LiknojKm68SIUBnP35iZz-JPO1WG8eLPiZTk0/edit#heading=h.xnee1t3imh4r.
- Hobbhahn, Marius, and Tamay Besiroglu. "Trends in GPU price-performance." Epoch, June 27, 2022. https://epochai.org/blog/trends-in-gpu-price-performance.
- Hobbhahn, Marius, and Tamay Besiroglu. "Predicting GPU performance." Epoch, December 1, 2022. https://epochai.org/blog/predicting-gpu-performance.
- Hobbhahn, Marius, et al. "Trends in Machine Learning Hardware." Epoch, November 9, 2023. https://epochai.org/blog/trends-in-machine-learning-hardware.
- Hoffman, Jordan, et al. *Training Compute-Optimal Large Language Models*. arXiv.org, March 29, 2022. https://arxiv.org/abs/2203.15556.
- Kaplan, Jared, et al. *Scaling Laws for Neural Language Models*. arXiv.org, January 23, 2020. https://arxiv.org/abs/2001.08361.
- Khan, Saif M., and Alexander Mann. *AI Chips: What They Are and Why They Matter*. Washington DC: Center for Security and Emerging Technology, April 2020. https://cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/.
- Lohn, Andrew J., and Micah Musser. *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* Washington: Center for Security and Emerging Technology, January 2022. https://cset.georgetown.edu/publication/ai-and-compute/.
- Mostaque, Emad (@EMostaque). "We actually used 256 A100s for this per the model card, 150k hours in total so at market price $600k." X (formerly Twitter), August 28, 2022. https://twitter.com/EMostaque/status/1563870674111832066.
- Ngo, Richard. "Visualizing the Deep Learning Revolution." Medium, January 5, 2023. https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5.
- OpenAI. "GPT-4 System Card." March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.
- Owen, David. "How Predictable Is Language Model Benchmark Performance?" Epoch, June 9, 2023, updated January 11, 2024. https://epochai.org/blog/how-predictable-is-language-model-benchmark-performance.
- Peleg, Yam (@Yampeleg). "GPT-4's details are leaked. It is over. Everything is here." X (formerly Twitter), July 10, 2023. https://archive.is/2RQ8X.
- *Preparing the Federal Response to Advanced Technologies: Hearing Before U.S. Senate Committee on Homeland Security and Governmental Affairs, Subcommittee on Emerging Threats and Spending Oversight*, 118th Cong. 2023. Statement of Jeff Alstott, Senior Information Scientist; Professor of Policy Analysis, Pardee RAND Graduate School. https://www.rand.org/pubs/testimonies/CTA2953-1.html.
- Sevilla, Jaime, and Edu Roldán. "An Interactive Model of AI Takeoff Speeds." Epoch, January 24, 2023. https://epochai.org/blog/interactive-model-of-takeoff-speeds.
- Sevilla, Jaime, et al. *Compute Trends across Three Eras of Machine Learning*. arXiv.org, March 9, 2022 https://arxiv.org/abs/2202.05924.
- Sharir, Or, et al. *The Cost of Training NLP Models*. arXiv.org, April 2020. https://arxiv.org/pdf/2004.08900.pdf?.
- Tayar, David (@davidtayar5). "Remainder of note." X (formerly Twitter), February 20, 2023. https://twitter.com/davidtayar5/status/1627690520456691712.

- Villalobos, Pablo and Anson Ho. "Trends in Training Dataset Sizes." Epoch, September 20, 2022. https://epochai.org/blog/trends-in-training-dataset-sizes.
- Villalobos, Pablo, et al. *Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning*. arXiv.org, October 26, 2022. https://arxiv.org/abs/2211.04325.
- Wang, Ken. "DeepMind Achieved StarCraft II GrandMaster Level, but at What Cost?" Medium, January 4, 2020. https://medium.com/swlh/deepmind-achieved-starcraft-ii-grandmaster-level-but-at-what-cost-32891dd990e4.


### U.S. Government Policy


- Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification. 87 Fed. Reg. 62,186 (Oct. 13, 2022) (to be codified at 15 C.F.R. pt. 734, 736, 740, 742, 744, 762, 772, and 774). https://www.federalregister.gov/public-inspection/2022-21658/additional-export-controls-certain-advanced-computing-and-semiconductor-manufacturing-items.
- Export Controls on Semiconductor Manufacturing Items Interim Final Rule. 88 Fed. Reg. 73,424 (Oct. 25, 2023) (to be codified at 15 C.F.R. pt. 734, 736, 740, 742, 744, 772, 774). https://www.federalregister.gov/documents/2023/10/25/2023-23049/export-controls-on-semiconductor-manufacturing-items.
- Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections. 88 Fed. Reg. 73,458 (Oct. 25, 2023) (to be codified at 15 C.F.R. pt. 732, 734, 736, 740, 742, 744, 746, 748, 758, 770, 772, and 774). https://www.federalregister.gov/documents/2023/10/25/2023-23055/implementation-of-additional-export-controls-certain-advanced-computing-items-supercomputer-and.
- Rasser, Martijn. "A Conversation with Under Secretary of Commerce Alan F. Estevez." CNAS, October 27, 2022. https://www.cnas.org/publications/transcript/a-conversation-with-under-secretary-of-commerce-alan-f-estevez.
- Taking Additional Steps to Address the National Emergency with Respect to Significant Malicious Cyber-Enabled Activities. 89 Fed. Reg. 5,698 (Jan. 29, 2024) (to be codified at 15 C.F.R. pt. 7). https://www.federalregister.gov/documents/2024/01/29/2024-01580/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious.
- The White House. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." October 30, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
- The White House. "National Artificial Intelligence Research Resource Task Force Releases Final Report." January 24, 2023. https://www.whitehouse.gov/ostp/news-updates/2023/01/24/national-artificial-intelligence-research-resource-task-force-releases-final-report/.

1. Jaime Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*, arXiv.org, March 9, 2022, https://arxiv.org/abs/2202.05924.

2. This estimate comes from a similar methodology to Cottier (2023), updated with the most recent estimates for compute growth and hardware price-performance. For more on this estimate, see this paper's section, Current Best Estimates and Assumptions: Cost Growth. Ben Cottier, "Trends in the dollar training cost of machine learning systems," Epoch, January 31, 2023, https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.

3. OpenAI's CEO Sam Altman stated that the company spent $100M to train GPT-4, although he did not give further details, and this may include earlier experiments in addition to the final training run. Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, April 17, 2023, https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

4. For example, OpenAI's then-CEO Sam Altman stated in November 2023 that the company was working on GPT-5. Madhumita Murgia, "OpenAI CEO Sam Altman wants to build AI 'superintelligence,'" Ars Technica, November 14, 2023, https://arstechnica.com/ai/2023/11/openai-ceo-sam-altman-wants-to-build-ai-superintelligence/. See also David Tayar (@davidtayar5), "Remainder of note," X (formerly Twitter), February 20, 2023, https://twitter.com/davidtayar5/status/1627690520456691712; Kyle Wiggers et al., "Anthropic's $5B, 4-year plan to take on OpenAI," *Tech Crunch*, April 6, 2023, https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai/.

5. Nvidia reportedly is expected to ship 550,000 state-of-the-art H100 graphics processing units (GPUs) in 2023. Madhumita Murgia et al., "Saudi Arabia and UAE race to buy Nvidia chips to power AI ambitions," *Financial Times*, August 14, 2023, https://www.ft.com/content/c93d2a76-16f3-4585-af61-86667c5090ba.

6. "National Research Cloud Call to Action," Stanford University Human-Centered Artificial Intelligence, https://hai.stanford.edu/national-research-cloud-joint-letter.

7. Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over."

8. "Expanding access to safer AI with Amazon," Anthropic, September 25, 2023, https://www.anthropic.com/index/anthropic-amazon; Krystal Hu, "Google agrees to invest up to $2 billion in OpenAI rival Anthropic," Reuters, October 27, 2023, https://www.reuters.com/technology/google-agrees-invest-up-2-bln-openai-rival-anthropic-wsj-2023-10-27/.

9. The White House, "National Artificial Intelligence Research Resource Task Force Releases Final Report," January 24, 2023, https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf; The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

10. Meta, "Statement of Support for Meta's Open Approach to Today's AI," July 18, 2023, https://about.fb.com/news/2023/07/llama-2-statement-of-support/.

11. OpenAI "GPT-4 System Card," March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf; Anthropic, "Anthropic's Responsible Scaling Policy," September 19, 2023, https://www.anthropic.com/news/anthropics-responsible-scaling-policy; OpenAI, "Preparedness Framework (Beta)," https://cdn.openai.com/openai-preparedness-framework-beta.pdf, accessed December 19, 2023.

12. Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence, August 18, 2021, https://arxiv.org/pdf/2108.07258.pdf.

13. Marie Lamensch, "Generative AI Tools Are Perpetuating Harmful Gender Stereotypes," Centre for International Governance Innovation, June 14, 2023, https://www.cigionline.org/articles/generative-ai-tools-are-perpetuating-harmful-gender-stereotypes/.

14. OpenAI, "GPT-4 System Card"; Daniil A. Boiko et al., *Emergent autonomous scientific research capabilities of large language models*, arXiv.org, April 11, 2023, https://arxiv.org/pdf/2304.05332.pdf; "Frontier Threats Red Teaming for AI Safety," Anthropic, July 26, 2023, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety; Emily H. Soice et al., *Can large language models democratize access to dual-use biotechnology*, arXiv.org, June 6, 2023, https://arxiv.org/pdf/2306.03809.pdf; Jonas B. Sandbrink, *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools*, arXiv.org, June 24, 2023, https://arxiv.org/abs/2306.13952; Fabio Urbina et al., "Dual use of artificial-intelligence-powered drug discovery," *Natural Machine Intelligence* 4 (March 2022) 189–191, https://www.nature.com/articles/s42256-022-00465-9.epdf.

15. Emad Mostaque, "Stable Diffusion Public Release," Stability AI, August 22, 2022, https://stability.ai/blog/stable-diffusion-public-release; Benj Edwards, "With Stable Diffusion, You May Never Believe What You See Online Again," Ars Technica, September 6, 2022, https://arstechnica.com/information-technology/2022/09/with-stable-diffusion-you-may-never-believe-what-you-see-online-again/3/; Steve Dent, "Stable Diffusion update removes ability to copy artist styles or make NSFW works," *Engadget*, November 25, 2022, https://www.engadget.com/stable-diffusion-version-2-update-artist-styles-nsfw-work-124513511.html.

16. OpenAI "GPT-4 System Card;" Anthropic, "Model Card and Evaluations for Claude Models," July 6, 2023, https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf; Hugo Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," Meta, July 18, 2023, https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

17. Toby Shevlane, *Structured access: an emerging paradigm for safe AI deployment*, arXiv.org, January 13, 2022, https://arxiv.org/abs/2201.05159.

18. Benj Edwards, "With Stable Diffusion, You May Never Believe What You See Online Again"; Market Trends, "Stable Diffusion NSFW and Its Alternatives," Analytics Insight, May 17, 2023, https://www.analyticsinsight.net/stable-diffusion-nsfw-and-its-alternatives/.

19. georgesung, "Llama-2 7B uncensored - QLoRA fine-tune on wizard_vicuna_70k_unfiltered," reddit, https://www.reddit.com/r/LocalLLaMA/comments/154rqay/llama2_7b_uncensored_qlora_finetune_on_wizard/; Anjali Gopal et al., *Will releasing the weights of future large language models grant widespread access to pandemic agents?* arXiv.org, October 25, 2023, https://arxiv.org/pdf/2310.18233.pdf.

20. "Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification," document no. 2022-21658, October 13, 2022, https://www.federalregister.gov/public-inspection/2022-21658/additional-export-controls-certain-advanced-computing-and-semiconductor-manufacturing-items; "Export Controls on Semiconductor Manufacturing Items Interim Final Rule," October 17, 2023, https://www.bis.doc.gov/index.php/documents/federal-register-notices-1/3352-10-16-23-semiconductor-equipment-controls/file.

21. Martijn Rasser, "A Conversation with Under Secretary of Commerce Alan F. Estevez," CNAS, October 27, 2022, https://www.cnas.org/publications/transcript/a-conversation-with-under-secretary-of-commerce-alan-f-estevez.

22. "Tech Insights Finds SMIC 7nm (N+2) in Huawei Mate 60 Pro," Tech Insights, https://www.techinsights.com/blog/techinsights-finds-smic-7nm-n2-huawei-mate-60-pro.

23. Meta's open-source model Llama 2 was released approximately seven months after GPT-3.5, to which it is roughly comparable in terms of performance. This time lag may not be generalizable. The time lag from an initial closed model to an open-source model of approximately equivalent capabilities has varied widely by model. It took 35 months for an open-source equivalent of AlphaFold to be released and 14 months for an open-source equivalent of GPT-3 to be released. Nathan Benaich and Ian Hogarth, "State of AI Report 2022," presentation Air Street Capital, October 11, 2022, https://www.stateof.ai/2022, slides 34–36; Ankur A. Patel and Saleem Maroof, "LLaMA 1 vs LLaMA 2: A Deep Dive into Meta's LLMs," Ankur's Newsletter, August 15, 2023, https://www.ankursnewsletter.com/p/llama-1-vs-llama-2-a-deep-dive-into.

24. Rainer Strack et al., "The Future of Jobs in the Era of AI," Boston Consulting Group, March 2021, https://web-assets.bcg.com/f5/e7/9aa9f81a446198ac5402aaf97a87/bcg-the-future-of-jobs-in-the-era-of-ai-mar-2021-r-r.pdf; Annie Lowrey, "How ChatGPT Will Destabilized White-Collar Work," *The Atlantic*, January 20, 2023, https://www.theatlantic.com/ideas/archive/2023/01/chatgpt-ai-economy-automation-jobs/672767/; and Claire Cain Miller and Courtney Cox, "In Reversal Because of A.I., Office Jobs Are Now More at Risk," *The New York Times*, August 24, 2023, https://www.nytimes.com/2023/08/24/upshot/artificial-intelligence-jobs.html.

25. Marvin Minsky and Seymour A. Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge MA: MIT Press, 2017), https://direct.mit.edu/books/book/3132/PerceptronsAn-Introduction-to-Computational; Jeffrey Dean, *The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design*, arXiv.org, November 13, 2019, https://arxiv.org/pdf/1911.05289.pdf.

26. Ben Buchanan, "The AI Triad and What It Means for National Security Strategy" (Washington: Center for Security and Emerging Technology, August 2020), https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/.

27. Richard Ngo, "Visualizing the deep learning revolution," *Medium*, January 5, 2023, https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5.

28. Jared Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv.org, January 23, 2020, https://arxiv.org/abs/2001.08361; Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models*, arXiv.org, March 29, 2022, https://arxiv.org/abs/2203.15556.

29. Aarohi Srivastava et al., *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models,* arXiv.org, June 12, 2023, https://arxiv.org/abs/2206.04615; Dan Hendrycks et al., *Measuring Massive Multitask Language Understanding,* arXiv.org, January 12, 2021, https://arxiv.org/abs/2009.03300.

30. David Owen, "How Predictable Is Language Model Benchmark Performance?" Epoch, June 9, 2023, updated January 11, 2024, https://epochai.org/blog/extrapolating-performance-in-language-modelling-benchmarks.

31. Shana Lynch, "AI Benchmarks Hit Saturation," Stanford University, April 3, 2023, https://hai.stanford.edu/news/ai-benchmarks-hit-saturation; Maslej et al., "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf, 12; Andrew J. Lohn, "Scaling AI: Cost and Performance of AI at the Leading Edge," (Washington: Center for Security and Emerging Technology, December 2023), https://cset.georgetown.edu/publication/scaling-ai.

32. Ali Alvi and Paresh Kharya, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model," Microsoft Research Blog, October 11, 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/; Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, arXiv.org, December 31, 2020, https://arxiv.org/pdf/2101.00027.pdf; Aakanksha Chowdhery et al., *PaLM: Scaling Language Modeling with Pathways*, arXiv.org, April 19, 2022, https://arxiv.org/pdf/2204.02311.pdf, 3, 36, 65.

33. Nisha Arya, "GPT-4 Details Have Been Leaked!" *KDNuggets*, July 19, 2023, https://www.kdnuggets.com/2023/07/gpt4-details-leaked.html; Yam Peleg (@Yampeleg), "GPT-4's details are leaked. It is over. Everything is here," X (formerly Twitter), July 10, 2023, https://archive.is/2RQ8X.

34. For example, Anthropic is reportedly raising a billion dollars to train its next-generation model, "Claude Next." (Kyle Wiggers et al., "Anthropic's $5B, 4-year plan to take on OpenAI.") Leaked investor reports have stated that OpenAI is training GPT-5. (Tayar, "Remainder of note.")

35. Hardware performance is measured in floating-point operations per second (FLOP/s). Compute is measured in floating- point operations (FLOP), or FLOP/s multiplied by time (in seconds).

36. Konstantin Pilz et al., *Increased Compute Efficiency and the Diffusion of AI Capabilities*, arXiv.org, November 26, 2023, https://arxiv.org/abs/2311.15377.

37. Danny Hernandez and Tom B. Brown, *Measuring the Algorithmic Efficiency of Neural Networks*, arXiv.org, May 8, 2020, https://arxiv.org/abs/2005.04305; Florian E. Dorner, *Measuring Progress in Deep Reinforcement Learning Sample Efficiency*, arXiv.org, February 9, 2021, https://arxiv.org/abs/2102.04881; Ege Erdil and Tamay Besiroglu, *Algorithmic progress in computer vision*, arXiv.org, December 16, 2022, https://arxiv.org/abs/2212.05153; and Anson Ho et al., *Algorithmic Progress in Language Modeling*, forthcoming.

38. Some researchers refer to this metric of compute times algorithmic efficiency as "effective compute" or "effective FLOP." See Ajeya Cotra, "Forecasting TAI with biological anchors," July 2020, https://docs.google.com/document/d/1IJ6Sr-gPeXdSJugFulwIpvavc0atjHGM82QjIfUSBGQ; Tom Davidson, "What a compute-centric framework says about takeoff speeds," Open Philanthropy, June 27, 2023, https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/; and Jaime Sevilla and Edu Roldán, "An Interactive Model of AI Takeoff Speeds," Epoch, January 24, 2023, https://epochai.org/blog/interactive-model-of-takeoff-speeds.

39. Dario Amodei and Danny Hernandez, "AI and Compute," OpenAI, May 16, 2018, https://openai.com/blog/ai-and-compute/.

40. Ryan Carey, "Interpreting AI Compute Trends," AI Impacts, July 10, 2018, https://aiimpacts.org/interpreting-ai-compute-trends/.

41. Ajeya Cotra, "Appendices to biological anchors report," July 2020, https://docs.google.com/document/d/1qjgBkoHO_kDuUYqy_Vws0fpf-dG5pTU4b8Uej6ff2Fg.

42. Andrew J. Lohn and Micah Musser, *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* (Washington: Center for Security and Emerging Technology, January 2022), 9, https://cset.georgetown.edu/publication/ai-and-compute/.

43. Amodei and Hernandez, "AI and Compute."

44. Lohn and Musser, *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?*

45. Jaime Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*.

46. Marius Hobbhahn and Tamay Besiroglu, "Trends in GPU price-performance," Epoch, June 27, 2022, https://epochai.org/blog/trends-in-gpu-price-performance.

47. Cottier, "Trends in the dollar training cost of machine learning systems."

48. Lennart Heim, "This can't go on(?)—AI Training Compute Costs," blog.heim.xyz, June 1, 2023, https://blog.heim.xyz/this-cant-go-on-compute-training-costs/.

49. Marius Hobbhahn et al., "Trends in Machine Learning Hardware," Epoch, November 9, 2023, https://epochai.org/blog/trends-in-machine-learning-hardware.

50. Epoch, ML Inputs Visualization, accessed November 11, 2023, https://epochai.org/mlinputs/visualization?startDate=2015-9-1&startLargeScaleEra=2015-9-1&largeScaleAction=isolate&preset=Large%20scale%20models%20-%20compute&labelEras=false&labelPoints=true.

51. Hernandez and Brown, *Measuring the Algorithmic Efficiency of Neural Networks*.

52. Erdil and Besiroglu, *Algorithmic progress in computer vision*.

53. Florian E. Dorner, *Measuring Progress in Deep Reinforcement Learning Sample Efficiency*.

54. 95 percent confidence interval of 5.3 to 13.0 months. Ho et al., *Algorithmic Progress in Language Modeling*.

55. Katja Grace, *Algorithmic Progress in Six Domains* (Berkeley, CA: Machine Intelligence Research Institute, 2013), https://intelligence.org/files/AlgorithmicProgress.pdf.

56. Cotra, "Forecasting TAI with biological anchors;" Marius Hobbhahn, "Marius' disagreements with bio anchors," 2022, https://docs.google.com/document/d/1EKI7nU1LiknojKm68SIUBnP35iZzJPO1WG8eLPiZTk0/edit#heading=h.xnee1t3imh4r.

57. Cotra, "Forecasting TAI with biological anchors."

58. Carey, "Interpreting AI Compute Trends;" Lohn and Musser, *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?*; Tamay Besiroglu et al., "Projecting compute trends in Machine Learning," Epoch, March 7, 2022, https://epochai.org/blog/projecting-compute-trends.

59. Pablo Villalobos and Anson Ho, "Trends in Training Dataset Sizes," Epoch, 2022, https://epochai.org/blog/trends-in-training-dataset-sizes.

60. Epoch, ML Inputs Visualization.

61. Cottier, "Trends in the dollar training cost of machine learning systems."

62. Subtracting the 2.1-year doubling rate (0.143 OOMs/yr) for machine learning GPU price-performance from the 7.0-month compute doubling rate (0.516 OOMs/yr) yields a cost growth rate of 0.373 OOMs/yr, or a doubling of costs every 9.7 months. (0.516 OOMs/yr compute growth – 0.143 OOMs/yr hardware improvements = 0.373 OOMs/yr cost growth.)

63. Using a similar approach, high and low cost growth estimates are derived using the error propagation formula $\sigma_z = \sqrt{(\sigma_x^2 + \sigma_y^2)}$.

64. Ho et al., *Algorithmic Progress in Language Modeling*.

65. "Announcing Epoch's Updated Parameter, Compute and Data Trends Database," Epoch, October 23, 2023, https://epochai.org/blog/announcing-updated-pcd-database.

66. Epoch, Cost Estimates for GPT-4, 2023, https://colab.research.google.com/drive/1O99z9b1I5O66bT78r9ScslE_nOj5irN9#scrollTo=CytB-eruRSGB.

67. Epoch, Cost Estimates for GPT-4.

68. Epoch notes, "This number is an estimate based on limited evidence," and cites a 95 percent confidence interval of between $2.4 \times 10^{25}$ to $2.0 \times 10^{26}$ FLOP. Epoch, Epoch Database, accessed February 26, 2024, https://epochai.org/data/epochdb/table; Epoch, Gemini_compute, accessed February 26, 2024, https://colab.research.google.com/drive/1sfG91UfiYpEYnj_xB5YRy07T5dv-9O_c#scrollTo=ZW80nL_0Yp5O.

69. The 95 percent confidence intervals for effective compute combine algorithmic efficiency and compute growth high and low estimates using the error propagation formula $\sigma_z = \sqrt{(\sigma_x^2 + \sigma_y^2)}$.

70. Ben Blanchard and Sarah Wu, "TSMC cuts capex on tool delays, demand woes; cautious on outlook," Reuters, October 13, 2023, https://www.reuters.com/technology/tsmc-q3-profit-jumps-80-beats-market-expectations-2022-10-13/; Chang Chien-chung and Frances Huang, "TSMC Expects Q4 Sales Rise 11.1% from Q3, Leaves Capex Plan Unchanged," Focus Taiwan, October 19, 2023, https://focustaiwan.tw/business/202310190017.

71. Meta, "Meta Reports Fourth Quarter and Full Year 2022 Results" February 1, 2023, https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx; Dan Swinhoe, "Meta's CapEx Drops Almost $3bn during Data Center Construction Pause," Data Center Dynamics, October 26, 2023, https://www.datacenterdynamics.com/en/news/metas-capex-drops-almost-3bn-during-data-center-construction-pause/.

72. Sebastian Moss, "AWS revenues increase 12% 'significant capital expense' goes to generative AI," Data Center Dynamics, August 4, 2023, https://www.datacenterdynamics.com/en/news/aws-revenues-increase-12-significant-capital-expense-goes-to-generative-ai/.

73. Grace Dean, "Meta has pumped $36 billion into its metaverse and VR businesses since 2019. These 4 charts show the scale of its extreme spending—and huge losses," *Business Insider,* October 29, 2022, https://www.businessinsider.com/charts-meta-metaverse-spending-losses-reality-labs-vr-mark-zuckerberg-2022-10.

74. "Manhattan Project," Comprehensive Nuclear-Test-Ban Treaty Organization, https://www.ctbto.org/nuclear-testing/history-of-nuclear-testing/manhattan-project/, captured by the Internet Archive at https://web.archive.org/web/20220423080628/https://www.ctbto.org/nuclear-testing/history-of-nuclear-testing/manhattan-project/.

75. John Curatola, "Delivering the Atomic Bombs: The Silverplate B-29," The National World War II Museum, August 11, 2023, https://www.nationalww2museum.org/war/articles/delivering-atomic-bombs-silverplate-b-29.

76. "Apollo Program Budget Appropriations ($000)," NASA, https://history.nasa.gov/SP-4029/Apollo_18-16_Apollo_Program_Budget_Appropriations.htm.

77. Office of the Under Secretary of Defense (Comptroller), *National Defense Budget Estimates for FY 2021*, U.S. Department of Defense, April 2020, https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2021/FY21_Green_Book.pdf, Table 7-7.

78. Jonathan Masters and Will Merrow, "How much Aid Has the U.S. Sent Ukraine? Here are Six Charts.," Council on Foreign Relations, September 21, 2023, https://www.cfr.org/article/how-much-aid-has-us-sent-ukraine-here-are-six-charts.

79. DoD spending on "overseas contingency operations," a budgetary category that includes direct expenditures for Iraq, Afghanistan, and other overseas operations, was $2.1T from fiscal year 2001 to 2022. This does not include long-term costs such as veterans' care. Neta C. Crawford, "The U.S. Budgetary Cost of the Post-9/11 Wars," (Providence RI: Brown University, September 1, 2021), https://watson.brown.edu/costsofwar/figures/2021/BudgetaryCosts.

80. "COVID-19 Relief: Funding and Spending as of Jan. 31, 2023," U.S. Government Accountability Office, February 2023, https://www.gao.gov/assets/gao-23-106647.pdf.

81. Davidson, "What a compute-centric framework says about takeoff speeds." Sevilla and Roldán have released an interactive online tool to explore various scenarios using Davidson's analysis. Sevilla and Roldán, "An Interactive Model of AI Takeoff Speeds."

82. Marius Hobbhahn and Tamay Besiroglu, "Predicting GPU performance," Epoch, December 1, 2022, https://epochai.org/blog/predicting-gpu-performance.

83. Anson Ho et al., *Limits to the Energy Efficiency of CMOS Microprocessors*, arXiv.org, December 14, 2023, https://arxiv.org/abs/2312.08595.

84. For example, the Biden administration's October 2023 executive order raised concern about "substantial security risks" from dual-use foundation models with "widely available model weights" and solicited input on "appropriate policy and regulatory approaches." The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."

85. The 95 percent confidence intervals for effective compute combine algorithmic efficiency and compute growth high and low estimates using the error propagation formula $\sigma_z = \sqrt{(\sigma_x^2 + \sigma_y^2)}$.

86. Meta's open-source model Llama 2 was released approximately seven months after GPT-3.5, to which it is roughly comparable in terms of performance. Nathan Benaich and Ian Hogarth, "State of AI Report 2022," slides 34–36; Ankur A. Patel and Saleem Maroof, "LLaMA 1 vs LLaMA 2: A Deep Dive into Meta's LLMs."

87. Emad Mostaque (@EMostaque), "We actually used 256 A100s for this per the model card, 150k hours in total so at market price $600k," X (formerly Twitter), August 28, 2022, https://twitter.com/EMostaque/status/1563870674111832066.

88. Hobbhahn et al., "Trends in Machine Learning Hardware."

89. The 95 percent confidence intervals for the combined algorithmic efficiency and hardware improvement rates are derived using the error propagation formula $\sigma_z = \sqrt{(\sigma_x^2 + \sigma_y^2)}$.

90. Lohn, "Scaling AI: Cost and Performance of AI at the Leading Edge."

91. For example, see the GPT-4 leak. Arya, "GPT-4 Details Have Been Leaked!"

92. Tim Fist et al., "China Firms are Evading Chip Controls," *Foreign Policy*, June 21, 2023, https://foreignpolicy.com/2023/06/21/china-united-states-semiconductor-chips-sanctions-evasion/; Tim Fist and Erich Grunewald, "Preventing AI Chip Smuggling to China," CNAS, October 24, 2023, https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china; and Josh Ye et al., "Focus: Inside China's underground market for high-end Nvidia AI chips," Reuters, June 20, 2023, https://www.reuters.com/technology/inside-chinas-underground-market-high-end-nvidia-ai-chips-2023-06-19/.

93. Hanna Dohmen et al., "Controlling Access to Advanced Compute via the Cloud: Options for U.S. Policy Makers, Part 1," (Washington: Center for Security and Emerging Technology, May 15, 2023), https://cset.georgetown.edu/article/controlling-access-to-advanced-compute-via-the-cloud/; Hanna Dohmen et al., "Controlling Access to Advanced Compute via the Cloud: Options for U.S. Policy Makers, Part 2," (Washington: Center for Security and Emerging Technology, June 5, 2023), https://cset.georgetown.edu/article/controlling-access-to-compute-via-the-cloud-options-for-u-s-policymakers-part-ii/. The Bureau of Industry and Security is also seeking public comments for suggestions on how to address issues of access to "infrastructure-as-a-service." Bureau of Industry and Security, Department of Commerce, "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections Interim Final Rule (AC/S IFR)," October 27, 2023, https://www.bis.doc.gov/index.php/documents/federal-register-notices-1/3353-2023-10-16-advanced-computing-supercomputing-ifr/file. The Biden administration's October 2023 AI executive order also orders proposals for regulations of foreign use of compute through infrastructure-as-a-service that would prevent their use for training potentially malicious models. These proposals would include tracking the identity of the customer. The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The Bureau of Industry and Security has also requested comment on a proposed rule requiring infrastructure-as-a-service providers to verify the identity of foreign customers and report to the U.S. government if foreign customers attempt to train a large AI model that could be used for cyber attacks. Bureau of Industry and Security, Department of Commerce, "Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities," January 29, 2024, https://www.federalregister.gov/documents/2024/01/29/2024-01580/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious.

94. Francisco Pires, "China's ByteDance has Gobbled Up $1 Billion of Nvidia GPUs for AI This Year," Tom's Hardware, June 16, 2023, https://www.tomshardware.com/news/chinas-bytedance-has-gobbled-up-dollar1-billion-of-nvidia-gpus-for-ai-this-year.

95. Josh Ye et al., "Focus: Inside China's underground market for high-end Nvidia AI chips."

96. Reuters, "Nvidia tweaks flagship H100 chip for export to China as H800," *South China Morning Post*, March 22, 2023, https://www.scmp.com/tech/tech-war/article/3214379/nvidia-tweaks-flagship-h100-chip-export-china-h800.

97. Qianer Liu and Hannah Murphy, "China's Internet Giants Order $5bn of Nvidia Chips to Power AI Ambitions," *Financial Times*, August 9, 2023, https://www.ft.com/content/9dfee156-4870-4ca4-b67d-bb5a285d855c.

98. Rasser, "A Conversation with Under Secretary of Commerce Alan F. Estevez."

99. Tim Kelly et al., "As Japan aligns with U.S. chip curbs on China, some in Tokyo feel uneasy," Reuters, July 24, 2023, https://www.reuters.com/technology/space/japan-aligns-with-us-chip-curbs-china-some-tokyo-feel-uneasy-2023-07-24; Pieter Haeck, "Dutch slap new restrictions on chips exports to China," *Politico*, June 30, 2023, https://www.politico.eu/article/dutch-impose-export-controls-on-chips-printing-equipment-to-china/.

100. "TechInsights Confirming SMIC N+2 7nm in Huawei Mate 60 Pro," Tech Insights, https://www.techinsights.com/blog/techinsights-confirming-smic-n2-7nm-huawei-mate-60-pro.

101. Matthew Schleich and William Alan Reinsch, "Contextualizing the National Security Concerns of China's Domestically Produced High-End Chip," Center for Strategic and International Studies, September 26, 2023, https://www.csis.org/analysis/contextualizing-national-security-concerns-over-chinas-domestically-produced-high-end-chip.

102. Saif M. Khan and Alexander Mann, *AI Chips: What They Are and Why They Matter* (Washington: Center for Security and Emerging Technology, April 2020), 26, https://cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/.

103. Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, arXiv.org, September 4, 2023, https://arxiv.org/pdf/2307.03718.pdf; *Artificial Intelligence Challenges and Opportunities for the Department of Defense: Hearing Before U.S. Senate Committee on Armed Services, Subcommittee on Cybersecurity*, 118th Cong. (2023) (statement of Jason Matheny, President and Chief Executive Officer, RAND Corporation), https://www.rand.org/pubs/testimonies/CTA2723-1.html; *Preparing the Federal Response to Advanced Technologies: Hearing Before U.S. Senate Committee on Homeland Security and Governmental Affairs, Subcommittee on Emerging Threats and Spending Oversight*, 118th Cong. (2023) (statement of Jeff Alstott, Senior Information Scientist; Professor of Policy Analysis, Pardee RAND Graduate School), https://www.rand.org/pubs/testimonies/CTA2953-1.html; "A Responsible AI Act," Center for AI Policy, https://www.aipolicy.us/work.

104. The $10^{26}$ operations threshold applies to both integer and floating-point operations. The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," section 4.2(b)(i).

105. The 95 percent confidence intervals for the combined algorithmic efficiency and hardware improvement rates are derived using the error propagation formula $\sigma_z = \sqrt{(\sigma_x^2 + \sigma_y^2)}$.

106. Anthropic, "Anthropic's Responsible Scaling Policy"; METR, "Responsible Scaling Policies (RSPs)," September 26, 2023, https://metr.org/blog/2023-09-26-rsp/.

107. The Biden administration's executive order tasked the Commerce Department to solicit input on "appropriate policy and regulatory approaches related to dual-use foundation models for which the model weights are widely available." The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," section 4.6.

108. Philip E. Ross, "5 Commandments: The Rules Engineers Live By Weren't Always Set in Stone," *IEEE Spectrum*, December 1, 2003, https://spectrum.ieee.org/semiconductors/materials/5-commandments.

109. Arjun Kharpal, "Apple Supplier TSMC to Build a $12 Billion Chip Factory in the U.S.," CNBC, May 15, 2020, https://www.cnbc.com/2020/05/15/tsmc-to-build-us-chip-factory.html; Mark Lapedus and Ann Steffora Mutschler, "Regaining the Edge in U.S. Chip Manufacturing," Semiconductor Engineering, October 26, 2020, https://semiengineering.com/can-the-u-s-regain-its-edge-in-chip-manufacturing/; "Chipmaking Is Being Redesigned. Effects Will Be Far-Reaching," *The Economist*, January 21, 2021, https://www.economist.com/business/2021/01/23/chipmaking-is-being-redesigned-effects-will-be-far-reaching; AleksandarK, "TSMC Completes Its Latest 3 nm Factory, Mass Production in 2022," TechPowerUp, November 27, 2020, https://www.techpowerup.com/275255/tsmc-completes-its-latest-3-nm-factory-mass-production-in-2022; and "Samsung Considers Austin for $17 Billion Chip Plant, Seeks Tax Breaks of at Least $806 Million," Reuters, February 4, 2021, https://www.cnbc.com/2021/02/05/samsung-considers-austin-for-17-billion-chip-plant.html.

110. Hobbhahn and Besiroglu, "Predicting GPU performance."

111. Ho et al., *Limits to the Energy Efficiency of CMOS Microprocessors*.

112. Hoffman et al., *Training Compute-Optimal Large Language Models*.

113. Kaplan et al., *Scaling Laws for Neural Language Models*.

114. Villalobos and Ho, "Trends in Training Dataset Sizes."

115. Pablo Villalobos et al. *Will We Run Out of Data? An analysis of the limits of scaling datasets in Machine Learning*, arXiv.org, October 26, 2022, https://arxiv.org/abs/2211.04325.

116. Ali Alvi and Paresh Kharya, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model;" Amir Gholami, "AI and Memory Wall," Medium, March 29, 2021, https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8.

117. Lohn and Musser, *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?*

118. Epoch, ML Inputs Visualization.

119. Epoch, ML Inputs Visualization.

120. Hobbhahn et al., "Trends in Machine Learning Hardware."

121. Ho et al., *Algorithmic Progress in Language Modeling*.

122. Villalobos and Ho, "Trends in Training Dataset Sizes."

123. Villalobos and Ho, "Trends in Training Dataset Sizes."

124. Dan H., "How Much Did AlphaGo Zero Cost?," Dansplaining, updated June 2020, https://www.yuzeh.com/data/agz-cost.html.

125. Carey, "Interpreting AI Compute Trends."

126. Ken Wang, "DeepMind achieved StarCraft II Grandmaster Level, but at what cost?," Medium, January 4, 2020, https://medium.com/swlh/deepmind-achieved-starcraft-ii-grandmaster-level-but-at-what-cost-32891dd990e4.

127. Or Sharir et al., *The Cost of Training NLP Models*, arXiv.org, April 2020, https://arxiv.org/pdf/2004.08900.pdf.

128. Chuan Li, "OpenAI's GPT-3 Language Model: A Technical Overview," Lambda Labs, June 3, 2020, https://lambdalabs.com/blog/demystifying-gpt-3/.

129. While Lohn and Musser use the lower estimate of $1.65M to train GPT-3 for their cost projections so as to not overstate the rising cost of compute, they assess Li's estimate of $4.6M as more reasonable.

130. Lennart Heim, "Estimating PaLM's training cost," blog.heim.xyz, April 5, 2022, https://blog.heim.xyz/palm-training-cost/.

131. Epoch, Cost Estimates for GPT-4.

132. Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over."

133. Epoch, Cost Estimates for GPT-4.

## About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts and the public with innovative, fact-based research, ideas and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, DC, and was established in February 2007 by co-founders Kurt M. Campbell and Michèle A. Flournoy. CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and non-partisan.



### CNAS Editorial

**DIRECTOR OF STUDIES**
Paul Scharre

**DEPUTY DIRECTOR OF STUDIES**
Katherine L. Kuzminski

**PUBLICATIONS & EDITORIAL DIRECTOR**
Maura McCarthy

**CREATIVE DIRECTOR**
Melody Cook

**DESIGNER**
Rin Rothback

### Cover Art & Production Notes

**COVER ILLUSTRATION**
Rin Rothback

**PRINTER**
CSI Printing & Graphics
Printed on an HP Indigo Digital Press

**Center for a New American Security**
1701 Pennsylvania Ave NW
Suite 700
Washington, DC 20006

CNAS.org

@CNASdc



**Contact Us**

202.457.9400

info@cnas.org


CNAS