Introduction

In the field of AI, not checking the news for a few months is enough to become "out of touch." Occasionally, this breakneck speed of development is driven by revolutionary theories or original ideas. More often, the newest state-of-the-art model doesn't rely on any new conceptual advances at all, rather just a larger neural network and more powerful computing systems than were used in previous attempts.

In 2018, researchers at OpenAI attempted to quantify the rate at which the largest models in AI research were growing in terms of their demands for computing power (often referred to as "compute" for short).[1] By examining the amount of compute required to train some of the most influential AI models over the history of AI research, they identified two trend lines for the rate of compute growth.

They found that prior to 2012, the amount of compute used to build a breakthrough model grew at roughly the same rate as Moore's law, the long-standing observation that the computational power of an individual microchip has tended to double every two years. In 2012, however, the release of the image recognition system AlexNet sparked interest in the use of deep learning methods—the computationally expensive methods that have been behind most of the AI advances of the past decade. Following the release of AlexNet, the compute demands of top models began to climb far faster than the previous trend, doubling not every two years but rather every 3.4 months between 2012 and 2018, as visualized in Figure 1.

The largest models in the early years of deep learning were devoted to image classification, where researchers quickly realized that increasing computing power reliably led to better performances.[2] After image recognition systems began to surpass human-level performance on some tasks, research shifted to new priorities even as the same trend in rising compute needs continued. Around the middle of the 2010s, larger AI models were playing games like Atari or Go using reinforcement learning algorithms.[3] Then, the emergence of a new architecture known as

we assume that compute per dollar is likely to double roughly every four years (solid line), or even every two years (lower bound of shaded region), the compute trendline still quickly becomes unsustainable before the end of the decade.

Figure 2: Extrapolated costs will soon become infeasible

Source: CSET. Note: The blue line represents growing costs assuming compute per dollar doubles every four years, with error shading representing no change in compute costs or a doubling time as fast as every two years. The red line represents expected GDP at a growth of 3 percent per year from 2019 levels with error shading representing growth between 2 and 5 percent.

Without any changes in the price of compute, the cost of a cutting-edge model is expected to cross the U.S. GDP threshold in June of 2026. If the amount of compute that can be performed for a dollar doubles every four years, this point is only pushed back by five months to November of 2026. Even if compute per dollar doubled at the rapid pace of every two years, this point is only delayed until May of 2027, less than a year after it would be reached with no changes in the price of compute. Relaxing the assumption that compute per dollar is a stable value, then, likely buys the original trendline only a few additional months of sustainability.

The Availability of Compute

Rather than fall, price per computation may actually rise as demand outpaces supply. Excess demand is already driving GPU prices to double or triple retail prices.[18] Chip shortages are stalling the automotive industry and delaying products like iPhones, PlayStations, and Xboxes, while creating long wait lists for customers across the board.[19] Whether budgets grow fast enough to continue buying them does not matter if there are not enough chips to continue the trend.

Estimates for the number of existing AI accelerators are imprecise. Once manufactured, most GPUs are used for non-AI applications such as personal computers, gaming, or cryptomining. The large clusters of accelerators needed to set AI compute records are mostly managed in datacenters, but many of those accelerators are better suited for low-power inference than high-performance training.[20] In what follows, our estimates attempt to count the accelerators managed across all cloud datacenters without separating inference chips from training chips, an approach that likely overstates the number of accelerators actually available for AI training.

Overall, 123 million GPUs shipped in the second quarter of 2021, with Nvidia accounting for 15.23 percent of the total, which suggests Nvidia sells approximately 75 million GPU units per year.[21] Thirty-seven percent of Nvidia's revenue came from the datacenter market, and if we likewise assume that approximately 37 percent of its units went to datacenters, this translates to about 28 million Nvidia GPUs going to datacenters annually.[22] Nvidia GPUs are not the only AI accelerators going into datacenters, but they reportedly make up 80 percent of the market.[23] Based on all these figures, we estimate the total number of accelerators reaching datacenters annually to be somewhere in the ballpark of 35 million. This figure is likely a substantial, but it does not need to

be precise.[*] As in the previous section, large errors in estimating the total available supply only result in small changes in the dates at which large-scale models on the compute demand trendline become unattainable.

Following the conventional three-year lifespan for accelerators, we find that by the end of 2025, the compute demand trendline predicts that a single model would require the use of every GPU in every datacenter for a continuous period of three years in order to fully train.[†] Since such a model would need to begin training at the end of 2022 with the full utilization of all accelerators already in cloud datacenters at that time, it would need to use all datacenter accelerators produced since 2019. Just over two years have passed since then, so it is natural to wonder: is the compute demand trend even still alive today, and how much more compute growth is possible if it is not?

---

[*] For this calculation we assumed that 37 percent of Nvidia's revenue coming from the datacenter market implies that 37 percent of its units are shipped to datacenters, but high-end AI processors are more expensive than most consumer GPUs, which means that fewer Nvidia accelerators likely end up in cloud datacenters each year than what we have calculated.

[†] Specifically, December 2025. Even if our estimate for the number of accelerators available in the cloud to train on is off by an order of magnitude, this breaking point would still be reached by December of 2026. The reality may even be more pessimistic than we claim here, because for our calculations we assume that every accelerator in the cloud is capable of operating continuously with a throughput of 163 teraFLOPs per second, a figure that has been obtained experimentally on Nvidia A100 GPUs but that likely overestimates the average performance of all accelerators available in the cloud. See Deepak Narayanan et al., "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM," *arXiv [cs.CL]* (April 2021): arXiv:2104.04473.

Managing Massive Models

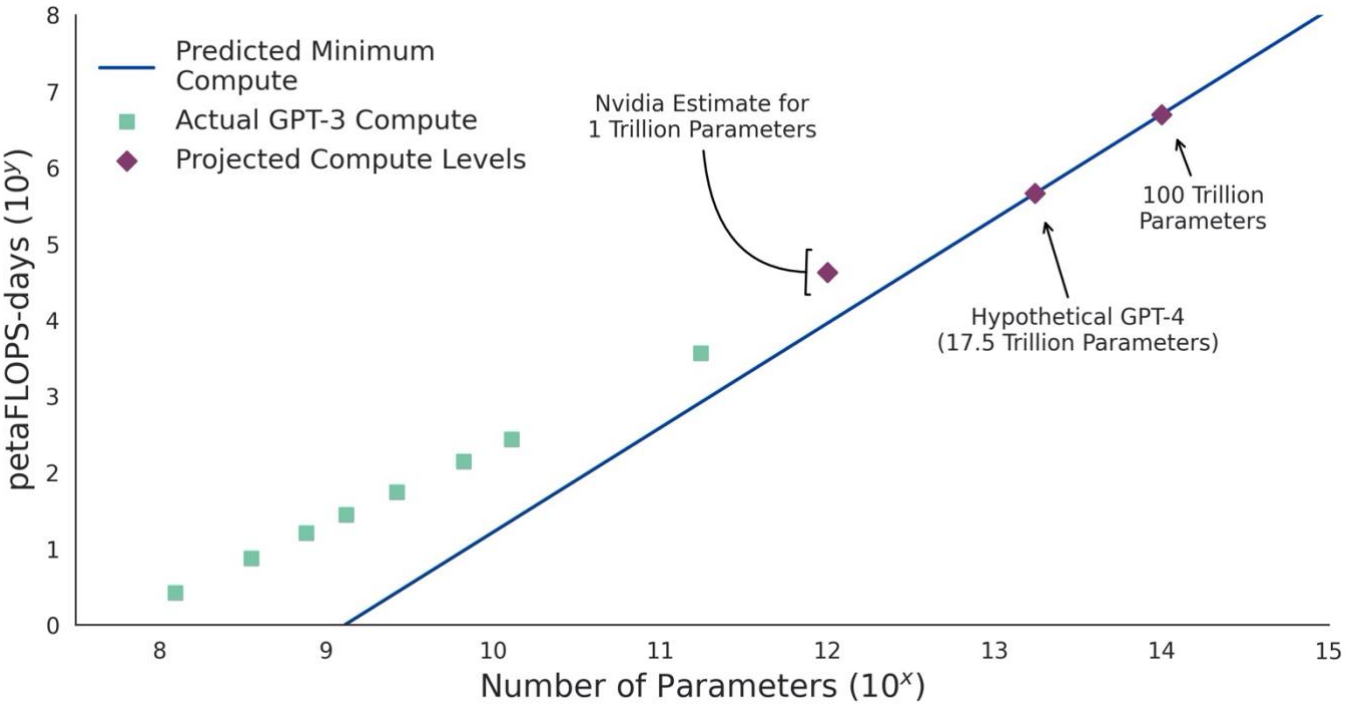
The only major increases in model size since GPT-3's release in 2020 have been a 530 billion parameter model called Megatron-Turing NLG, which was announced in October 2021, and a 280 billion parameter model called Gopher, which was announced in December 2021. The developers of Megatron-Turing NLG reported the size of their compute infrastructure, but they did not report how long the model was trained for, making it impossible to infer a total compute requirement for the model's training process.[24] A useful estimate for how much compute such a model might require to train came five months earlier, when the same developers outlined a similar approach for training models with up to one trillion parameters and included estimates for total training time.[25] They concluded that training a trillion parameter model would take 42,000 petaFLOPS-days, which we conservatively estimate would cost $19.2 million dollars on Google's TPUs training continuously at maximum performance. Had such a model been released in October 2021, it would have fallen a year behind the projected compute demand trend line. This, combined with the fact that GPT-3 likewise fell below the curve, suggests that the compute demand trend may have already started to slow down.

In other research from 2020, OpenAI derived a series of mathematical equations to predict the minimum amount of compute needed to train a variety of models, based on factors like their number of parameters and dataset size.[26] These equations factor in how machine learning training requires the data to pass through the network several times, how compute for each pass grows as the number of parameters grows, and how the data needs to grow as the number of parameters grows.

The blue line in Figure 3 shows OpenAI's equation representing the minimal amount of compute required to effectively train language models of various sizes extrapolated to very large models.[27] The green squares show the amount of compute that was used to train several smaller versions of GPT-3—each of which used larger training datasets than the optimal minimum, and which therefore used more compute than the theoretical minimum. Nvidia's projection for a one trillion parameter model is

shown as a purple diamond along with projections for GPT-4 and a 100 trillion parameter model. For now, assuming that developers can achieve near optimal efficiency, the equation estimates that building GPT-4—which we define as one hundred times bigger than GPT-3 (17.5 trillion parameters)—would take at least 450,000 petaFLOPS-days. That would require 7,600 GPUs running for a year and would cost about $200 million. Training a 100 trillion parameter model would need 83,000 GPUs running for a year and would cost over $2 billion.[28]

Figure 3: Anticipated compute needs for potential AI milestones



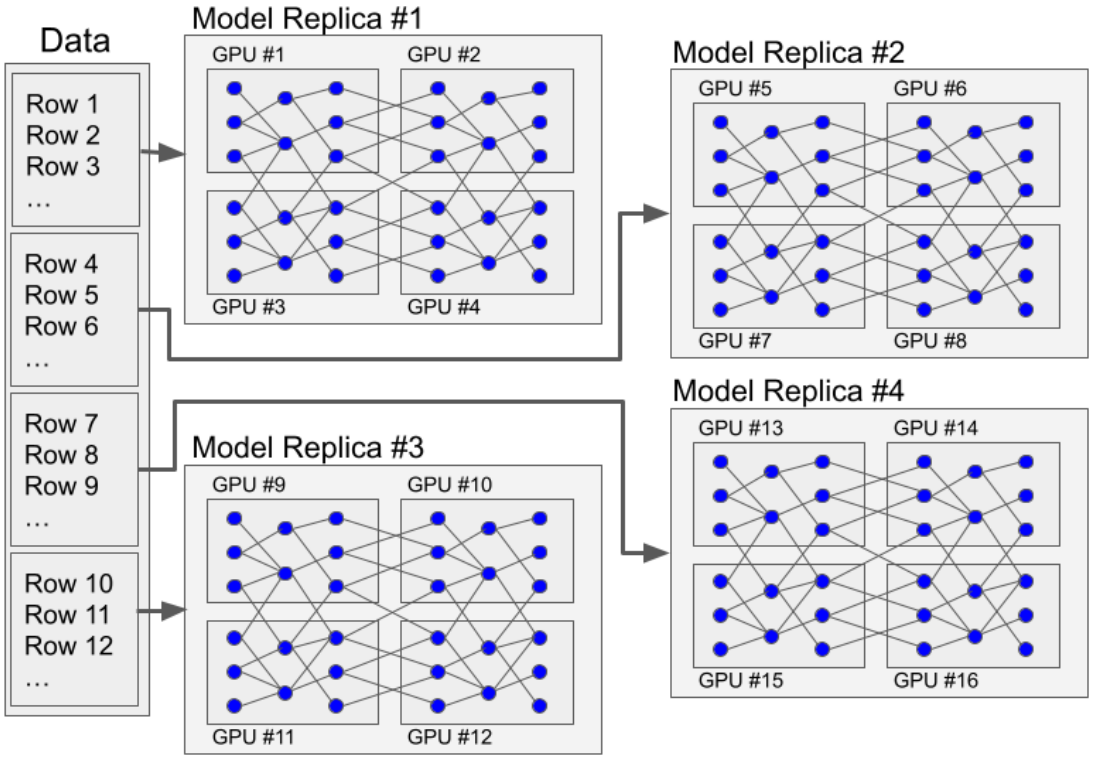
Source: OpenAI, Nvidia, and CSET.

83,000 GPUs represents only 0.2 percent of the 35 million accelerators we estimate go into the cloud every year, and $2 billion is a very high sticker price, though well within the budgetary capacity of a nation-state. But for models over roughly one trillion parameters to be trained at all, researchers will have to overcome an additional series of technical challenges driven by a simple problem: models are already getting too large to manage. The largest AI models no longer fit on a single processor, which means that even inference requires clusters of processors to function. This requires careful orchestration on a technical level to

ensure that multiple processors can run in parallel with one another.

Parallelization for AI is not new. In prior years, AI training often used data parallelization methods, in which many processors worked simultaneously on separate slices of the data, but each processor still stored a full copy of the model. Despite increases in processor memory, this is no longer possible. To train these cutting-edge models, the layers of a deep neural network are held on different processors and even individual layers may be split across processors, as illustrated in Figure 4.

Figure 4: Representation of highly parallelized model training



Source: CSET.

As one example, the 530 billion parameter Megatron-Turing model used 4,480 GPUs in total. Eight different copies of the model ran simultaneously on different slices of the data, but each copy of the model was so big that it was stored across 280 GPUs. The layers of the neural network were split across 35 servers, with each layer itself being spread across eight GPUs.[29] This example shows the complexity of the problem, which only gets more

difficult as the size of the model increases. Moreover, coordinating all of this activity places additional compute requirements on the training process while also requiring significant technical expertise to manage.

Splitting the training process across multiple processors means that the results of computations performed on one processor must be passed to others. At large enough scales, that communication can take significant time, and traffic jams arise. Managing the flows so that traffic does not grind to a halt is arguably the main impediment for continuing to scale up the size of AI models. Some experts question whether it is even possible to significantly increase the parallelization for transformer models like the one used in GPT-3 beyond what has already been accomplished.[30]

Where Will Future Progress Come From?

If the rate of growth in compute demands is already slowing down, then future progress in AI cannot rely on just continuing to scale up model sizes, and will instead have to come from doing more with more modest increases in compute. Unfortunately, although algorithms have been exponentially improving their efficiency, the rate of improvement is not fast enough to make up for a loss in compute growth. The number of computations required to reach AlexNet's level of performance in 2018 was a mere 1/25th the number of computations that were required to reach the same level of performance in 2012.[31] But over the same period, the compute demand trend covered a 300,000 times increase in compute usage. Although algorithms improved dramatically over the last decade, the growth in compute usage has in general been a larger factor in improving the performance of cutting-edge models.[32]

Estimating the rate of improvement in algorithmic efficiency is much harder than estimating the growth in compute usage because it varies across applications, with many major architectures or subfields having only become popular recently.[33] Over short time periods, some domains have improved at nearly the same rate as the compute growth trend.[34] Nonetheless, an end or even partial slowdown to the historical rate of increase in compute usage would require major and continual improvements to algorithmic efficiency in order to compensate. Additionally, efficiency improvements have already been happening throughout the deep learning boom. Making up for a reduced ability to simply scale up compute usage would require not only finding major additional gains in efficiency, but doing so at a rate that is faster than researchers have already been doing. These improvements would need to increase substantially from an already impressively high rate.

Although these results may seem bleak, AI progress will not grind to a halt. The trend in growing compute consumption that drove many of the headlines for the past decade cannot last for much longer, but it will probably slow rather than end abruptly. We should also not discount ingenuity and innovations that could lead

to new breakthroughs in algorithms or techniques, particularly when financial incentives are so large. Indeed, the focus on parallelization that enabled the compute explosion in the first place is largely a byproduct of the looming end of Moore's law and the resulting fears of stagnating compute growth. Some current and future theoretical approaches offer promise for advancing AI research.

Leading algorithms—like the transformer—may be losing training efficiency at the largest sizes, but other architectures are starting to sustain larger models. For instance, Mixture of Experts (MoE) methods allow for more parameters by combining many smaller models together (which may themselves be transformers), each of which are individually less capable than a single large model. This approach permits models that are larger in the aggregate to be trained on less compute, with Google and the Beijing Academy of Artificial Intelligence both releasing trillion-parameter models in the past year trained using MoE methods.[35] MoE approaches offer some advantages but are not as capable in any one area as the largest single models. Both compute and parameter size are critical ingredients for increasing the performance of a model under the current deep learning paradigm, and there are diminishing returns associated with scaling up one without the other.

More importantly, not all progress requires record-breaking levels of compute. AlphaFold is revolutionizing aspects of computational biochemistry and only required a few weeks of training on 16 TPUs—likely costing tens of thousands of dollars rather than the millions that were needed to train GPT-3.[36] Similarly, the current top performing image classifier only needed two days to train on 512 TPUs.[37] In part, these relative efficiencies are due to using algorithms and approaches that have become more efficient over time.[38] But in part, these efficiencies come from simply focusing more on application-centric problems (like protein folding) and tailoring the approach to the task rather than simply throwing more compute at the problem.

Major overhauls of the computing paradigm like quantum computing or neuromorphic chips might one day allow for vast

amounts of plentiful new compute.[39] But these radically different approaches to designing computing chips are still largely theoretical and are unlikely to make an impact before we project that the compute demand trendline will hit fundamental budgetary and supply availability limits. In the meantime, progress will likely involve more incremental improvements to the algorithms and architectures that already exist.

In the nearer term, where the extremes of compute power are needed, that investment can be shared. It may take years, centuries, or millennia of computing time to train a very generalized model, but far less time is needed to fine-tune such a model for newer, more specific applications.[40] This provides an alternate explanation for why GPT-4 has been slow to arrive: rather than simply training a newer, bigger model, OpenAI appears to have shifted its attention to adapting GPT-3 for more carefully scoped, financially viable products such as the code-generating program, Codex.

This shift from a focus on training massive "foundation" models to fine-tuning and deploying them for specific applications is likely to continue.[41] But this type of shift in focus mainly benefits a privileged few if such foundation models are kept as the carefully guarded secrets of a small handful of companies or governments. There may be some security benefits to having these models controlled by a trusted few organizations, which would make it more difficult for malicious actors to misuse models or develop methods of attacking them.[42] On the other hand, if continued AI research requires access to the largest models and those are held by only the wealthiest or most powerful organizations, then AI research will become increasingly difficult for the larger part of the AI community.

Conclusion and Policy Recommendations

For nearly a decade, buying and using more compute each year has been a primary factor driving AI research beyond what was previously thought possible. This trend is likely to break soon. Although experts may disagree about which limitation is most critical, continued progress in AI will soon require addressing major structural challenges such as exploding costs, chip shortages, and parallelization bottlenecks. Future progress will likely rest far more on a shift towards efficiency in both algorithms and hardware rather than massive increases in compute usage. In addition, we anticipate that the future of AI research will increasingly rely on tailoring algorithms, hardware, and approaches to sub-disciplines and applications.

This is not to say that progress towards increasingly powerful and generalizable AI is dead; only that it will require a partial re-orientation away from the dominant strategy of the past decade—more compute—towards other approaches. If correct, this finding has a number of implications for policymakers interested in promoting AI progress. We discuss a few of these implications below:

(1) **Shift focus towards talent development, both by increasing investment in AI education at home and by actively competing to attract highly skilled immigrants from abroad.** Improving algorithmic efficiency and overcoming parallelization bottlenecks in training are difficult problems that require significantly more human expertise than simply purchasing more compute. This suggests that the path towards continued progress in the future rests far more on developing, attracting, and retaining talent than merely outspending competitors. Correspondingly, policymakers who want to encourage AI progress at home should invest significant resources in (a) bolstering AI and computer science education, (b) increasing the number of H1-B visas available for AI researchers specifically, and (c) striving to make the United States a more attractive destination for immigrants generally. CSET already has publications addressing each of these topics.[43]

(2) **Support AI researchers with technical training, not just compute resources.** The National Artificial Intelligence Research Resource (NAIRR) Task Force is currently exploring the types of support that it can provide to bolster AI research in the United States, especially in the broad categories of "computational resources, high-quality data, educational tools, and user support."[44] Compute remains an extremely important factor in AI progress, and the NAIRR should take steps where possible to expand the access of researchers to compute resources—especially academics, students, and those without access to multi-million-dollar budgets.

It is unlikely that the NAIRR can provide sufficient compute to researchers to keep the compute demand trendline alive, or even to compete with the quantities of compute already used by major research centers. Nonetheless, impactful results and educational experience can come from even moderately sized models. Significant attention should be paid to developing educational tools that can help researchers build the skills necessary to innovate with more efficient algorithms and better-scaling parallelization methods. Programs that promote interdisciplinary work between machine learning and other areas of computer science such as distributed systems and programming languages may be especially fruitful for generating broad efficiency gains.

(3) **Promote openness and access to large-scale models throughout the research community, especially for researchers who cannot train their own.** The future of AI research may come to focus heavily on the intermittent release of massive, compute-intensive "foundation models" that then become the basis for extensive follow-on research and development. If this general depiction is right, then the United States has an interest in ensuring that these foundation models are not monopolized by only a small handful of actors. There are likely to be other researchers or entrepreneurs who could contribute meaningfully to our understanding or application of these models even though they may lack the compute resources to build similarly sized models themselves.

Policymakers, where appropriate, should seek to encourage the owners of large foundation models to permit appropriately vetted researchers access to these models. In many cases, however, this must be balanced against the need to promote the security of the models themselves, especially those with potentially dangerous uses.[45] Regrettably, there are unlikely to be hard or fast rules that can govern when models should be made as public as possible or when they should be deliberately made difficult to access. At this stage, we limit ourselves to noting that efforts should be made to ensure that AI remains a field where researchers of many backgrounds can usefully contribute and where access to a few key models does not rest entirely in the hands of a coterie of powerful institutions.

Authors

Andrew Lohn is a senior fellow with the CyberAI Project at CSET, where Micah Musser is a research analyst.

Acknowledgments

For feedback and assistance, we would like to thank John Bansemer, Girish Sastry, Deepak Narayanan, Jared Kaplan, and Neil Thompson, all of whose experience and analytical comments were tremendously helpful in developing this project. Melissa Deng and Alex Friedland provided editorial support.

Document Identifier: doi: 10.51593/2021CA009

Endnotes

[1] Dario Amodei et al., "AI and Compute," *OpenAI*, May 16, 2018, https://openai.com/blog/ai-and-compute/.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM* 60, no. 6 (June 2017): 84-90.

[3] Volodymyr Mnih et al., "Playing Atari with Deep Reinforcement Learning," arXiv preprint arXiv:1312.5602 (2013); David Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," arXiv preprint arXiv:1712.01815 (2017).

[4] Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805 (2018); Yinhan Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692 (2019); Colin Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv preprint arXiv:1910.10683 (2019); Alec Radford et al., "Language Models are Unsupervised Multitask Learners," *Papers With Code*, 2019, https://paperswithcode.com/paper/language-models-are-unsupervised-multitask; Tom Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165 (2020).

[5] Brown, "Language Models."

[6] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258 (2021).

[7] For this paper, we mean to say that this trend is unsustainable in the sense that the trend itself cannot continue. But it is worth mentioning that spiraling compute demands are also unsustainable in an environmental sense. Training GPT-3 released an estimated 552 tons of CO_2 equivalent into the atmosphere—the equivalent of 460 round-trip flights between San Francisco and New York. David Patterson et al., "Carbon Emissions and Large Neural Network Training," arXiv [cs.LG] (April 2021): arXiv:2104.10350. It is easy to overstate the importance of this value: even in 2020, roughly six times this many people flew between San Francisco and New York every day. "SFO Fact Sheet," FlySFO, accessed December 4, 2021, https://www.flysfo.com/sfo-fact-sheet. This energy consumption is also miniscule compared to other emerging technologies that require enormous amounts of computing, most notably cryptocurrencies. In 2018, Bitcoin alone was estimated to have generated 100,000 times that volume of CO_2 emissions, and by 2021 the energy requirements of Bitcoin had nearly doubled relative to 2018. That is more electricity than the entire nation of

[22] Jim Chien, "Nvidia, AMD see rising sales from server sector," *DigiTimes Asia*, June 30, 2020, https://www.digitimes.com/news/a20200630PD213.html.

[23] "NVIDIA maintains dominant position in 2020 market for AI processors for cloud and data center," *Omdia*, August 4, 2021, https://omdia.tech.informa.com/pr/2021-aug/nvidia-maintains-dominant-position-in-2020-market-for-ai-processors-for-cloud-and-data-center.

[24] Paresh Kharya and Ali Alvi, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model," *NVIDIA Developer Blog*, October 11, 2021, https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.

[25] Deepak Narayanan et al., "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM," arXiv preprint arXiv:2104.04473 (2021).

[26] Jared Kaplan et al., "Scaling Law for Neural Language Models," arXiv preprint arXiv:2001.08361 (2020); Tom Henighan et al., "Scaling Laws for Autoregressive Generative Modeling," arXiv preprint arXiv:2010.14701 (2020).

[27] It is important to note that this scaling law applies specifically to the single-transformer architecture. There are other approaches—for instance, mixture of expert models, discussed a bit later—for which more parameters can be trained using less compute but sacrificing aspects of performance. We analyze the transformer architecture, as it is currently the favored approach for language models and is versatile enough to perform well across a number of other domains. Eventually, the transformer will likely be supplanted by other models, but this discussion helps to ground the amount of compute that would be required under the current paradigm to reach models of various sizes.

[28] Lower prices per GPU-hour are available for long term commitments, but long training times are less useful than splitting the model or increasing the batch size, so it is not clear that long term commitments are beneficial. See Jared Kaplan et al., "Scaling Law for Neural Language Models," arXiv preprint arXiv:2001.08361 (2020).

[29] Kharya and Alvi, "Using DeepSpeed."

[30] Bommasani et al., "Opportunities and Risks."

[31] Danny Hernandez and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," arXiv preprint arXiv:2005.04305 (2020).

[32] Note that efficiency improvements can also be due to improvements other than those in the algorithms used to train models, such as improvements in

methods for data collection, curation, or usage. See Sebastian Borgeaud et al., "Improving language models by retrieving from trillions of tokens," arXiv preprint arXiv:2112.04426 (2021) for an example of research in this area.

[33] There have also been some analyses of algorithmic efficiency improvements in fields beyond AI. See Yash Sherry and Neil C. Thompson, "How Fast do Algorithms Improve?," *Proceedings of the IEEE* 109, no. 11 (November 2021): 1768-1777.

[34] Hernandez and Brown, "Measuring Algorithmic Efficiency."

[35] Noam Shazeer et al., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," arXiv preprint arXiv:1701.06538 (2017); William Fedus, Barret Zoph, and Noam Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," arXiv preprint arXiv:2101.03961 (2021); Alberto Romero, "GPT-3 Scared You? Meet Wu Dao 2.0: A Monster of 1.75 Trillion Parameters," *Towards Data Science,* June 5, 2021, https://towardsdatascience.com/gpt-3-scared-you-meet-wu-dao-2-0-a-monster-of-1-75-trillion-parameters-832cd83db484.

[36] John Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature* 596 (August 2021): 583-589.

[37] Hieu Pham et al., "Meta Pseudo Labels," arXiv preprint arXiv:2003.10580 (2020).

[38] Hernandez and Brown, "Measuring Algorithmic Efficiency."

[39] "The cost of training machines is becoming a problem," *The Economist*, June 13, 2020, https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem.

[40] Danny Hernandez et al., "Scaling Laws for Transfer," arXiv preprint arXiv:2102.01293 (2021).

[41] Bommasani et al., "Opportunities and Risks."

[42] Andrew J. Lohn, "Poison in the Well" (Center for Security and Emerging Technology, June 2021), https://doi.org/10.51593/2020CA013; Benjamin Buchanan et al., "Truth, Lies, and Automation: How Language Models Could Change Disinformation" (Center for Security and Emerging Technology, May 2021), https://doi.org/10.51593/2021CA003.

[43] Diana Gehlhaus et al., "U.S. AI Workforce: Policy Recommendations" (Center for Security and Emerging Technology, October 2021), https://doi.org/10.51593/20200087; Dahlia Peterson, Kayla Goode, and Diana Gehlhaus, "AI Education in China and the United States" (Center for Security

and Emerging Technology, September 2021), https://doi.org/10.51593/20210005; Zachary Arnold et al., "Immigration Policy and the U.S. AI Sector: A Preliminary Assessment" (Center for Security and Emerging Technology, September 2019), https://doi.org/10.51593/20190009; Tina Huang and Zachary Arnold, "Immigration Policy and the Global Competition for AI Talent" (Center for Security and Emerging Technology, June 2020), https://doi.org/10.51593/20190024; Tina Huang, Zachary Arnold, and Remco Zwetsloot, "Most of America's 'Most Promising' AI Startups Have Immigrant Founders" (Center for Security and Emerging Technology, October 2020), https://doi.org/10.51593/20200065; Remco Zwetsloot et al., "Keeping Top AI Talent in the United States" (Center for Security and Emerging Technology, December 2019), https://doi.org/10.51593/20190007; Remco Zwetsloot, Roxanne Heston, and Zachary Arnold, "Strengthening the U.S. AI Workforce: A Policy and Research Agenda" (Center for Security and Emerging Technology, September 2019), https://doi.org/10.51593/20190003.

[44] "The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force," *The White House*, June 10, 2021, https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.

[45] See, e.g., Buchanan et al., "Truth, Lies, and Automation."