Research Report

KONSTANTIN F. PILZ, YUSUF MAHMOOD, LENNART HEIM

AI's Power Requirements Under Exponential Growth

Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness

For more information on this publication, visit **www.rand.org/t/RRA3572-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

Limited Print and Electronic Distribution Rights

About This Report

An exponential increase in computational resources (compute) used for artificial intelligence (AI) training and deployment has recently enabled rapid advances in AI models' capabilities and their widespread use. The resulting unprecedented demand for AI data centers is already posing challenges for U.S. data center construction, primarily because it is difficult to find adequate power grid capacity. To allow policy researchers, government agencies, and elected officials to anticipate the impacts of future growth in demand, we ask the following questions:

- How much power would the United States have to provide if it wanted to host a majority of upcoming AI chip production?
- How much power will data centers running the largest training runs require at the current rate of growth?
- What would the consequences on U.S. competitiveness be if the United States cannot meet this demand for AI data centers?

In this report, we provide two extrapolations of recent AI trends to assess future power needs, summarize current bottlenecks for rapid data center construction, and discuss what a failure to resolve them could mean for U.S. competitiveness.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

This research was, at its inception, independently initiated and conducted within the Technology and Security Policy Center using income from operations and gifts from philanthropic supporters, which have been made or recommended by DALHAP Investments Ltd., Effektiv Spenden, Ergo Impact, Founders Pledge, Fredrik Österberg, Good Ventures, Jaan Tallinn, Longview, Open Philanthropy, and Waking Up Foundation. This research was published through funding provided by DALHAP Investments Ltd., as recommended by Ergo Impact and Fathom. A complete list of donors and funders is available at www.rand.org/TASP. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

We would like to thank the leadership of the RAND Technology and Security Policy Center, Jeff Alstott and Casey Dugan, for their guidance on this publication, Alison Hottes for managing the quality assurance process, and our reviewers Joel Predd, Aimee Curtright, and Ben Cottier for their thoughtful feedback.

Summary

Larger training runs and widespread deployment of future artificial intelligence (AI) systems may demand a rapid scale-up of computational resources (compute) that require unprecedented amounts of power. We find that globally, AI data centers could need ten gigawatts (GW) of additional power capacity in 2025 alone, which is more than the total power capacity of the state of Utah. If exponential growth in chip supply continues, AI data centers will need 68 GW in total by 2027—almost a doubling of global data center power requirements from 2022 and close to California's 2022 total power capacity of 86 GW.

Given recent training compute growth, data centers hosting large training runs pose a particular challenge. Training could demand up to 1 GW in a single location by 2028 and 8 GW—equivalent to eight nuclear reactors—by 2030, assuming that current training compute scaling trends persist.

The United States currently leads the world in data centers and AI compute, but unprecedented demand leaves the industry struggling to find the power capacity needed for rapidly building new data centers. Failure to address current bottlenecks may compel U.S. companies to relocate AI infrastructure abroad, potentially compromising the U.S. competitive advantage in compute and AI and increasing the risk of intellectual property theft.

More research is needed to assess bottlenecks for U.S. data center build-out and identify solutions, which may include simplifying permitting for power generation, transmission infrastructure, and data center construction.

Contents

Figures and Tables

Figures

Tables

Chapter 1

Projecting Power Requirements for AI Data Centers

Artificial intelligence's (AI's) demand for computational resources has grown exponentially,[1] driven by increasing training requirements and a rapidly expanding user base. Both developing and deploying frontier models take tens of thousands—and soon hundreds of thousands—of AI chips (Sevilla et al., 2022). Hence, training and deployment require massive amounts of power (Burkacky et al., 2024; Scharre, 2024). For instance, xAI's Colossus supercomputer in Memphis, Tennessee, contains 100,000 AI chips and requires 150 megawatts (MW) of power—the generation capacity of around 55 modern wind turbines and the equivalent of about 53,000 U.S. households (Trueman, 2024; U.S. Geological Survey, 2022).[2]

Exponential growth in both (1) the total number of AI chips deployed in the United States and (2) the compute used for the largest training runs over the past few years are already leaving the data center industry struggling to accommodate record demand (U.S. Department of Energy, 2024). In this report, we assess the potential scale of AI data center power requirements under continued exponential growth by extrapolating two key trends: (1) growth in total AI chip production and (2) growth in compute used for training notable AI models. We then summarize already existing challenges for data center construction in the United States and discuss the potential geopolitical implications if U.S. AI data center construction cannot meet demand.

Both our projections assume that recent exponential growth in compute demand will continue. Yet, technical limits, chip production bottlenecks, or geopolitical events could disrupt these trends (Appendix B).

Total Power Requirements for AI Infrastructure in the United States

Continued exponential growth in demand would far outpace previous data center expansion. AI data center power demand grew tenfold over the last three years—from 0.4 gigawatts (GW) in 2020 to 4.3 GW in 2023 (Patel, Nishball, and Ontiveros, 2024).[3] In 2025, total AI data center demand will likely reach about 21 GW of total power capacity, more than a fourfold increase from

[1] In the rest of the report, we refer to *computational resources* as *compute*.

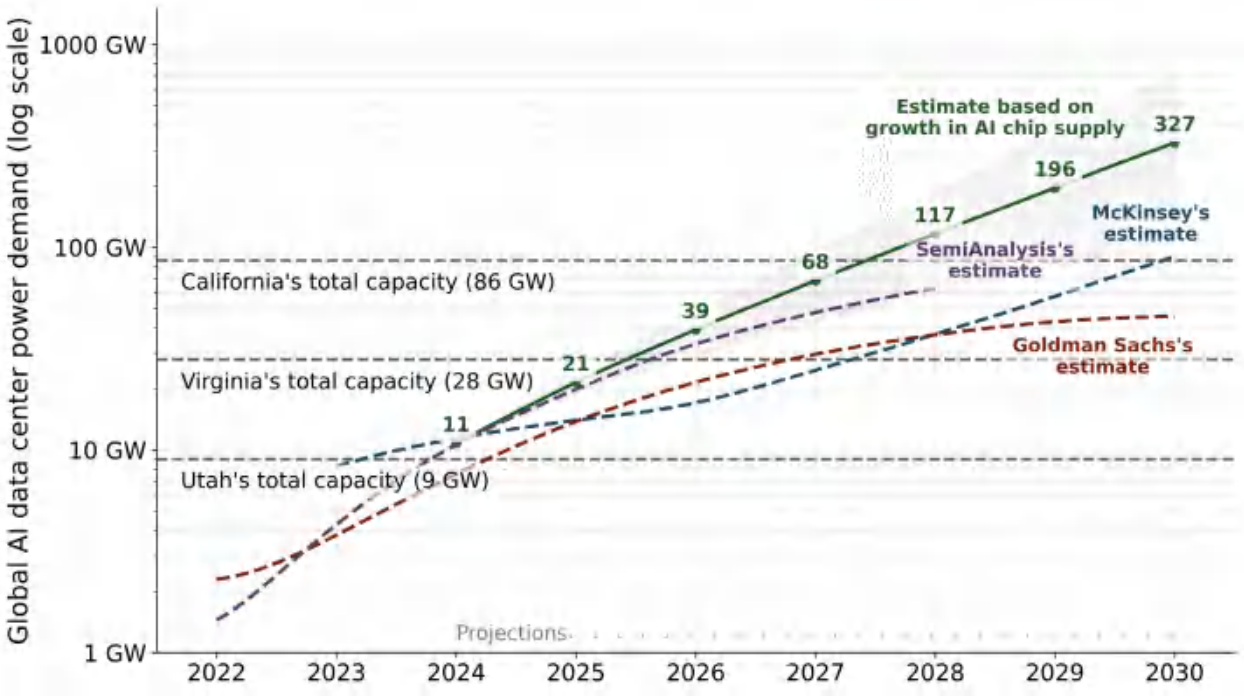
[2] The average wind turbine generates about 2.75 MW of power. It takes 150 MW / 2.75 MW = 55 wind turbines to power the Memphis data center, which could supply about 53,000 households.

[3] For 2020, 0.3 GW × 1.25 PUE = 0.375 GW; for 2023, 3.3 GW × 1.25 PUE = 4.125 GW.

2023 and twice the total power capacity of the state of Utah (Patel, Nishball, and Ontiveros, 2024; U.S. Energy Information Administration [EIA], 2024).

Sevilla et al. (2024) estimate that AI chip production could increase between 1.3 and 2 times annually until 2030 if demand continues to grow at the current rate. We find that this expansion would require about 68 GW of AI data center capacity globally by 2027 and 327 GW by 2030, even when accounting for increases in data center power usage effectiveness (PUE) (Figure 1.1).

Figure 1.1. Estimates of Data Center Power Capacity Required to Host All AI Chips, 2024–2030



SOURCES: Authors' analysis of data from Patel, Nishball, and Ontiveros, 2024; Srivathsan et al., 2024; Goldman Sachs, 2024; and EIA, 2023.

NOTE: We extrapolated power demand of AI data centers by assuming that AI chip supply could grow between 1.3 and 2 times per year, given continued rapidly growing demand (Sevilla et al., 2024). AI data center power demand in 2024 was estimated to be about 11 GW (Patel, Nishball, and Ontiveros, 2024). The green line displays the median estimate for AI data centers' future power demand at a 1.7-times annual growth rate in chip supply; the grey uncertainty range indicates demand growth between 1.3 and 2 times. We display three external estimates: (1) "SemiAnalysis," by Patel, Nishball, and Ontiveros (2024); (2) "McKinsey," by Srivathsan et al. (2024); and (3) Goldman Sachs (2024). We further display reference lines showing the total power capacity of select U.S. states (as of 2022). See Appendix B for full sources and methodology. Note that this figure includes only AI-specific data center demand. Traditional data centers grow by about 5 GW per year, which would be added to the demand displayed (Patel, Nishball, and Ontiveros, 2024). For illustration: 68 GW power capacity could power about 38 million NVIDIA H100s—the most-used AI chips in 2024. 327 GW could power about 184 million H100s.[a]

[a] 68 GW / 1.25 PUE / 1,419 watts (W)/H100 = 38 million H100s; 327 GW / 1.25 PUE/ 1,419 W/H100 = 184 million H100s. See Appendix A for sources.

To put this number into perspective, in 2022, global data center power capacity was about 88 GW, indicating that AI alone will drive data center growth of 180 percent by 2027 and 460 percent by 2030 (International Energy Agency [IEA], 2024).[4] For another comparison, California (the most populated U.S. state) has a total power capacity of 86 GW, and the total power capacity of the United States was 1,105 GW in 2022 (EIA, 2023).

Other estimates of AI data center power needs expect slower demand growth globally: Patel, Nishball, and Ontiveros (2024) expect 62 GW of total demand by 2027; Goldman Sachs (2024) predicts about 24 GW of total demand by 2030; and Srivathsan et al. (2024) expect about 90 GW by 2030 (see Figure 1.1). Although details on the authors' methods were not provided for any of these estimates, each likely assumes that AI is experiencing only a brief period of exponential growth. We discuss this divergence and other potential limitations in Appendix B.

It is not certain that the United States is adding power quickly enough to accommodate AI data center demand. If the United States wanted to retain a majority (say, 75 percent) of AI compute within its borders, it would have to make about 51 GW available to AI data centers by 2027. This is in addition to the growing energy demand of non-AI data centers and other consumers in the United States, such as electric vehicles, residential homes, and manufacturing (Goldman Sachs, 2024).

It is difficult to assess whether the United States is building power generation capacity quickly enough. This is because even though the United States is projected to add almost 160 GW of theoretical power capacity between 2024 and 2028, most of that capacity will come from wind and solar, both of which are available only a fraction of the time, greatly reducing the power delivered (EIA, undated). Assessing the potential gap between power supply and demand growth in the United States would require modeling supply in a more detailed way, as well as additional sources of demand besides AI data centers. A comprehensive analysis is beyond the scope of this report, but current evidence shows that U.S. companies already face increasing difficulty in finding sufficient grid capacity to build data centers that meet AI demand, indicating a challenge that will likely worsen with increased training and deployment of AI (Moss, 2024a).

Making renewable energy sources suitable for AI data centers remains a challenge. The vast majority of generation capacity added to the grid in 2024 comes from wind and solar energy (EIA, 2024). However, because of daily and seasonal variations, power generation from these sources is less consistent than traditional sources (such as nuclear, gas, or coal) unless they are combined with energy storage (Tozzi, 2023). Current data center design requires power being available more than 99 percent of the time (Uptime Institute, undated). Although some developers have indicated that AI training workloads could accommodate reduced power reliability, this is not yet done in practice (U.S. Department of Energy, 2024). Given these constraints and additional permitting challenges associated with using renewable energy, some industry experts emphasize gas power plants, in addition to renewables, as an important source of additional capacity to ensure reliability (U.S. Department of Energy, 2024).

To use another carbon-neutral source of power, major AI compute providers have signed deals to bring retired nuclear power plants, such as Three Mile Island, back online or are planning to build

[4] Assuming an average power utilization of 60 percent for data centers, 460 terawatt-hours (TWh) of power consumption in a year corresponds to a power capacity of about 88 GW (460 TWh / 8,760 h / 0.6 = 87.5 GW).

small modular reactors (Wong, 2023; Moss, 2024b). However, it will likely take years until additional nuclear capacity from these sources will come online (Fist and Datta, 2024).

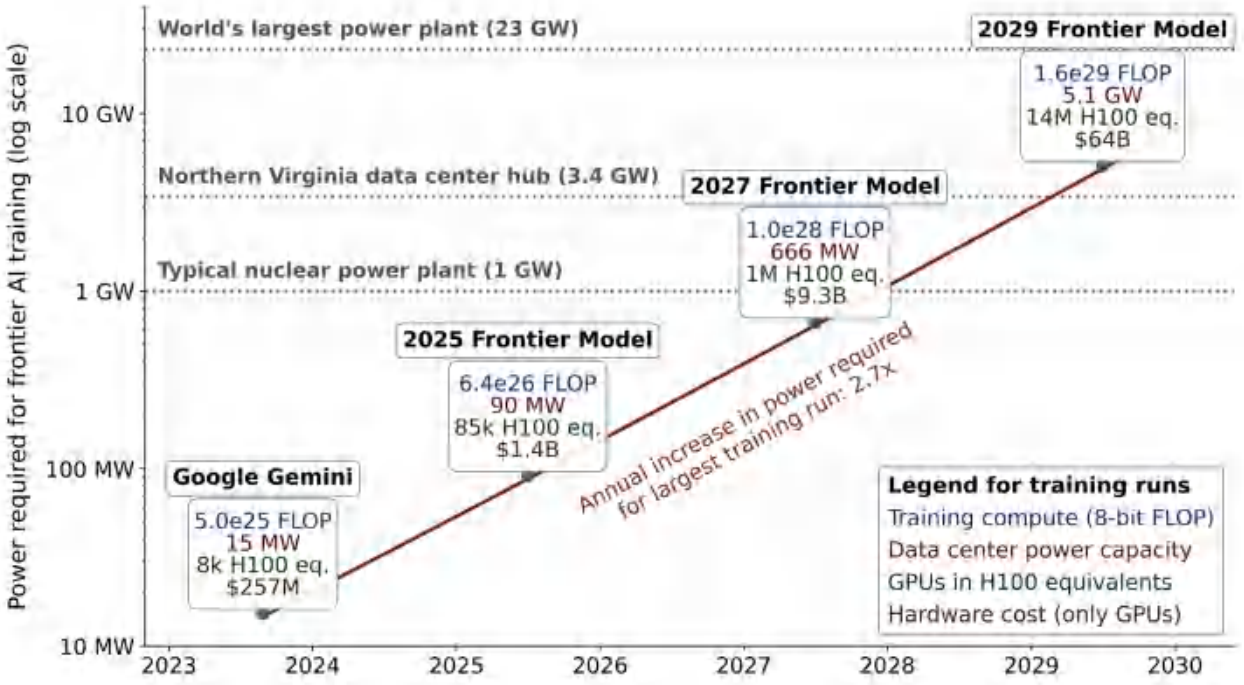
Given the reliability issues for wind and solar and long lead times for nuclear and geothermal energy, there will likely be a trade-off between decarbonization goals and AI data center build-out in the United States (Fist and Datta, 2024). AI data centers' demand for reliable supply may require delaying the retirement of coal and gas power plants and potentially even establishing additional gas power plants.

Power Requirements for AI Training

AI training presents a unique challenge because it requires large amounts of power capacity available at a single location. The compute used to train the most advanced models is rapidly increasing by about 4 to 5 times every year (Sevilla and Roldán, 2024a; Sevilla et al, 2022). Meanwhile, the energy efficiency of AI chips has grown much more slowly (only about 1.3 times per year), and data center PUE has only moderately improved recently (Hobbhahn, Heim, and Aydos, 2023; Taylor, 2024). Extrapolating these trends highlights that, in 2027, a data center hosting AI training may require as much power as a typical nuclear power plant can generate (Figure 1.2). In 2029, the power required for a single AI training cluster may even surpass all data centers in Northern Virginia—the world's largest data center hub—combined. While decentralized training across multiple data centers could reduce the burden on a single location, in practice, most AI developers still build large, centralized clusters.[5] Furthermore, even if a 2029 training run could be split among, say, four clusters, each of these clusters would still require more than 1 GW of power capacity.

[5] Recent examples, such as Google's approach with Gemini training (Anil et al., 2024), show attempts at multi–data center distribution. However, details are limited (such as the number of data centers), and Google may have used only a small number of closely located facilities. Meanwhile, other reported efforts for large clusters rely on a single centralized location (Kahn, 2024; Trueman, 2024).

Figure 1.2. Extrapolation of the Power Required for the Data Center Hosting the Largest AI Training Run



NOTE: We model power requirements for future frontier model training runs and estimate they will increase by 2.7 times per year. We assume the floating point operations (FLOP) used for training the largest AI models continue to grow fourfold every year. We model graphics processing unit (GPU) power efficiency improvements of 1.3 times per year and data center PUE to improve from 1.25 to 1.05 by 2030. We further assume that training run duration increases from 100 days for Google's Gemini to 200 days by 2030. The y-axis shows the data center power capacity required for hosting the supercomputer running the training of future frontier models. We include four data points: a hypothetical cluster used for training Google's Gemini model and three clusters for hypothetical future models. (Google's Gemini was first released in December 2023, but given that it trained for 100 days, the cluster it trained on must have existed at least 100 days before release, so we display the data point 100 days before release [Epoch AI, 2024].) For each model, we display the training compute in 8-bit FLOP, the data center power capacity required, the number of H100 equivalents required to achieve the same performance, and the hardware price, assuming that an H100 cost $30,000 in 2023 and the price per FLOP improved by 2 times every 2.1 years. We display three reference points for context: a typical nuclear power plant (Office of Nuclear Energy, 2022); the combined capacity of all data centers in the Northern Virginia region, the world's largest data center hub (Clabaugh, 2023); and the capacity of China's Three Gorges Dam, the world's largest power plant (U.S. Geological Survey, undated). See Appendix B for methods and data sources. eq. = equivalents; k = thousand.

Limitations of Extrapolations

Both projections assume that current exponential trends will continue for the next six years. This is because compute scaling has been consistent for more than a decade and because in 2024 several hyperscalers announced ambitious plans to continue growing their compute capacity (Vanian, 2024; Metz and Mickle, 2024; Herrera, 2024, Sevilla et al., 2024). Yet, diminishing returns to AI scaling, geopolitical events, or a lower growth in chip supply could all reduce the investment available for infrastructure build-out.

We further assume only moderate energy efficiency improvements, consistent with recent trends. Breakthroughs in energy efficiency could reduce the power required for future AI chips. We discuss these and other limitations in Appendix B.

Chapter 2

Challenges for Rapid AI Data Center Construction in the United States and Their Implications

To assess current challenges for rapidly expanding power generation and data center construction, we summarize common causes for delays cited in recent reports and analyses. We find that permitting challenges for power generation, distribution, and data center construction are common causes that delay data center projects.

1. Insufficient power generation: A lack of power generation addition is increasing wait times for grid connections. For instance, in Virginia, the state with the largest share of data centers, connection requests now take between four and seven years (Saul, 2024). This is in part because of permitting challenges for various power generation projects (Bauer et al., 2024).
2. Inadequate transmission infrastructure: Even when power is available, regions often lack transmission lines to deliver the power to sites suitable for data center construction. Projects to expand transmission infrastructure are difficult to coordinate and can take years (Potter, 2024).
3. Data center permitting issues: Large data center projects face a variety of permitting challenges on the local, state, and federal level, limiting suitable sites and sometimes causing project delays and cancelations (Kurtz, 2023; Spivack, 2023).
4. Supply chain delays: Data centers need a wide variety of inputs. Some, such as emergency power generators, now have waiting times of more than one year. Supply chain issues greatly delay data center construction (Dotan and Fitch, 2024).
5. Environmental commitments and regulations: Data centers are subject to government regulations that limit their use of certain forms of energy, such as natural gas (Department of Ecology, State of Washington, undated). Given the constraints that these commitments and regulations place on data center construction, compute providers' ability to procure sufficient power can trade off against environmental considerations (Coleman, 2023).

In combination, these factors cause an increasing gap between data center supply and demand, as evidenced by historically low vacancy rates for colocation data centers that host hardware for business customers. As a consequence, almost 80 percent of upcoming data center space is already pre-leased ahead of completion (CBRE, 2024).

A shortage of data center infrastructure could lead to less compute being available to companies and researchers, potentially slowing down AI development and deployment. It could further provide an advantage to large compute providers that have the capital to invest in future infrastructure build-

out years ahead and thus outcompete smaller market participants. Finally, a lack of data center infrastructure in the United States could cause data center construction to shift to other countries. In the next section, we discuss the potential geopolitical implication of such a shift.

Geopolitical Implications

Inadequate power may reduce the U.S. lead in AI compute. The United States leads the world in number of data centers and market share of compute providers (Sastry et al, 2024). Although no direct estimates exist, this indicates that the United States likely hosts a significant majority of all AI chips. However, the current challenges in power availability for rapid data center construction could reduce this advantage. An increasing number of U.S. companies are considering expanding their AI infrastructure to countries that offer more power availability, faster permitting, and additional financial incentives (Shoaib, 2024). For instance, Microsoft recently invested $1.5 billion in the United Arab Emirates' G42 for AI and AI infrastructure projects (Microsoft, 2024a; Microsoft, 2024b; "Abu Dhabi's US$302 Billion Fund Deepens AI and Private Credit Push," 2024; World Nuclear Association, 2024). This trend of companies looking abroad for AI infrastructure development raises several concerns for the United States' position in the global AI landscape. The following sections outline why this shift is problematic and its potential implications for U.S. leadership in AI technology and innovation.

Compute is a primary enabler of AI progress; leading in compute enables leading in AI. Recent breakthroughs in AI have largely relied on a massive increase in compute used for training AI models (Sevilla and Roldán, 2024a). In other words, the success of a country's AI industry relies on access to specialized compute and the infrastructure needed to host it (Khan and Mann, 2020). Furthermore, the more compute a country has access to, the wider it can deploy its AI models. This may allow the country to derive economic and military advantages by deploying relevant AI models on a larger scale than competitors (Pavel et al., 2023). Some AI experts further predict that future AI systems will possess capabilities highly relevant to national security, such as automated warfare and biological weapons (Bengio et al., 2024). Controlling a large share of AI infrastructure could allow the United States to mitigate such risks and help ensure that AI use is aligned with democratic values.

Supplying AI chips to additional countries could complicate enforcement of semiconductor export controls. Controlling AI chip supply allows the United States to impose export controls on China and other international competitors to slow their advances in AI (Allen, 2022). However, the more chips the United State exports beyond close allies, the more opportunities it gives China to smuggle them (Fist and Grunewald, 2023). Retaining chips within the United States and providing cloud access could allow the United States to provide AI compute to other countries without increasing the risk of smuggling (Fist and Scharre, 2023).

As AI models get more capable, securing AI compute becomes increasingly challenging, particularly abroad. The infrastructure hosting advanced AI models will likely become a target of increasingly sophisticated cyberattacks (Nevo et al., 2024). This risk is significantly amplified when compute resources are located outside U.S. borders, where oversight and control are more limited.

Potential Options to Limit National Security Risks of Compute Exports

Given the geopolitical importance of compute, it may be in the U.S. interest to retain a large share of it within U.S. borders. However, if a lack of power availability and permitting issues mean that the United States can host only a limited number of AI chips, the following measures can mitigate some of the downsides of compute exports.

The United States could track AI chip exports to assess the global distribution of compute. While the United States leads in data center capacity and AI chip supply, the exact amount of AI compute within its control remains unclear (Pilz and Heim, 2023; Sastry et al., 2024). The United States currently tracks exports only to a limited number of jurisdictions (Dohmen and Feldgoise, 2023). To better assess global compute distribution, the United States could implement a tracking system for national AI computing power. Such a system could collect sales data from AI chip suppliers and inventory information from major data center providers. Data about the global distribution of AI chips would allow the United States to assess it relative position compared with other countries and help address national security risks from misuse of AI chips by foreign governments.

Projects run by U.S. companies and strict security and reporting requirements could reduce security risks. One option to maintain U.S. control over AI chips exported to countries that are not U.S. allies or partners could be to favor projects led by U.S. companies to limit foreign government influence (Chou, 2024). Additionally, the U.S. government could set conditions for export licenses. For instance, it could require companies purchasing AI chips to demonstrate compliance with rigorous physical and cybersecurity standards and to commit to reporting security incidents—such as cyberattacks, sabotage, or misuse of infrastructure and models—to relevant U.S. security agencies. These requirements could help protect advanced AI models and other intellectual property of U.S. companies running on infrastructure abroad from theft. Finally, to prevent misuse of compute resources, export licenses could further require AI chip owners to adopt know-your-customer practices and regularly report data on compute users and use types to the U.S. government (Egan and Heim, 2023; Nevo et al., 2024).

Although these strategies can reduce some of the risks compute exports introduce, such exports still reduce the U.S. strategic advantage in compute overall and thus limit U.S. diplomatic, economic, and innovation power, as outlined previously. In the next section, we summarize further research that could help more comprehensively assess AI data center power needs and approaches to reducing power supply bottlenecks.

Suggestions for Future Research

We summarize some topics for research that could help better quantify power supply challenges and identify and assess causes and solutions for bottlenecks in AI data center build-out.

Quantifying Data Center Power Demand and Supply

- **Model future increases in power grid supply and compare them with data center demand.** A comprehensive model of power supply in the United States could help predict increases in supply and demand. This model should factor in the reliability and availability of different energy sources to adequately account for the fact that data centers need reliable power around the clock.
- **Assess factors that may reduce future power demand of AI data centers.** Research on decentralized training could help understand how much power AI training will require in a single location. Research on potential efficiency increases in AI chips could further increase accuracy of demand forecasts. Finally, modeling scenarios for future AI chip production would directly inform future data center demand.
- **Continue studying potential bottlenecks to scaling of training and inference compute.** One key uncertainty is the extent to which frontier AI developers will continue scaling the compute used to train and run inference on AI systems. For example, it is possible that training will soon be bottlenecked by latency issues (Erdil, 2024) or that data scarcity could undermine the business case for continued scaling (Villalobos et al., 2024). Although previous work suggests that power capacity bottlenecks will be reached before other constraints, further research is needed (Sevilla et al., 2024).

Causes and Solutions for Data Center Construction Challenges

- **Identify which state or federal regulations may limit the expansion of U.S. energy capacity for AI training runs and assess the feasibility and desirability of streamlining these regulatory processes.** Regulations—such as environmental review processes, permitting requirements, and other regulatory standards—affect the development and implementation of projects that generate and distribute energy, such as the development of transmission lines (Datta and Coleman, 2024). Recent commentators have analyzed how these regulatory processes may limit clean energy generation (Potter, Datta, and Stapp, 2022). Future research could similarly assess the extent to which these regulations limit power availability for future AI development. If regulatory barriers do pose obstacles to energy construction, future work could investigate ways that these regulatory processes may be reformed or streamlined. For example, in the past, federal agencies have used programmatic environmental assessments and impact statements (National Institute of Justice, 2019) to expedite environmental review for large infrastructure projects (Bureau of Land Management and U.S. Department of Energy, 2010). Future work could identify similar approaches that could simplify permitting for data center–related projects.
- **Investigate opportunities for expanding power generation and distribution for data centers.** AI compute providers have indicated particular interest in natural gas (Kimball, 2024), small modular nuclear reactors (Ohnsman, 2024), and geothermal energy ("Meta Platforms Strikes Geothermal Energy Deal to Power US Data Centers," 2024) to scale up their compute resources. Compute providers have also indicated interest in improving the

electrical grid's capacity to deliver energy to data centers (Petersen, 2024). The Department of Energy identified several opportunities for energy generation and distribution to be expanded to meet the power demands of data centers (U.S. Department of Energy, 2024). Future research could assess the feasibility, scalability, and environmental impact of the Department of Energy's suggestions.

- Develop and evaluate options for state and federal responses to energy shortfalls. The federal government has historically leveraged government-owned resources to accelerate power generation and transmission (Raby, 2017). Future research could assess the extent to which these resources (e.g., unused federal land) could be used for data center–related projects. Additionally, compute and energy for AI may be critical for national security (Pavel et al., 2023). Federal authorities related to emergency response or defense preparedness could be leveraged to expand U.S. energy production (Majkut and Nakano, 2023). For example, the Defense Production Act has been used to expand clean energy production (U.S. Department of Energy, 2022). Future research could determine whether similar authorities could be used to bolster energy production and distribution for data centers.
- Assess the ability of private companies to meet energy shortfalls. AI companies have successfully raised billions of dollars (Ghaffary et al., 2024), enabling them to spend tens of millions of dollars on AI training runs (Cottier et al., 2024). However, some commentators project that frontier models could cost billions of dollars to train by 2030 (Scharre, 2024). As this paper suggests, energy demand is a key driver of AI training and inference costs. Future research could assess AI and data center companies' ability to generate and distribute enough energy to satisfy future demand, with or without government support.

Appendix A

Approach, Methods, and Sources

In this appendix, we describe the assumptions, sources, and methods used to generate Figures 1.1 and 1.2, as well as our approach to identifying current challenges for data centers and further research.

Method and Sources for Figure 1.1

In Figure 1.1, we display an estimate of the future power needs of AI data centers, assuming continuing exponential growth in chip supply. Patel, Nishball, and Ontiveros (2024) report that data center information technology (IT) power demand grew from 3.3 GW in 2023 to 8.5 GW in 2024, which corresponds to a data center power capacity—the power required for the entire data center, including not just the hardware but also cooling and power conversion—of 4.3 GW and 10.6 GW, respectively (see Table A.1). Sevilla et al. (2024) estimate that global chip supply could grow between 1.3 and 2 times annually between 2024 and 2030, with a median of 1.7 times.

For our main projection (the curve labeled "Estimate based on growth in AI chip supply" in Figure 1.1), we display the median estimate, which anchors on the SemiAnalysis estimate for 2024 of 10.6 GW and projects it forward by multiplying the chip production growth of the previous year by 1.7 for each year. We further display an uncertainty interval around the median estimate, derived from 1.3 and 2 times growth, respectively. We estimated the total data center power capacity by multiplying by the PUE for each year (see Table A.1). To account for efficiency gains, we assumed continual improvement in PUE from 1.25 in 2024 to 1.05 in 2030 (Patel, Nishball, and Ontiveros, 2024).

We further display three external estimates by Goldman Sachs (2024); Srivathsan et al. (2024), labeled "McKinsey"; and Patel, Nishball, and Ontiveros (2024), labeled "SemiAnalysis." We display the y-axis on a log scale to better capture exponential increases in demand.

To contextualize the scale of power demand, we add the capacity of three U.S. states: Utah, Virginia, and California (EIA, 2023).

Tables A.1 through A.4 provide an overview of the data used in Figure 1.1.

Table A.1. SemiAnalysis Data on AI Data Center Power Demand

Year	IT Power (MW)	PUE (with Efficiency Increases)	Data Center Facility Power Demand (GW)
2020	318	1.38	0.440
2021	640	1.35	0.864
2022	1,102	1.32	1.451
2023	3,332	1.28	4.28
2024	8,499	1.25	10.6
2025	16,356	1.22	19.9
2026	28,140	1.18	33.3
2027	41,337	1.15	47.5
2028	56,280	1.12	62.8

SOURCE: Authors' analysis of data from Patel, Nishball, and Ontiveros, 2024.

Table A.2. AI Data Center Power Demand Data for Estimate Based on Growth in AI Chip Supply

	Data Center IT Power Demand Based on Exponential Chip Growth				Data Center Facility Power Demand Based on Exponential Chip Growth		
Year	Lower Confidence Interval	Median	Upper Confidence Interval	PUE (with Efficiency Increases)	Lower Confidence Interval	Median	Upper Confidence Interval
2024	5.2	5.2	5.2	1.25	6.5	6.5	6.5
2025	6.7	8.8	10.3	1.22	8.2	10.7	12.6
2026	8.7	14.9	20.7	1.18	10.3	17.7	24.5
2027	11.4	25.4	41.3	1.15	13.1	29.2	47.5
2028	14.8	43.2	82.7	1.12	16.5	48.2	92.3
2029	19.2	73.4	165.3	1.08	20.8	79.5	179.1
2030	24.9	124.7	330.7	1.05	26.2	131.0	347.2

SOURCES: Growth factors are derived from Sevilla et al., 2024; starting value in 2024 is derived from Patel, Nishball, and Ontiveros, 2024.

Table A.3. AI Data Center Power Demand Data from Goldman Sachs

Year	U.S. AI (TWh)	Other AI (TWh)	Sum (TWh)	Sum (GW)
2020	2	3	5	0.6
2021	2	4	6	0.7
2022	3	5	8	0.91
2023	4	8	12	1.37
2024	11	19	30	3.4
2025	22	36	58	6.6
2026	37	57	94	10.7
2027	53	78	131	15.0
2028	69	95	164	18.7
2029	83	107	190	21.7
2030	93	116	209	23.9

SOURCE: Authors' analysis of data from Goldman Sachs, 2024.

Table A.4. AI Data Center Power Demand from McKinsey

Year	AI Power Demand (GW)
2023	8.3
2024	11.4
2025	14
2026	17.1
2027	24.7
2028	37.2
2029	57.2
2030	89.9

SOURCE: Authors' analysis of data from Srivathsan et al., 2024.

Method and Sources for Figure 1.2

For Figure 1.2, we extrapolate the power required for future AI compute clusters based on trends in training compute growth and efficiency improvements in AI accelerators.

To estimate the power required for a training run each year, we first calculate the AI accelerator efficiency:

$$Efficiency = Performance \ / \ Power\ Capacity$$

We assume a hardware efficiency based on NVIDIA's DGX H100—the most used AI chip for frontier AI training at the time of this writing—which has a performance of 1.98e15 8-bit FLOP per second (FLOP/s) and requires 1,419 W per chip, considering all cluster components (NVIDIA, 2023). To model efficiency increases, we assume a continuation of the 1.3-times FLOP per watt improvement trend for AI chips each year (Hobbhahn, Heim, and Aydos, 2023). For a given year, we calculate the efficiency as follows:

$$Efficiency = Performance \ / \ Power\ Capacity \times 1.3^{(Year\ -\ 2023)}$$

We further estimated the cluster performance required in FLOP/s for a given training run by dividing the total training compute in FLOP by the training duration, which we assume to be 100 days in December 2023 and to linearly increase to 200 days by 2030 (Epoch AI, 2024).

$$Cluster\ Performance = Training\ Compute \ / \ Training\ Duration$$

At the time of this writing, Google Gemini, released on December 6, 2023, was the largest known training run at 5e25 FLOP (Epoch AI, 2024). The FLOP used to train frontier AI models has increased by 4 to 5 times each year between 2010 and 2024 (Sevilla and Roldán, 2024). For estimating the size of the largest training run for a given date, we anchor on Gemini's training run and then increase the training compute by 4 times per year. Because Gemini likely trained for about 100 days, we display the cluster needed to train it 100 days before its release in December 2023 (Epoch AI, 2024).

We then estimate the total data center power capacity required, accounting for the fact that, for large training runs, AI hardware typically has a low utilization factor: 0.34 in the case of GPT-4 (Patel and Wong, 2023). We further multiply by a PUE factor, which is the conversion factor from IT power capacity to total data center power capacity. Because of efficiency improvements, we assume that PUE linearly decreases from 1.25 in 2023 to 1.05 by 2030 (Patel, Nishball, and Ontiveros, 2024; Fist and Datta, 2024).

$$Data\ Center\ Power\ Capacity = Cluster\ Performance \ / \ (Efficiency \times 0.34) \times PUE$$

Using the data center power capacity required over time, we display data for a hypothetical cluster required for an illustrative training run every two years. The first data point is a hypothetical cluster used for training Google's Gemini. The following three data points are clusters required for training models consistent with training compute and hardware trends. For each data point, we show the

number of 8-bit FLOP used. We further show the number of H100 that would have the same performance.[6] For the GPU cost, we assumed that an H100 cost $30,000 in 2023 and that price-performance in cost per FLOP/s improves by 2 times every 2.1 years (Hobbhahn, Heim, and Aydos, 2023; Eadline, 2023).

To put the scale of required power into perspective, we add three horizontal lines as reference points:

- The generation capacity of a typical nuclear reactor (Office of Nuclear Energy, 2022).
- The total data center capacity of the Northern Virginia data center hub in 2023, the currently largest aggregation of data centers in the world (Clabaugh, 2023).
- China's Three Gorges Dam, the world's largest power plant, supposed to provide an upper bound on how large power plants can be (U.S. Geological Survey, undated).

We run our code in Google Colab to estimate the increase in power requirements each year and to generate the plot.

Methods for Summarizing Data Center Construction Challenges and Suggesting Topics for Further Research

To understand the challenges that AI data center construction currently faces, we consulted multiple sources. We collected academic publications through Google Scholar, using such search terms as "AI data center challenges," "AI data center power," and "AI data center permitting." However, because the rapid growth in AI data center power requirements is a recent phenomenon, academic research is currently limited. While existing studies analyze and project data center power requirements, they predate the current surge in AI-driven demand. Although some publications explore specific energy efficiency improvements, they do not address the macro-level trends in AI power requirements that are central to our analysis.

To capture current industry developments, we expanded our literature search to include industry sources and publications. We monitored leading industry news sources—including Data Center Dynamics, Data Center Knowledge, and Data Centre Magazine—throughout our research process. We conducted targeted searches using such terms as "AI data center challenges," "AI data center power," and "AI data center permitting." Drawing on our previous research experience with compute supply chains and data centers, we identified authoritative sources, prioritizing government publications (such as Department of Energy reports) and industry sources with transparent methodologies. We also analyzed reports from SemiAnalysis, an industry intelligence group specializing in AI hardware and data center coverage, to understand emerging trends and challenges.

To find topics for further research, we additionally surveyed the literature for recent permitting challenges for data centers, power generation, and grid infrastructure. We also drew from discussions and readings suggested by other researchers throughout the research project.

[6] Note that the H100 equivalents are equivalent in performance but would require significantly more power, given the power trendline factors in efficiency improvements for future AI accelerators.

Appendix B

Limitations of Estimates

In this appendix, we discuss the limitations of our model approaches.

Total AI Data Center Power Needs (Figure 1.1)

General Trend

Our exponential growth scenario is more aggressive than the other three scenarios provided (Patel, Nishball, and Ontiveros, 2024; Srivathsan et al., 2024; and Goldman Sachs, 2024). Given that none of those reports fully explain its approach, it is not possible to directly assess the different results. However, our central assumption is that exponential demand growth will continue and the chip industry will grow exponentially to accommodate this growth (Sevilla et al., 2024). At least in the short term, this assumption is consistent with IEA (2024), which estimates that data center demand will increase tenfold between 2023 and 2026.

We assumed exponential growth because compute scaling has now lasted for more than a decade, and spending has increased by several orders of magnitude (Cottier et al., 2024; Sevilla and Roldán, 2024). Furthermore, key companies in AI compute—such as Meta, Amazon Web Services, and OpenAI—are planning to invest hundreds of billions into AI infrastructure (Vanian, 2024; Metz and Mickle, 2024; Herrera, 2024). Our extrapolation would be much more modest if investment increased less rapidly or should there be supply chain bottlenecks for AI chips. Yet, given sustained growth in both investment and chip supply, these two trends seem unlikely to occur within the next few years.

Efficiency Improvements

The U.S. Department of Energy (2024) suggests that future breakthroughs in energy efficiency of training and inference may reduce the power required for AI data centers. Our analysis already accounts for potential increases in data center efficiency by assuming that PUE will fall from 1.25 to 1.05. Additional improvements would be marginal because 1.0 is the lowest value PUE can be, indicating no power overhead for the data center infrastructure. Our analysis assumes that chip production volume will grow but power use per chip will remain constant. This is consistent with the overall trend in power required per chip of machine learning GPUs (Hobbhahn, Heim, and Aydos, 2023). However, although data on the most used GPUs is too scarce to find a statistically significant trend, Hobbhahn, Heim, and Aydos (2023) also note that the steady increase in power required for NVIDIA's last four AI chips (the V100, A100, H100, and B100) could suggest that power requirements are increasing. Yet, if power requirements of future chips fell while price-performance improvements continued, this would mean lower power demand from AI chips.

Lack of Approaches

Another limitation of our estimate is that it uses only one of several potential methods of estimating power requirements. We chose chip production primarily because it is simpler than such methods as aggregating power procurement by data center companies or extrapolating investment and because there were public estimates available for chip production, whereas the other data sources were not publicly available.

General Skepticism of Forecasts

Finally, there is also a broader reason to be skeptical of power estimates for new technologies, given that previous forecasts have overestimated power needs (Meyer, 2024). Yet, uncertainty about technology could lead our estimates to be skewed in either direction and thus does not necessarily indicate that our extrapolation is an overestimate.

Data Centers Hosting the Largest Training Runs (Figure 1.2)

Continued Rate of Compute Scaling

Our extrapolation assumes that AI companies will continue to increase computational resources at current rates until 2030 (Sevilla et al., 2024). However, various factors could lead to slower compute scaling rates, resulting in lower power requirements for the largest data centers. These factors include diminishing returns for scaling AI, bottlenecks (such as running out of training data), or lack of chip supply.

Despite these potential challenges, we believe the current scaling trends are likely to continue for several reasons:

- A 2024 Epoch AI investigation assessed potential limitations in training data, chip supply, and network architecture and found it plausible that current scaling rates could persist until 2030, given continued improvements in AI (Sevilla et al., 2024).
- The compute scaling trend has been consistent for more than a decade, suggesting a robust pattern (Sevilla et al., 2024).
- Empirical scaling laws suggest continued improvement of AI capabilities with increasing training compute (Villalobos, 2023).
- Major AI companies are planning significant investments in AI infrastructure, indicating their commitment to expanding compute resources (Vanian, 2024; Metz and Mickle, 2024; Herrera, 2024).

Although unforeseen events or technological shifts could alter this trajectory, the current evidence and industry behavior support our assumption of continued scaling in computational resources for AI training.

Efficiency Improvements

We factored in improvements in PUE from 1.25 to 1.05, and we accounted for AI chip efficiency increases by 1.3 times each year (Hobbhahn, Heim, Aydos, 2023; Fist and Datta, 2024). Yet, future hardware developments could lead to higher or lower efficiency gains, adding additional uncertainty.

Decentralized Training

Our extrapolation disregards the potential for training across several data center locations. Recent examples, such as Google's approach with Gemini training (Anil et al., 2024), show attempts at multi–data center distribution. However, details are limited (such as the number of data centers), and Google may have used only a small number of closely located facilities. Meanwhile, other reported efforts for large clusters rely on a single centralized location (Kahn, 2024; Trueman, 2024). If by 2030, the largest training run was conducted across four clusters, it would require about 1 GW of power for each of the four. Although this would still be an unprecedented amount of power for each of the four clusters, it could greatly reduce challenges with finding adequate capacity in a single location.

Abbreviations

AI	artificial intelligence
EIA	U.S. Energy Information Administration
FLOP	floating point operations
FLOP/S	floating point operations per second
GPU	graphics processing unit
GW	gigawatt
IEA	International Energy Agency
IT	information technology
MW	megawatt
PUE	power usage effectiveness
TWh	terawatt-hour
W	watt

Bibliography

"Abu Dhabi's US$302 Billion Fund Deepens AI and Private Credit Push," *Business Times*, May 17, 2024.

Allen, Gregory C., *Choking Off China's Access to the Future of AI: New U.S. Export Controls on AI and Semiconductors Mark a Transformation of U.S. Technology Competition with China*, Center for Strategic and International Studies, October 2022. As of November 12, 2024: https://www.csis.org/analysis/choking-chinas-access-future-ai

Anil, Rohan, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, et al., "Gemini: A Family of Highly Capable Multimodal Models," arXiv, arXiv:2312.11805, last updated June 17, 2024. As of September 18, 2024: https://arxiv.org/abs/2312.11805

Bauer, Lauren, Wendy Edelberg, Cameron Greene, Olivia Howard, and Linsie Zou, "Eight Facts About Permitting and the Clean Energy Transition," Brookings Institution, May 22, 2024. As of September 18, 2024: https://www.brookings.edu/articles/eight-facts-about-permitting-and-the-clean-energy-transition/

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atilim Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann, "Managing Extreme AI Risks amid Rapid Progress," *Science*, Vol. 384, No. 6698, May 20, 2024.

Bureau of Land Management and U.S. Department of Energy, "Description of Alternatives and Reasonably Foreseeable Development Scenario," in *Draft Programmatic Environmental Impact Statement for Solar Energy Development in Six Southwestern States*, December 2010. As of September 18, 2024: https://solareis.anl.gov/documents/dpeis/Solar_DPEIS_Chapter_2.pdf

Burkacky, Ondrej, Mark Patel, Klaus Pototzky, Diana Tang, Rutger Vrijen, and Wendy Zhu, "Generative AI: The Next S-Curve for the Semiconductor Industry?" McKinsey & Company, March 29, 2024. As of September 18, 2024: https://www.mckinsey.com/industries/semiconductors/our-insights/generative-ai-the-next-s-curve-for-the-semiconductor-industry

Calacanis, Jason, "CoreWeave's Brannin McBee on the Future of AI Infrastructure, GPU Economics, and Data Centers," *This Week in Startups* podcast, April 3, 2024. As of September 18, 2024: https://thisweekinstartups.com/episodes/V0CIfa08GrX

CBRE, "North America Data Center Trends H1 2024," August 19, 2024. As of November 12, 2024: https://www.cbre.com/insights/reports/north-america-data-center-trends-h1-2024

Chou, Brenda, "Risks of the Microsoft-G42 Deal," American Security Project, July 15, 2024. As of September 18, 2024: https://www.americansecurityproject.org/risks-of-the-microsoft-g42-deal/

Clabaugh, Jeff, "Northern Virginia Remains the World's Data Center Capital—Here's How It Got There," WTOP, April 24, 2023. As of September 18, 2024: https://wtop.com/business-finance/2023/04/northern-virginia-remains-the-worlds-data-center-capital-how-it-got-there/

Coleman, Jude, "AI's Climate Impact Goes Beyond Its Emissions," *Scientific American*, December 7, 2023.

Cottier, Ben, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen, "How Much Does It Cost to Train Frontier AI Models?" Epoch AI, June 3, 2024. As of September 18, 2024: https://epochai.org/blog/how-much-does-it-cost-to-train-frontier-ai-models

Datta, Arnab, and James Coleman, "We Must End the Litigation Doom Loop," Institute for Progress, May 29, 2024. As of September 18, 2024: https://ifp.org/we-must-end-the-litigation-doom-loop/

Department of Ecology, State of Washington, "Data Centers," webpage, undated. As of October 11, 2024: https://ecology.wa.gov/air-climate/air-quality/data-centers

Dohmen, Hanna, and Jacob Feldgoise, "A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports," Center for Security and Emerging Technology, December 4, 2023. As of September 18, 2024: https://cset.georgetown.edu/article/bis-2023-update-explainer/

Dotan, Tom, and Asa Fitch, "Why the AI Industry's Thirst for New Data Centers Can't Be Satisfied," *Wall Street Journal*, April 24, 2024.

Eadline, Doug, "Nvidia H100: Are 550,000 GPUs Enough for This Year?" *HPCwire*, August 17, 2023. As of January 17, 2025: https://www.hpcwire.com/2023/08/17/nvidia-h100-are-550000-gpus-enough-for-this-year/

Edwards, Benj, "Nvidia's Powerful H100 GPU Will Ship in October," Ars Technica, September 20, 2022. As of December 5, 2024: https://arstechnica.com/information-technology/2022/09/hopper-time-nvidias-most-powerful-ai-chip-yet-ships-in-october/

Egan, Janet, and Lennart Heim, *Oversight for Frontier AI Through a Know-Your-Customer Scheme for Compute Providers*, Center for the Governance of AI, October 25, 2023. As of September 18, 2024: https://www.governance.ai/research-paper/oversight-for-frontier-ai-through-kyc-scheme-for-compute-providers

EIA—*See* U.S. Energy Information Administration.

Epoch AI, "Notable AI Models," webpage, June 19, 2024. As of November 12, 2024: https://epoch.ai/data/notable-ai-models

Erdil, Ege, "Data Movement Bottlenecks to Large-Scale Model Training: Scaling Past 1e28 FLOP," Epoch AI, November 2, 2024. As of November 12, 2024: https://epochai.org/blog/data-movement-bottlenecks-scaling-past-1e28-flop

Fist, Tim, and Arnab Datta, *How to Build the Future of AI in the United States: Part Two of Compute in America*, Institute for Progress, October 23, 2024. As of November 12, 2024: https://ifp.org/future-of-ai-compute/

Fist, Tim, and Erich Grunewald, "Preventing AI Chip Smuggling to China: A Working Paper," Center for a New American Security, October 24, 2023. As of September 18, 2024: https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china

Fist, Tim, and Paul Scharre, "The Cloud Can Solve America's AI Problems," *Foreign Policy*, October 7, 2023.

Ghaffary, Shirin, Katie Roof, Rachel Metz, and Dina Bass, "OpenAI Raises $6.6 Billion in Funds at $157 Billion Value," Bloomberg, October 2, 2024.

Goldman Sachs, "AI Is Poised to Drive 160% Increase in Data Center Power Demand," May 14, 2024. As of November 12, 2024: https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

"Google Falling Short of Important Climate Target, Cites Electricity Needs of AI," NBC News, July 2, 2024.

Heim, Lennart, Markus Anderljung, Emma Bluemke, and Robert Trager, "Computing Power and the Governance of AI," Center for the Governance of AI, February 14, 2024. As of September 18, 2024: https://www.governance.ai/post/computing-power-and-the-governance-of-ai

Herrera, Sebastian, "Amazon, Built by Retail, Invests in Its AI Future," *Wall Street Journal*, June 30, 2024.

Hobbhahn, Marius, Lennart Heim, and Gökçe Aydos, "Trends in Machine Learning Hardware," Epoch AI, November 9, 2023. As of September 18, 2024: https://epochai.org/blog/trends-in-machine-learning-hardware

IEA—*See* International Energy Agency.

International Energy Agency, *Electricity 2024: Analysis and Forecast to 2026*, January 2024. As of October 10, 2024: https://www.iea.org/reports/electricity-2024

Kahn, Jeremy, "Is Microsoft's $100 Billion 'Stargate' OpenAI Supercomputer AI's 'Star Wars' Moment?" *Fortune*, April 2, 2024.

Khan, Saif M., and Alexander Mann, *AI Chips: What They Are and Why They Matter*, Center for Security and Emerging Technology, April 2020. As of October 11, 2024: https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/

Kimball, Spencer, "AI Could Drive a Natural Gas Boom as Power Companies Face Surging Electricity Demand," CNBC, May 5, 2024.

Kurtz, Josh, "Centers of Controversy: Is There Enough Energy for Md. to Meet Its Tech Ambitions?" Maryland Matters, December 11, 2023. As of December 6, 2024: https://marylandmatters.org/2023/12/11/centers-of-controversy-is-there-enough-energy-for-md-to-meet-its-tech-ambitions/

Majkut, Joseph, and Jane Nakano, "The Defense Production Act and the U.S. Race to Build Up Clean Energy Industrial Bases," Center for Strategic and International Studies, January 12, 2023. As of September 18, 2024: https://www.csis.org/analysis/defense-production-act-and-us-race-build-clean-energy-industrial-bases

McGeady, Cy, "Strategic Perspectives on U.S. Electric Demand Growth," Center for Strategic and International Studies, May 20, 2024. As of September 18, 2024: https://www.csis.org/analysis/strategic-perspectives-us-electric-demand-growth

McGrath, Glenn, "Power Blocks in Natural Gas-Fired Combined-Cycle Plants Are Getting Bigger," U.S. Energy Information Administration, February 12, 2019. As of September 18, 2024: https://www.eia.gov/todayinenergy/detail.php?id=38312

"Meta Platforms Strikes Geothermal Energy Deal to Power US Data Centers," Reuters, August 26, 2024.

Metz, Cade, and Tripp Mickle, "Behind OpenAI's Audacious Plan to Make A.I. Flow Like Electricity," *New York Times*, September 25, 2024.

Meyer, Robinson, "Is AI Really About to Devour All Our Energy?" HeatMap News, April 16, 2024. As of November 12, 2024: https://heatmap.news/technology/ai-energy-consumption

Microsoft, "Microsoft Invests $1.5 Billion in Abu Dhabi's G42 to Accelerate AI Development and Global Expansion," April 15, 2024a. As of September 18, 2024: https://news.microsoft.com/2024/04/15/microsoft-invests-1-5-billion-in-abu-dhabis-g42-to-accelerate-ai-development-and-global-expansion/

Microsoft, "Microsoft and G42 Announce $1 Billion Comprehensive Digital Ecosystem Initiative for Kenya," May 22, 2024b. As of September 18, 2024: https://news.microsoft.com/2024/05/22/microsoft-and-g42-announce-1-billion-comprehensive-digital-ecosystem-initiative-for-kenya/

Moss, Sebastian, "Meta's Mark Zuckerberg Says Energy Constraints Are Holding Back AI Data Center Buildout," Data Centre Dynamics, April 19, 2024a. As of September 18, 2024: https://www.datacenterdynamics.com/en/news/metas-mark-zuckerberg-says-energy-constraints-are-holding-back-ai-data-center-buildout/

Moss, Sebastian, "Three Mile Island Nuclear Power Plant to Return as Microsoft Signs 20-Year, 835MW AI Data Center PPA," Data Centre Dynamics, September 20, 2024b. As of November 12, 2024: https://www.datacenterdynamics.com/en/news/three-mile-island-nuclear-power-plant-to-return-as-microsoft-signs-20-year-835mw-ai-data-center-ppa/

National Institute of Justice, "NEPA Analysis: Programmatic Environmental Assessment," September 9, 2019. As of September 18, 2024: https://nij.ojp.gov/funding/nepa-analysis-programmatic-environmental-assessment

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024. As of September 18, 2024: https://www.rand.org/pubs/research_reports//RRA2849-1.html

NVIDIA, *NVIDIA DGX SuperPOD Data Center Design: Reference Guide, Featuring NVIDIA DGX H100 Systems*, DG-11301-001 v4, May 2023. As of September 18, 2024: https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/index.html

Office of Energy Efficiency and Renewable Energy, "Offshore Wind Energy: What Is Offshore Wind Energy?" webpage, U.S. Department of Energy, undated. As of September 18, 2024: https://windexchange.energy.gov/markets/offshore

Office of Nuclear Energy, "Nuclear Power Is the Most Reliable Energy Source and It's Not Even Close," blog post, March 24, 2021, updated July 2022a. As of September 18, 2024: https://www.energy.gov/ne/articles/nuclear-power-most-reliable-energy-source-and-its-not-even-close

Ohnsman, Alan, "Desperate for Power, AI Companies Look to the Nuclear Option," *Forbes*, June 10, 2024.

Patel, Dylan, Daniel Nishball, and Jeremie Elliahou Ontiveros, "AI Datacenter Energy Dilemma—Race for AI Datacenter Space," SemiAnalysis, March 13, 2024. As of September 18, 2024: https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race

Patel, Dylan, and Gerald Wong, "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE," SemiAnalysis, July 10, 2023. As of September 18, 2024: https://www.semianalysis.com/p/gpt-4-architecture-infrastructure

Pavel, Barry, Ivana Ke, Michael Spirtas, James Ryseff, Lea Sabbag, Gregory Smith, Keller Scholl, and Domenique Lumpkin, *AI and Geopolitics: How Might AI Affect the Rise and Fall of Nations?* RAND Corporation, PE-A3034-1, November 2023. As of October 11, 2024: https://www.rand.org/pubs/perspectives/PEA3034-1.html

Petersen, Melody, "Power-Hungry AI Data Centers Are Raising Electric Bills and Blackout Risk," *Los Angeles Times*, August 12, 2024.

Pilz, Konstantin, and Lennart Heim, "Compute at Scale: A Broad Investigation into the Data Center Industry," arXiv, arXiv:2311.02651, last updated November 22, 2023. As of September 18, 2024: https://arxiv.org/abs/2311.02651

Potter, Brian, *How to Save America's Transmission System*, Institute for Progress, February 2024. As of September 18, 2024: https://ifp.org/how-to-save-americas-transmission-system/

Potter, Brian, Arnab Datta, and Alec Stapp, *How to Stop Environmental Review from Harming the Environment*, Institute for Progress, September 13, 2022. As of October 11, 2024: https://ifp.org/environmental-review/

Raby, John K., "Designated Leasing Areas (DLAs) in Land Use Plans/Amendments," Bureau of Land Management, U.S. Department of the Interior, Instruction Memorandum No. MT-2017-028, April 21, 2017. As of October 11, 2024: https://www.blm.gov/policy/instruction-memorandum-no-mt-2017-028

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle, "Computing Power and the Governance of Artificial Intelligence," arXiv, arXiv:2402.08797, February 13, 2024. As of September 18, 2024: https://arxiv.org/abs/2402.08797

Saul, Josh, "Data Centers Face Seven-Year Wait for Dominion Power Hookups," Bloomberg, August 29, 2024.

Scharre, Paul, *Future-Proofing Frontier AI Regulation: Projecting Future Compute for Frontier AI Models*, Center for a New American Security, March 2024. As of September 18, 2024: https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation

Sevilla, Jaime, Tamay Besiroglu, Ben Cottier, Josh You, Edu Roldán, Pablo Villalobos, and Ege Erdil, "Can AI Scaling Continue Through 2030?" Epoch AI, August 20, 2024.

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning," arXiv, arXiv:2202.05924, March 9, 2022. As of September 18, 2024: https://arxiv.org/abs/2202.05924

Sevilla, Jaime, and Edu Roldán, "Training Compute of Frontier AI Models Grows by 4–5x per Year," Epoch AI, May 28, 2024. As of September 18, 2024: https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

Shoaib, Zaeem, "Big Tech Moves More AI Spending Abroad," *Wall Street Journal*, May 23, 2024.

Spivack, Miranda S., "More Data in the Cloud Means More Centers on the Ground to Move It," *New York Times*, June 27, 2023.

Srivathsan, Bhargs, Marc Sorel, Arjita Bhan, Pankaj Sachdeva, Haripreet Batra, Raman Sharma, Rishi Gupta, and Surbhi Choudhary, "AI Power: Expanding Data Center Capacity to Meet Growing Demand," McKinsey & Company, October 29, 2024. As of December 5, 2024: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand

Taylor, Petroc, "Data Center Average Annual Power Usage Effectiveness (PUE) Worldwide from 2007 to 2024" webpage, Statista, December 10, 2024. As of September 18, 2024: https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/

Tozzi, Christopher, "Why You Can't Power Your Data Center Only with Renewables—But Should Try Anyway," Data Center Knowledge, April 6, 2023. As of September 18, 2024: https://www.datacenterknowledge.com/cooling/why-you-can-t-power-your-data-center-only-with-renewables-but-should-try-anyway

Trueman, Charlotte, "xAI's Memphis Supercluster Has Gone Live, with up to 100,000 Nvidia H100 GPUs," Data Centre Dynamics, July 23, 2024. As of September 18, 2024: https://www.datacenterdynamics.com/en/news/xais-memphis-supercluster-has-gone-live-with-up-to-100000-nvidia-h100-gpus/

Uptime Institute, "Tier Classification System," webpage, undated. As of October 10, 2024: https://uptimeinstitute.com/tiers

U.S. Department of Energy, "President Biden Invokes Defense Production Act to Accelerate Domestic Manufacturing of Clean Energy," June 6, 2022. As of October 11, 2024: https://www.energy.gov/articles/president-biden-invokes-defense-production-act-accelerate-domestic-manufacturing-clean

U.S. Department of Energy, "Recommendations on Powering Artificial Intelligence and Data Center Infrastructure," July 30, 2024.

U.S. Energy Information Administration, "Annual Energy Outlook 2023—Table: Table 9. Electricity Generating Capacity—Case: Multiple Cases," webpage, undated. As of November 12, 2024: https://www.eia.gov/outlooks/aeo/data/browser/#/?id=9-AEO2023&region=0-0&cases=ref2023~highmacro~lowmacro&start=2021&end=2050&f=A&linechart=ref2023-d020623a.14-9-AEO2023~highmacro-d020623a.14-9-AEO2023~lowmacro-d020623a.14-9-AEO2023&ctype=linechart&sourcekey=0

U.S. Energy Information Administration, "US Electricity Profile 2022," webpage, November 2, 2023. As of December 30, 2024: https://www.eia.gov/electricity/state/archive/2022/

U.S. Energy Information Administration, "U.S. Power Grid Added 20.2 GW of Generating Capacity in the First Half of 2024," August 19, 2024. As of December 30, 2024: https://www.eia.gov/todayinenergy/detail.php?id=62864

U.S. Geological Survey, "Three Gorges Dam, China," webpage, undated. As of September 18, 2024: https://eros.usgs.gov/earthshots/three-gorges-dam-china

U.S. Geological Survey, "How Many Homes Can an Average Wind Turbine Power?" webpage, updated February 17, 2022. As of September 18, 2024: https://www.usgs.gov/faqs/how-many-homes-can-average-wind-turbine-power

Vanian, Jonathan, "Mark Zuckerberg Indicates Meta Is Spending Billions of Dollars on Nvidia AI Chips," CNBC, January 18, 2024.

Villalobos, Pablo, "Scaling Laws Literature Review," Epoch AI, January 26, 2023. As of October 11, 2024: https://epochai.org/blog/scaling-laws-literature-review

Villalobos, Pablo, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn, "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data," Epoch AI, June 6, 2024. As of November 12, 2024: https://epochai.org/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Wong, Wylie, "Going Nuclear: A Guide to SMRs and Nuclear-Powered Data Centers," Data Center Knowledge, November 29, 2023. As of November 12, 2024: https://www.datacenterknowledge.com/energy-power-supply/going-nuclear-a-guide-to-smrs-and-nuclear-powered-data-centers

World Nuclear Association, "Country Profiles: Nuclear Power in the United Arab Emirates," webpage, updated September 5, 2024. As of September 18, 2024: https://world-nuclear.org/information-library/country-profiles/countries-t-z/united-arab-emirates

About the Authors

Konstantin F. Pilz is a research assistant at RAND. His research focuses on data centers and AI supercomputers and their importance for training and deployment of frontier AI models. He holds a B.S. in biology and is a candidate for an M.A. in security studies.

Yusuf Mahmood is a technology and security policy fellow at RAND. He conducts policy research on the national security implications of AI. He holds B.A. degrees in economics and philosophy and is a J.D. candidate.

Lennart Heim is an information scientist at RAND and a professor of policy analysis at the Pardee RAND Graduate School. His research focuses on the role of compute for advanced AI systems and how compute can be leveraged as an instrument for AI governance, with an emphasis on policy development and security implications. Heim's publications cover the impacts and governance of advanced AI systems and empirical trends in machine learning, such as compute, data, and AI hardware. He holds an M.S. in electrical engineering.