EPRI

EPOCH AI

2025 White Paper

Scaling Intelligence: The Exponential Growth of AI's Power Needs

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The rapid advancement of artificial intelligence (AI)—particularly the training of large-scale "frontier models"—is driving renewed growth in electricity demand. This report analyzes the technical drivers of AI power consumption, projects future demand trajectories for individual training sites and broader AI needs, and highlights energy sector implications.

Key Findings

Power Growth for Frontier AI Training

Frontier AI training runs—the computationally intensive process of training large, advanced AI models—currently consume approximately **100–150 megawatts (MW)** each and are projected to reach **1–2 gigawatts (GW) each by 2028**, exceeding **4 GW per training run by 2030**. Figure 1 displays estimated power usage of recent frontier models, projected growth through 2030, and data on select AI data centers under construction or in planning. In the figure:

- *Historic baseline* reflects a continuation through 2030 of the observed **2.2× annual growth** in these models' peak power demand from 2018 to 2025.
- **Higher projections** are derived from a model incorporating three key drivers of training power demand: **training compute growth**, **hardware efficiency improvements**, and **training run duration**. The model projects faster growth in peak power needs than the historical baseline, primarily due to expected limits on training duration increases that spread computations over time.
  - *Training compute has grown at* **4.2× per year** *since 2018. This trend has persisted across architectures and modalities, driven by consistent performance gains from scaling.*
  - *Hardware efficiency is assumed to improve by* **33–52% annually**, *reflecting both historical trends and anticipated gains from lower precision numeric formats.*
  - *Training duration is assumed to grow* **10–20% annually**, *down from recent growth rates of* **25–50%**. *Increasing duration spreads the same energy demand across a longer period of time resulting in less peak power. As durations now exceed 100 days, further increases may face diminishing returns and competitive constraints.*

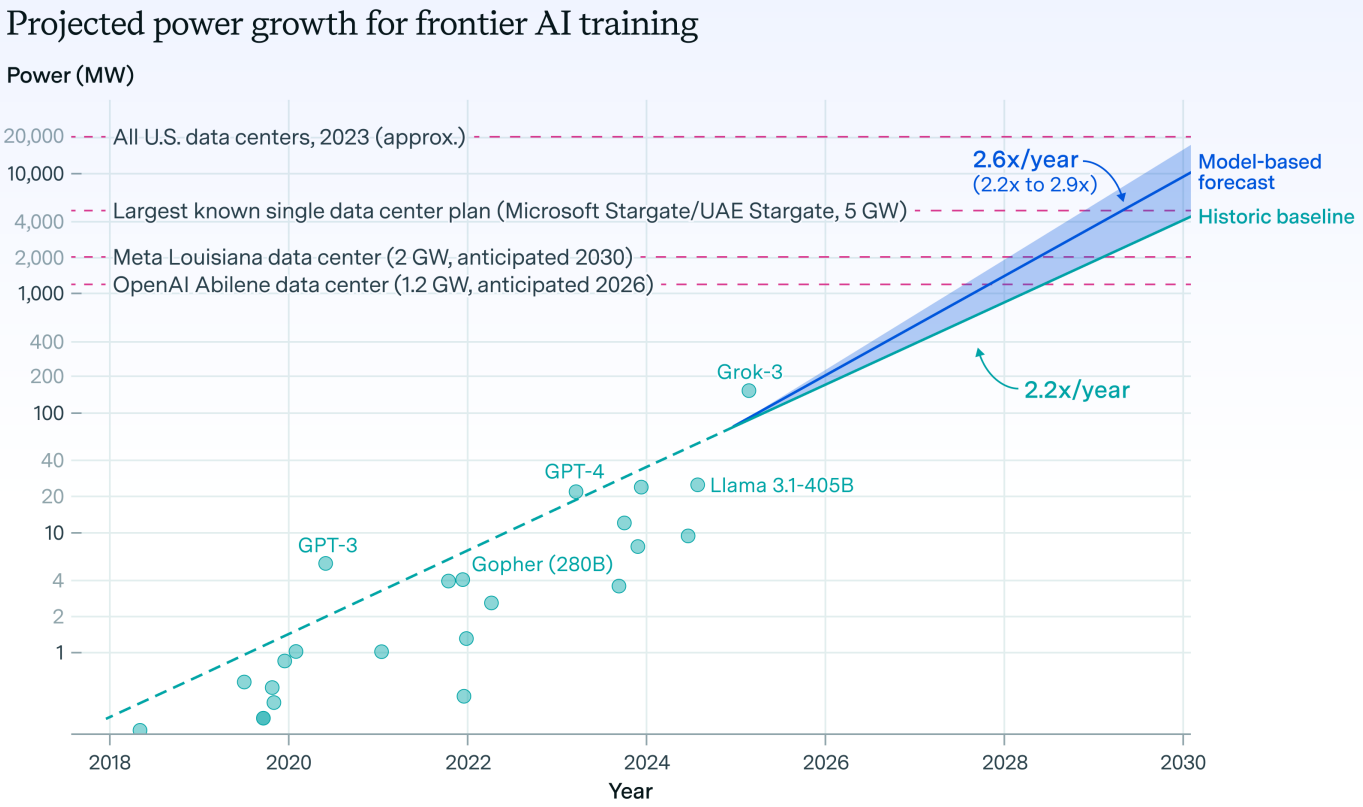


Figure 1. *Forecast for peak power demand required to train the largest frontier models, with historic frontier AI power growth and historic training runs highlighted for context. Graph prepared by Epoch AI.*

Training data centers—facilities specifically designed to support the intensive computation needs of training machine learning (ML) models—cannot keep doubling in size forever. Geographically distributed training is a possible strategy to overcome local power delivery limits. Synchronization across data centers separated by **15–50 miles** has been demonstrated, with potential for broader geographic spread.

Projected Total Power Capacity of AI Data Centers in the United States

U.S. AI power capacity is estimated at **5 GW** today and could reach more than **50 GW** by 2030, as shown in Figure 2. If multiple companies were to each develop a 4-GW training cluster, AI training could consume a significant fraction of that load.

- Despite uncertainty in both current data and projections, alternative growth estimates—based on chip deliveries, hyperscaler capital expenditures, and data center data—yield broadly similar results.
- **Training versus inference split is uncertain** and important. This split could affect the size, location, power demands, and potential flexibility of AI data centers.

In recent years, OpenAI and Google have reported similar power allocations for training and inference—the process of using a trained AI model—but the landscape is changing rapidly. Reasoning models may shift demand toward inference, but training remains a major driver due to continued scaling.

Implications for the Energy Sector

- **AI is the dominant near-term driver of data center power growth**, with hyperscaler capital expenditure (capex) exceeding $370 billion in 2025. Planned data centers (e.g., OpenAI's 1.2-GW Stargate, Meta's 2+-GW Louisiana campus) reflect this investment trajectory. Although materializing more slowly, power demands from electrification of the economy ultimately could be much larger; however, electrification creates different electric system development needs.
- **Forecasts suggest AI could consume over 5% of U.S. generation capacity by 2030**. Power demand from individual training runs may rival the output of major power plants, requiring new approaches to grid planning, permitting, and infrastructure investment.
- **Planning should account for both concentrated and distributed data center loads** as well as the potential for **real-time flexibility** in training and inference workloads and from on-site generation and storage assets.

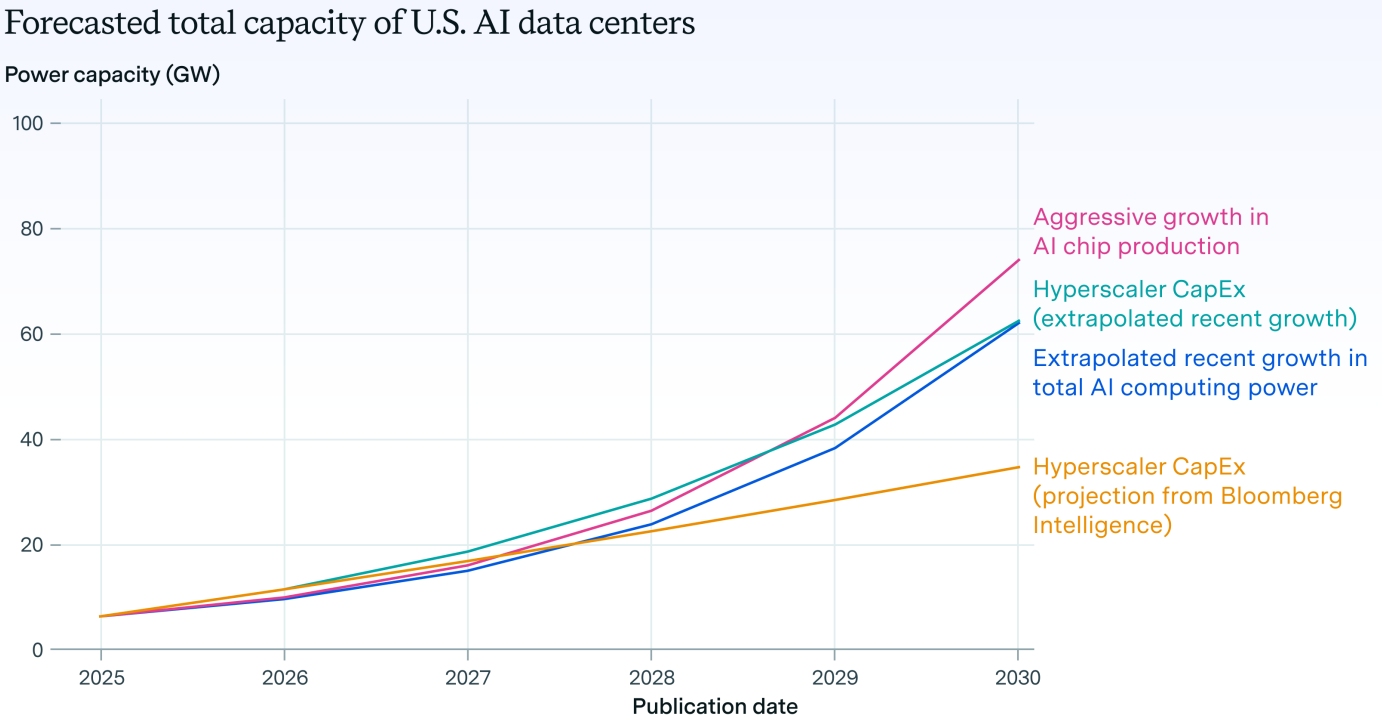


Figure 2. *Forecasted capacity of U.S. AI data centers using various approaches. Graph prepared by Epoch AI.*

INTRODUCTION

The AI industry has grown rapidly in recent years. Since early 2023, applications like ChatGPT have scaled from niche adoption to hundreds of millions of users. In response, leading AI companies are collectively allocating hundreds of billions of dollars annually to AI infrastructure. This surge in investment is driving substantial increases in the power demands of data centers that train and deploy AI.

While AI makes up only a small fraction of data center power demand today, it is the largest driver of both recent and projected growth. U.S. data center power demand doubled between 2018 and 2023, reaching 4.4% of the nation's total power consumption, and projections based on estimates of future chip sales indicate that data center power demand could more than double again over the next five years.[1]

Just as important from an electric systems perspective is data centers' concentrated power demand to train the largest and most capable AI models, also known as frontier models. The training of frontier models has historically required large, localized power supply. AI training consists of feeding a model vast amounts of data so it can learn patterns and relationships, and the largest training runs now exceed 100 MW. Power demand for frontier AI model training runs has been growing even faster than overall AI power demand, more than doubling every year, with some training clusters under development targeting capacities of 1 to 5 GW.

This report examines the trends driving AI power demand, with a focus on power demand for frontier AI model training.

Section 1 analyzes historical trends in power consumption of frontier AI models. It then assesses the factors—training compute growth, hardware efficiency, and training duration—that underpin their power demand and uses these factors to project possible future demands from individual sites.

Section 2 projects overall AI demands for power and apportions demand from the training of frontier models, inference, and other demands.

Section 3 presents key findings regarding overall AI demand.

[1] https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

1. GROWTH OF POWER DEMAND FOR FRONTIER AI MODEL TRAINING

Training large frontier AI models, like large language models, involves complex processes, massive datasets, and specialized hardware. These highly advanced, large-scale AI models often possess billions or even trillions of parameters and require immense computational capabilities. This section examines the requirements of frontier AI model training, explains the drivers behind the growth in power demand for training, and forecasts power demand for the training of individual frontier AI models through 2030.

- Section 1.1 reviews historic trends in power demand for training frontier AI models.
- In Section 1.2, training compute scaling is examined. Increasing the capacity of computational resources (like computing power, memory, and data) used to train a model—known as *training compute scaling*—is the primary driver of historic power demand increases. The evidence supporting the continued growth of training compute scaling is presented, and the relationship between growth in training compute and power demand is explained.
- In Section 1.3, other drivers of power demand are analyzed. These primarily include the efficiency of hardware—including AI chips, servers, and data centers—and training run durations.
- Section 1.4 brings together trends and projections for all the drivers to forecast growth in power demand for frontier AI training through 2030.

1.1 Historic Power Demand Growth for Frontier Training Runs

Power demand[2] for training frontier AI models has been increasing exponentially, growing by a factor of around 2.1x (90% confidence interval: 2.0x to 2.3x) annually over the past 15 years.[3] (See Figure 3.)

The power required to train frontier AI models is doubling annually

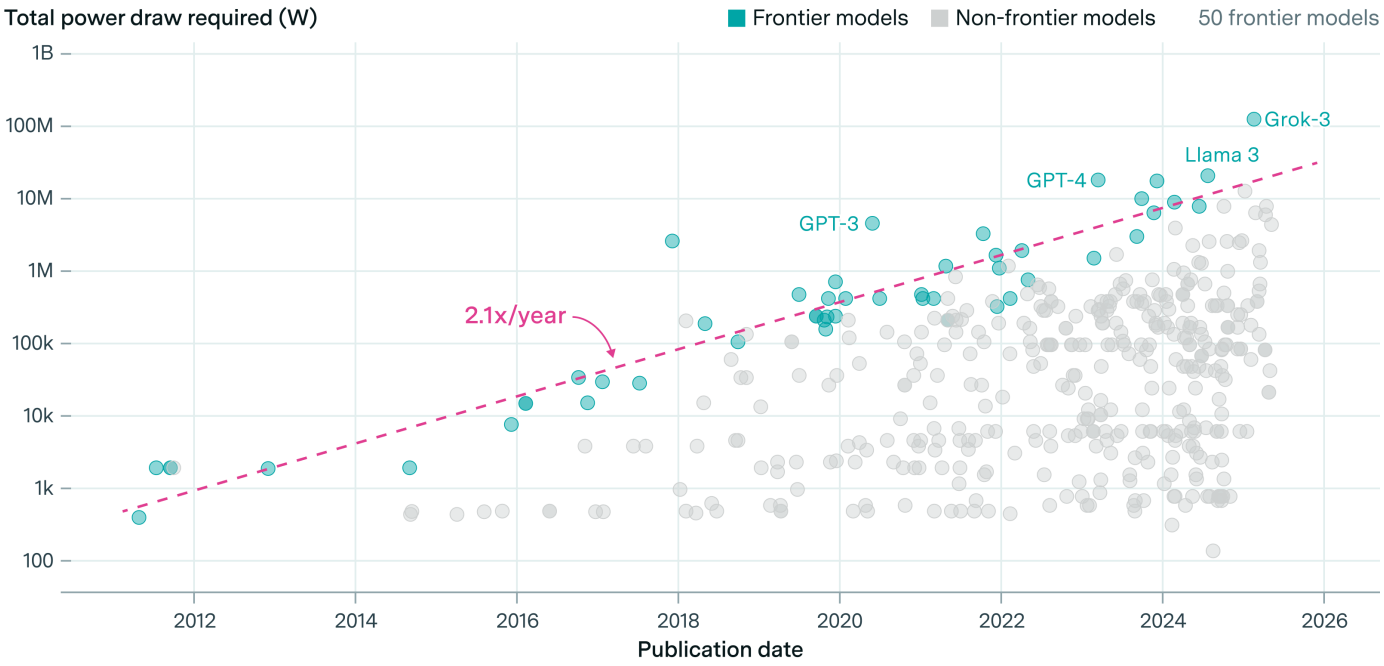


Figure 3. *Trend in the estimated power demand to train the Top 10 frontier models (at time of release) based on disclosed hardware information. This trend differs slightly from Figure 1 because it uses a more expansive definition of frontier model (Top 10 versus Top 5 in compute) and includes earlier years. Graph prepared by Epoch AI.*

[2] Power demand as discussed in this report refers to peak power demand of the data center, including IT hardware, cooling, and other infrastructure. For example, a data center may have a peak capacity of 100 MW but will likely not consume that much power 24/7.

[3] Frontier models are defined as the models trained on the largest compute scales over time. The power demand is estimated from the peak power capacity of the AI servers and data centers used to train these models. Epoch AI's full dataset of power draw for frontier models and the accompanying methodology is found here.

This result is corroborated by examining the growth rate in power demand at AI supercomputers,[4] or large clusters used for AI training, which has also doubled annually since 2019.

The fundamental driver of the growth in power demand for AI training is compute scaling. This compute is performed by AI chips[5] located in data centers and networked together in clusters. Training clusters have grown rapidly in computing performance and energy density over time to support the scale-up in frontier training runs. This scale-up of clusters has outpaced hardware and algorithmic efficiency gains, driving the growth in power demand for training.

To date, the most power-intensive training runs for released AI models were for xAI's Grok 3 and Grok 4 in 2025.[6] These models were trained using at least 100,000 Nvidia H100 graphical processing units (GPUs) in xAI's "Colossus" data center in Memphis, Tennessee, which had a peak power capacity of around 150 MW.[7] This power draw is at least seven times higher than a previous-generation frontier model, OpenAI's GPT-4, which was trained in early 2023 on a cluster drawing 22 MW. Although power demands for these clusters are expected to grow rapidly in the future, they are already significant at a local level. For example, xAI's training cluster drew 5% of the peak power demand of the local utility in Memphis.

1.2 Compute Scaling

Training compute for frontier AI models has grown rapidly at a relatively consistent rate of 4x to 5x per year since 2018, as shown in Figure 4. This growth rate is consistent across different model datasets. It holds for both a larger dataset of hundreds of groundbreaking or influential AI systems,[8] and for a subset that includes just the leading frontier models in terms of training compute.

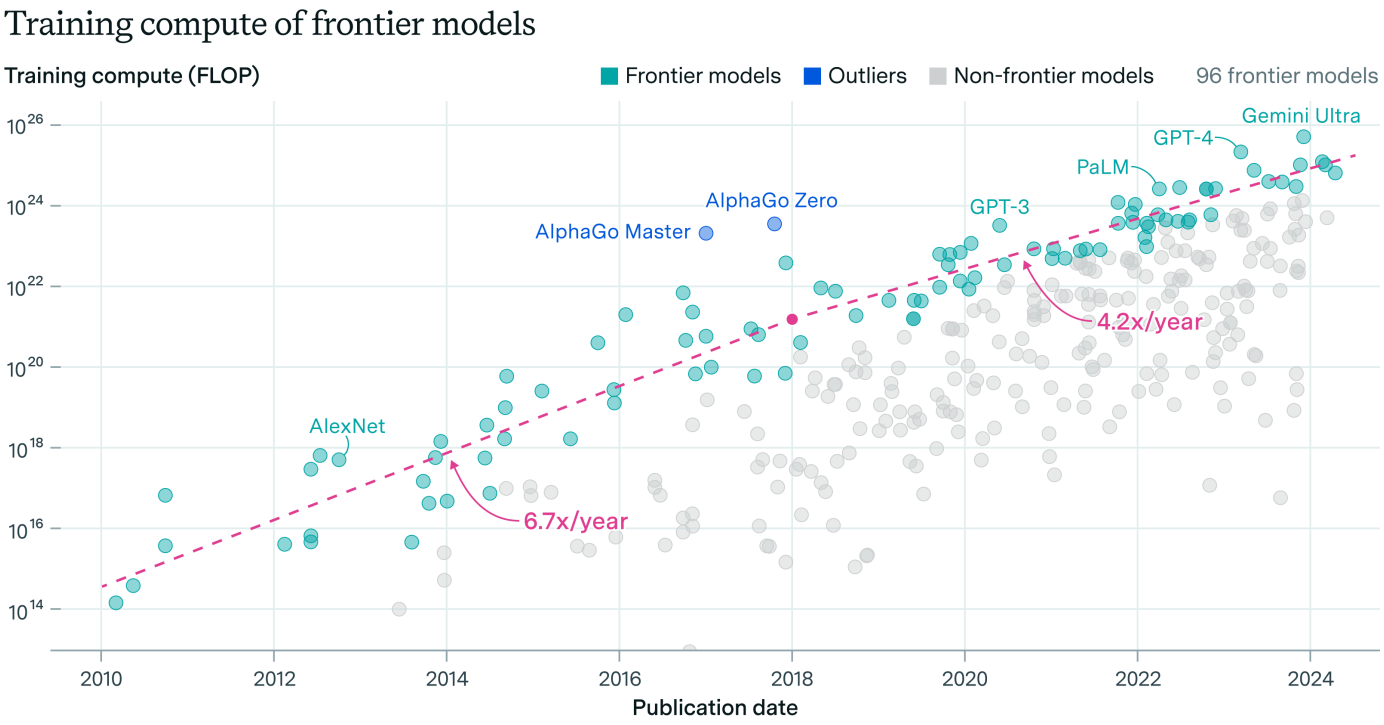


Figure 4. *Trends in the growth rate of training compute of the Top 10 frontier models, which have grown at a rate of 4.2x/year since 2018. Prior to 2018 these models grew at 6.7x/year. Graph prepared by Epoch AI.*

4 Epoch AI maintains data on over 500 large AI training clusters located around the world.

5 Chips perform the computations needed in AI. AI training requires specialized chips; these are also called "AI accelerators," "Graphical Processing Units (GPUs)," or simply "AI chips" or "AI hardware."

6 Not all models have enough disclosed training information to estimate power consumption, though it is unlikely that any model released to date required substantially more power to train than Grok 3 or 4.

7 Colossus has since doubled in size to 200,000 GPUs and is expected to draw 300 MW from the grid by Fall 2025, relying on supplemental gas generators in the meantime. It is not clear how many GPUs were online when Grok 4 was trained.

8 More information about this database can be found here.

This scaling, which is driving power demand, is motivated by the fact that researchers have discovered that AI models perform better as training compute increases, a phenomenon dubbed "scaling laws" due to a remarkable stability over many orders of magnitude of scale.

"The available evidence suggests that training compute scaling will likely continue in the near term, and it would be premature to predict a major shift in the compute growth trend."

1.2.1. Historic compute scaling

Since 2020, training compute growth has been led by *large language models* (LLMs), which can generate text or code and now drive widely used applications such as ChatGPT. Today's leading LLMs have continued the long-term growth trend in training compute. xAI's Grok-3, released in February 2025, was trained using roughly 4*10E26 floating point operations (FLOP)—a 20x increase in compute compared to the record-breaking GPT-4 in March 2023, for a growth rate of just over 4x per year.[9] OpenAI also recently released GPT-4.5, representing a "new order of magnitude of compute" compared to previous models, suggesting a similar compute scale as Grok-3.[10]

1.2.2 Future Compute Scaling

The rate of future compute scaling growth is difficult to predict. There are signs pointing to the continuance of such growth; but, at the same time, other factors could lead to a slowing of compute scaling.

The compute scaling trend may not hold—in other words, it may slow—for two broad reasons:

- **Escalating costs** may force scaling to slow or end. Maintaining a growth rate of 4x/year in compute scaling has been accompanied by escalating costs. Cottier et al. (2024) found that the cost to train frontier models has been growing somewhat slower than compute, by 2.4x/year.[11] Still, the upfront cost of the largest training clusters is already in the billions, with xAI's Memphis cluster costing an estimated $7 billion, so continued scaling at this pace would imply that individual training clusters could cost hundreds of billions by 2030.
  - Total AI investment across the industry is now hundreds of billions per year, and there are multiple plans (Appendix A) for data centers that are much more powerful than xAI's Memphis cluster, so rapid compute scaling is likely to continue in the short term. For example, OpenAI is planning to construct a 1.2-GW data center in Abilene, Texas by 2026 containing 400,000 of Nvidia's next-generation Blackwell GPUs.
  - However, beyond a certain point, scaling might not continue to pay off in terms of performance gains and potential AI applications revenue to justify further cost increases.[12] This is a key uncertainty about the future of AI, and fully resolving it is outside the scope of this report.
- **Efficiency improvements** could slow scaling if there is a more fundamental ceiling on AI capabilities or on the compute scales that enhance capabilities: if the benefits of compute scaling reach a plateau at some point, algorithmic innovations could bring this point closer in time. However, there is little evidence of a plateau so far.
  - The AI developer DeepSeek gained attention in early 2025 for training models using around 10% as much compute as similarly capable LLMs from the United States,[13] sparking debate on whether improved compute efficiency (meaning better performance for the same amount of compute) might disrupt scaling.

[9] See Epoch AI's database on AI models for more details on these training compute estimates.

[10] This suggests a roughly 10x scale-up from GPT-4, or around 2*10^26 FLOP, but this is uncertain due to a lack of public training details on GPT-4.5.

[11] This growth rate is similar for both the amortized cost of compute to do a training run and the upfront cost of the training cluster.

[12] Besides monetary costs, compute scaling may run up against hard resource constraints. Sevilla et al. (2024) estimated bottlenecks in chip production, power supply, and training data, finding that it would be feasible to maintain a 4x compute scaling growth through 2030, though these constraints may be binding shortly afterwards.

[13] DeepSeek-V3 matches or surpasses Meta's Llama 3 405B on benchmarks but was trained on around 10% as much training compute, requiring around 3 MW of power (more details here).

Factors that could lead to the continued growth of compute scaling include the following:

- **Compute growth has been relatively stable** historically, at 4x/year since 2018 despite shifts in model structures and rapid algorithmic improvements. Before 2018, compute grew even faster.
- **Model accuracy** improves with increased training compute as shown by scaling laws mentioned earlier.
- **Standardized benchmarks** such as speed and reasoning ability tend to improve with scaling (Owen, 2023).
- **Dramatic advances** have been seen with scaling. Qualitatively, scaling has occurred alongside dramatic advances, from crude early language models like GPT-2 to today's models that can handle a growing set of practically-useful tasks.
- **Improved efficiency** could actually motivate further scaling. Although mentioned above as a possible driver that could slow the growth rate, improved efficiency could conversely unlock higher capability levels, which could motivate further scaling.[14]
  - Compute efficiency gains have a long history of coexisting with total compute growth. Ho et al. (2024) estimated that over the past decade, algorithmic innovations have improved training efficiency for language models by approximately 3x annually, meaning that similar results can be achieved with 1/3 as much compute as the previous year.
  - DeepSeek may have achieved an acceleration in this trend, but efficiency gains have not slowed compute growth in the past. In terms of industry reaction, every U.S. hyperscaler announced substantial increases in AI investments for 2025 *after* DeepSeek's well-received R1 launch in January.

1.2.3. Potential Impact of Reasoning Models on Training Compute Scaling

Until recently, most compute used to train language models involved training on vast amounts of human-written data, a process called "pretraining." However, the emergence of *reasoning models*—models trained to "think"—has fueled speculation that the growth of training compute scaling may slow. This is because reasoning models benefit more from scaling up inference compute versus training compute, suggesting inference scaling could at least partially *replace* training scaling as the new paradigm of scaling.

However, the available evidence suggests that training compute scaling will likely continue in the near term despite the increase in reasoning models; therefore, predicting a major shift in the compute growth trend due to the emergence of reasoning models is premature. This holds true for several reasons:

- Despite recent discussion of pretraining "hitting a wall,"[15] previous work has estimated that there is enough text data to sustain multiple years of scaling at the current trend (Villalobos 2024). There are also large quantities of data that might be used for pretraining in other modalities such as video, images, and audio.
- Reasoning models themselves require training, and this training improves with scale.[16] So even if pretraining scaling slowed, overall training compute growth would continue due to the growth of reasoning model training.

"4x/year growth is very rapid and must slow down eventually, almost certainly by the 2030s. The key uncertainty is when, not if, this growth will slow down"

[14] For more discussion on how compute efficiency affects scaling, see this commentary from Epoch AI researcher Matthew Barnett on why algorithmic progress may spur more spending on compute.

[15] See this link for a summary of these reports. Another report notes that Ilya Sustkever (a cofounder of OpenAI and founder of Safe Superintelligence), claimed that gains from scaling pretraining have plateaued.

[16] Per OpenAI, reasoning performance "consistently improves with more reinforcement learning (train-time compute)."

- Inference scaling resulting from the use of reasoning models would likely be accompanied by training scaling because training compute is an upfront investment in improving model performance and thus the efficiency of inference. The growth of inference costs for reasoning models should, therefore, motivate more investments into training. In a simplified model of training-inference trade-offs, Epoch AI researchers estimated that spending on inference compute and training compute should be roughly equal.
- Senior staff at leading AI companies have said that reasoning models are complementary with scaling up pretraining.[17]
- Compute growth has endured dramatic shifts in paradigms before. The era of rapid scaling dates to the "deep learning" revolution of the early 2010s,[18] which substantially predates modern LLMs. For example, many of the early models were image recognition models such as AlexNet, and language models only came to dominate the field starting around 2020.

Nonetheless, the emergence of reasoning models does increase the uncertainty over the compute scaling trend going forward.

1.2.4 Conclusion

The growth in training compute for frontier models will likely continue, at least in the short term, for three key reasons:

- Scaling has achieved relatively consistent returns to date.
- AI investment levels are sufficient for continued scaling.
- Technical innovations such as efficiency improvements and reasoning models appear unlikely to significantly slow compute scaling, though they could change the nature of that scaling (e.g. shifting from pretraining to reasoning training).

For the purpose of projecting future power growth, the analysis of Sevilla and Roldán (2024) appears most relevant for the largest scale training runs. They determined an annual growth rate of 4.2x for the training compute of frontier models since 2018, with a 90% confidence interval of between 3.6x and 4.9x per year. This forecast is increasingly uncertain over time, even when using the 90% confidence interval, since this confidence interval is based on historic data. As training compute scaling continues, the high monetary and resource costs of sustaining this growth will rapidly escalate, and there will be more time for technical factors like the returns to scaling to change the picture. Consequently, extrapolating the current growth rate to 2030 should be viewed as a simplification that sets aside this uncertainty.

The 4x/year growth is very rapid and must slow eventually, almost certainly by the 2030s, absent a drastically transformed economy. The key uncertainty is when, not if, this growth will slow down.[19] While it would be feasible to scale training compute through 2030, we cannot rule out a slowdown before then due to technical innovations, data constraints, or diminishing returns to scaling.

For modeling trajectories for future power growth, the existing compute growth trend is a best-guess forecast, but this is a key uncertainty to watch.

1.3 Hardware Efficiency and Training Duration

While training compute has been quadrupling every year, power demand has grown slower than this, at a rate closer to 2x every year, because of increases in **training duration** and improvements in **hardware efficiency** (see Figure 5).

If training compute is held constant, increasing the **duration** of a training run reduces the computing throughput that a cluster must achieve to complete the training. This means a smaller cluster and less peak power because increasing duration spreads the same energy demand across a longer period of time. If cluster scaling becomes a key bottleneck, then allowing training runs to extend longer could be an important mitigation. Training run durations have been growing longer by around 26% per year. However, there may be limits to how long training runs will realistically last.

---

17 This includes OpenAI's Noam Brown and Anthropic's CEO, Dario Amodei. Amodei claimed that companies are now rapidly scaling up training for reasoning models, potentially requiring millions of AI chips. As with inference compute, the compute infrastructure used for reasoning training may differ from pretraining clusters, and this could help enable distributed training.

18 Before then, the compute growth trend for AI or machine learning models was much more modest, at around 1.4x per year.

19 As mentioned, even with hardware efficiency improvements, the hardware costs and power demands needed to sustain this scaling are both roughly doubling each year, implying a 1000-fold increase in costs every ten years.

Illustrating the impact of trends on AI power demand

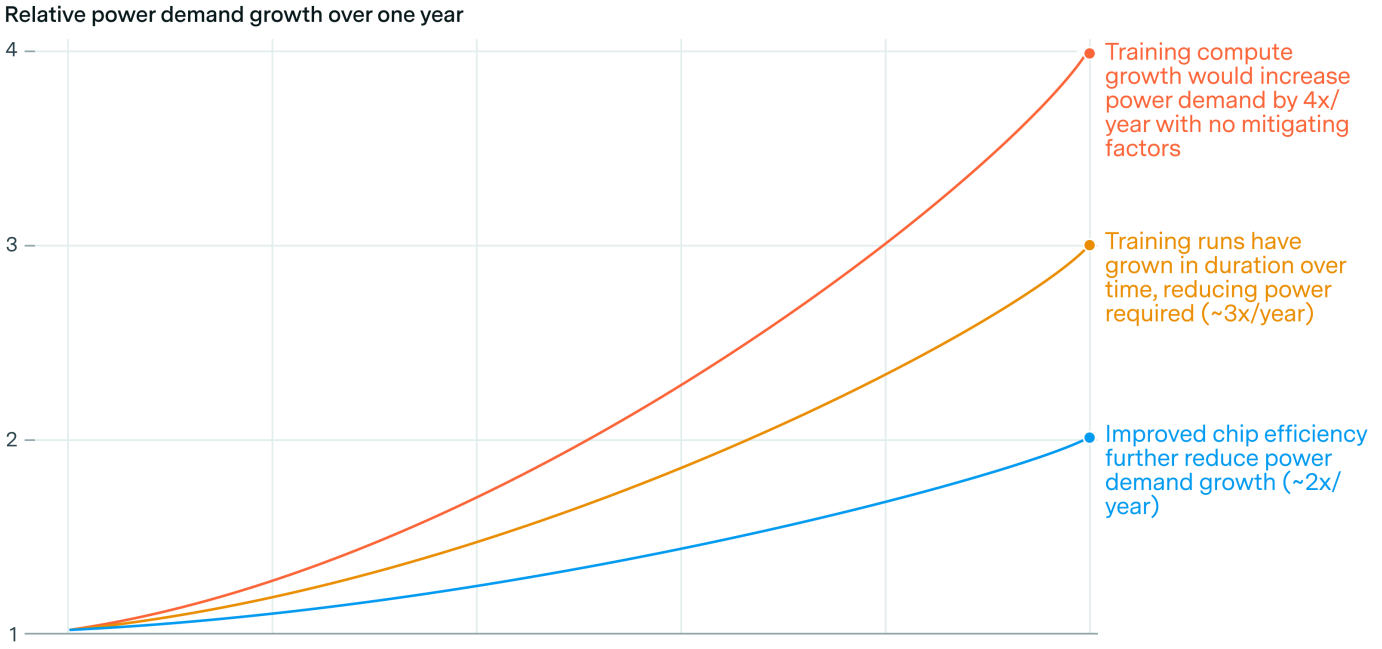


Figure 5. *Historical effect of different trends on power demand growth for AI training. 4x/year compute growth is reduced to 2x/year power growth due to increasing durations and improving chip efficiency. Graph prepared by Epoch AI.*

Improvements in hardware **energy efficiency** also reduce power demand, holding training compute constant. Energy efficiency is improving around 25–40% per year for leading AI chips. This is likely to continue, but the exact rate of improvement makes a significant difference to growth in training power demand.

Given these factors, we can build a simple model of power demand growth. The growth rate in electric power required to train frontier models is equal to the growth rate in overall training compute divided by the growth rates in training duration and energy efficiency:

Power demand = Training compute growth / (Training duration growth x Energy efficiency growth)

For example, if training compute grows at 4x per year, and frontier training runs are increasing in duration by 26% per year, and AI hardware is becoming 40% more efficient per year, then we can calculate the resulting power growth trend as follows:

***4x** per year training compute growth / **1.26x** per year training duration / **1.4x** per year efficiency growth = **2.27x** per year power growth*

This is similar to the estimated historic power growth trend, though the calculations don't align exactly because the relevant datasets are not identical.

This simple model translates training compute growth into the growth rate of power demand for frontier training runs. In addition, it can be used to help inform *forecasts* of power demand. Instead of simply extrapolating the historical growth in power demand, the trend can be deconstructed into the three factors of training compute, training duration, and efficiency growth, and evidence for how these growth rates might evolve can be considered.

LIMITS TO THIS MODEL AS A FORECASTING METHOD

The model of power demand shown here has limits as a forecast method because these factors are not necessarily independent. Improved chip efficiency and increased training run duration will reduce power demand for AI training, *if training compute is held constant*. However, changes in the growth rates of training run duration and chip efficiency could also change the trend in overall compute growth. The actual causal relationship between changes in duration/efficiency and changes in power demand is not necessarily straightforward.

For example, because duration is an input into total training compute, which is equal to the rate of computations performed multiplied by the duration of the run, a change in duration growth has a good chance of affecting compute growth as well. If duration growth slows, AI companies could compensate by scaling up their training clusters even faster, but they may instead accept a slower growth rate in total compute.

A similar concern could apply to hardware efficiency: a breakthrough in hardware energy efficiency could unlock faster compute growth instead of slowing down power demand growth. This would especially be true if AI training scale is limited by power availability, though the existing short-term plans for scaling up AI data centers provide some evidence against this.

Accordingly, the model may not accurately predict how changes in individual factors affect the overall rate of power growth due to these possible interactions. An alternative method, which is explored in the full forecast, would be to extrapolate the historic growth rate of power demand in training *clusters*, without considering the efficiency and duration factors that underlie it.

The case for this decomposition is that overall training compute growth has been quite stable over time despite many changes in the AI industry. There may also be technical factors that keep the growth rate at around 4x per year. Scaling up compute requires increasing the size of the model and training dataset, so even if a breakthrough in chip performance supports more rapid scaling, it could take time to prepare more training data and run preliminary experiments at scale.

A final limitation is that extrapolations become increasingly uncertain over time. Increasing power demand by 2x or more per year, which would imply over 1000-fold growth every decade, is clearly not sustainable in the long term. It is unclear how many gigawatts can actually be supplied to AI training runs, even if they're geographically distributed. This report does not model these constraints in detail.

1.3.1 Hardware Energy Efficiency Growth

1.3.1.1 Chip Efficiency

AI chips have become increasingly efficient over time, meaning that newer chips can produce more computational power per watt. However, these efficiency improvements have not been fast enough to keep up with compute growth, leading to increased power demand for AI training runs.

AI chip efficiency has historically improved around 26% to 40% per year, based on a variety of resources and examples:

- In a dataset of nine widely used AI chips since 2019, energy efficiency in compute power per watt has improved by 40% per year (confidence interval: 30–60%). (See Figure 6.)

"AI chip efficiency has historically improved around 26% to 40% per year… These efficiency improvements will likely continue in the coming years."

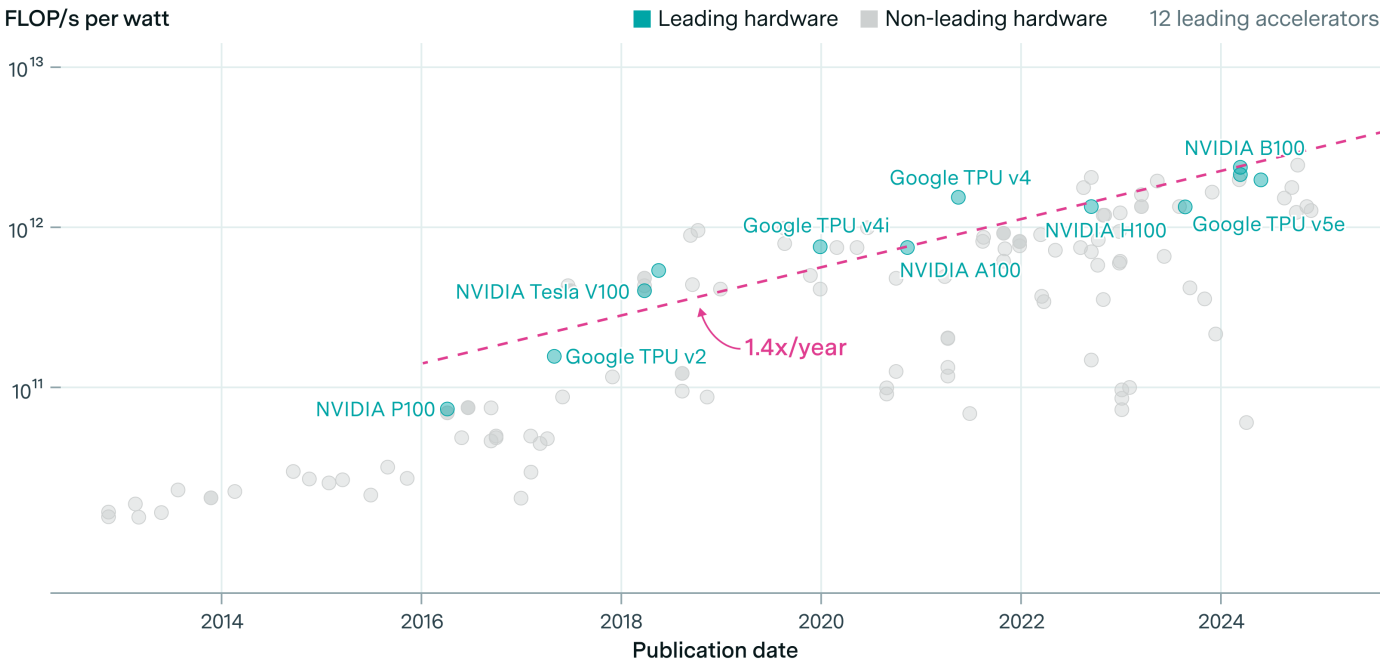


Figure 6. *Trend in the energy efficiency of the most widely used AI hardware since 2016. This efficiency has improved at 40% per year, a faster rate than GPUs as a whole. Graph prepared by Epoch AI.*

- Hobbhahn et al. (2023) found a slower growth trend of 26% (23–29%) per year using a dataset of over 47 ML GPU chips used from 2010 to 2023. Hobbhahn's 26% growth rate is based on a larger, less selective dataset than the 40% growth rate. This makes it more statistically robust—the 40% growth result has wide confidence intervals due to the small amount of data—but it may be less representative of the most popular, leading chips used in AI, which may be improving at a faster rate than the rest of the industry.
- The efficiency gains from the most recent hardware are consistent with these trends, but don't favor the higher or lower growth rate.
  - Nvidia has recently started delivering its new B200 GPU, which is around 62% more efficient than the H100, the leading chip of the previous generation.[20,21] This would be an annualized improvement of 27% per year.[22]
  - Google's latest AI chip, the Tensor Processing Unit (TPU) v6, is 67% more energy efficient than the TPUv5e, consistent with a faster rate of improvement since they were separated by just over a year.[23]
- Computing has a long history of efficiency gains. Under Koomey's law, computers have become 30–55% more energy efficient annually since 1950. The theoretical limits to microprocessor efficiency will not be binding for at least another decade (see Appendix B for more details).

Therefore, it would be reasonable to assume that efficiency continues to improve at a rate of at least 26% annually, and possibly around 40% per year.

[20] The B200 has a thermal design power of 1000 W, which is a 42% increase over the H100's TDP of 700 W, but its performance in FLOP/second is 127% higher, leading to a 62% improvement in energy efficiency. Source: SemiAnalysis.

[21] Nvidia has said that the B200 is 25x more efficient than the H100 for specific inference workloads. However, for large-scale training runs, which are heavily optimized for performance, one should compare using peak performance.

[22] The B200 first shipped in late 2024, and H100 entered full production and began shipping in late 2022, and a 62% improvement over two years is equivalent to two consecutive 27% annual improvements.

[23] TPUv5e and TPUv6 were made generally available in November 2023 and December 2024, respectively, so the time gap is closer to one year than two.

1.3.1.2 Server and Data Center-Level Efficiency and Utilization

The results above estimate the energy efficiency of individual AI chips when they are fully utilized. However, there are several additional factors that could affect the efficiency of training clusters in practice. These factors are not modeled in the main forecast due to a lack of evidence that they will change over time, but they could merit further consideration in future work. See Appendix B for more details. The following are further issues that impact efficiency:

- **Servers**: AI chips are configured in servers, and these servers have an overhead energy cost coming from cooling, networking, and other equipment. This overhead is significant, roughly doubling power consumption versus the power rating of the main AI chips inside the server.
- **Data centers**: There is also a smaller amount of overhead at the data center level. For large AI data centers, the non-IT loads contribute another 10–30% of power demand. The projections here simply assume that this additional load does not change.[24]
- **Utilization**: AI training runs typically achieve 30–40% compute utilization, so improved utilization could reduce the number of chips and consequently the amount of power required for training.
  - **Number formats**: Training compute is measured in operations on floating-point numbers, which are computer representations of numerical values. A floating-point number can be represented in more- or less-precise formats, similar to how one can round a decimal number to a certain number of digits. Switching to lower-precision formats can lead to large jumps in compute output (more here). Most recent models were trained in 16-bit precision, but a switch to 8-bit training is likely to happen soon, with some companies, such as DeepSeek and Nvidia, having already trained models at least partially in 8-bit precision.[25]
  - Switching to 8-bit training theoretically could double hardware energy efficiency based on AI chips' specifications on paper. However, it is not clear that a full doubling will be achieved in practice because lower precision could lead to worse training performance or hardware utilization may be lower.
  - Note that lower precision training could simply increase overall training compute instead of mitigating power demand, since it would boost compute output per chip in addition to output per watt.

1.3.1.3 Estimating Hardware Efficiency Growth

As our baseline for forecasting power demand growth, we average the 26% and 40% annual growth estimates for efficiency based on historic trends to arrive at **33%** growth per year. If we assume that a switch to 8-bit precision occurs over the next five years and leads to a further doubling of efficiency on top of other hardware improvements, this change would improve overall hardware energy efficiency growth over the next five years to **52%** per year.[26]

Our extrapolative model uses 33% as the low estimate of the rate of efficiency growth and 52% as the high estimate. These are two plausible reference scenarios, not lower or upper bounds on the efficiency trend.

1.3.2 Training Run Duration Growth

AI training runs have grown longer over time. This growth allows a given training run to be completed with less hardware, reducing power draw.

AI developers have scaled compute by both expanding their training clusters and by running them for longer. In a dataset of notable[27] machine learning models, training run durations have been growing at a rate of 26% per year since 2010 (see Figure 7). Among frontier models (those that were in the Top 10 of training compute at time of release), durations have been growing at a faster rate of around 50% per year since 2018.

---

24 PUE for modern AI data centers range from 1.1 to 1.3. Source: SemiAnalysis

25 DeepSeek's technical report and a subsequent paper documented their use of "mixed-precision" (using a mix of 8-bit and 16-bit precision,)in training DeepSeek-V3. Nvidia also recently trained a model in FP8 (8-bit) format.

26 Based on 33% annual growth due to chip improvements, plus an additional doubling over five years due to lower-precision training. 1.33 * 2^(1/5) = 1.52

27 Models that were state-of-the-art, highly cited, widely used, or historically significant. See here for more details.

The length of time spent training notable models is growing

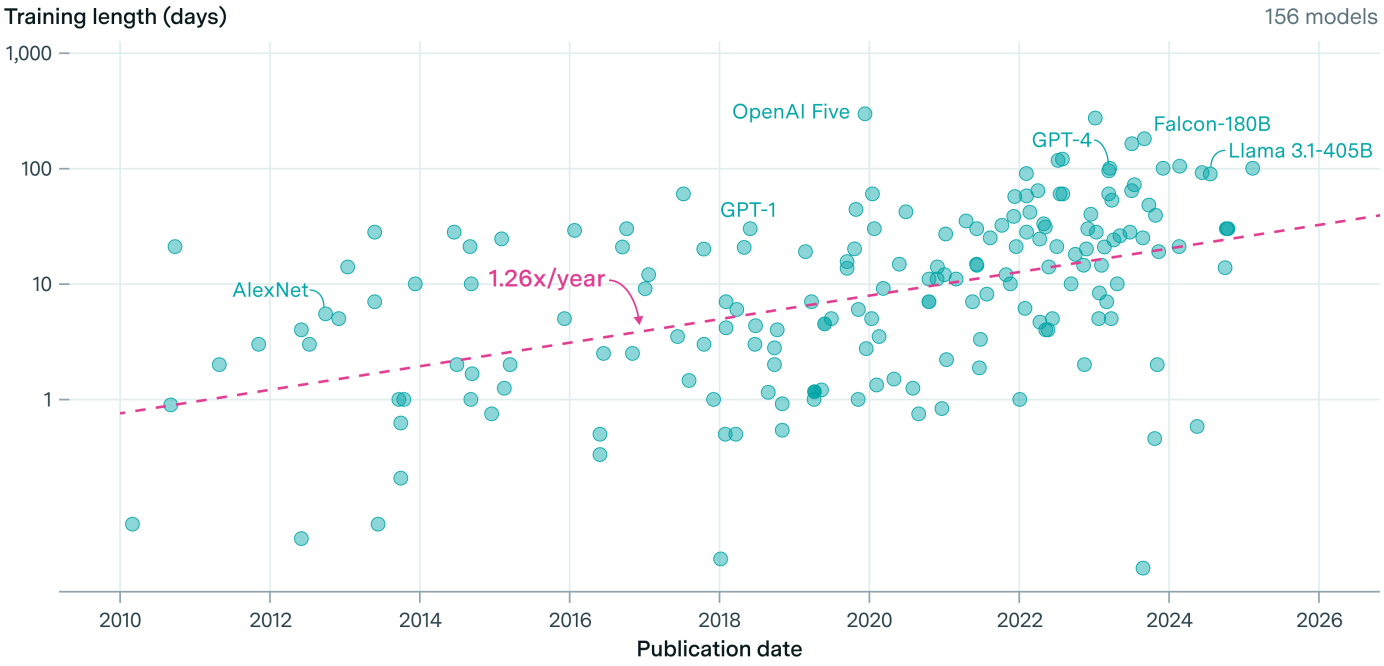


Figure 7. *Trend in the growth rate of training run durations for notable models. More information here. Graph prepared by Epoch AI.*

Many recent frontier models were trained using training runs that lasted close to 100 days. For example, GPT-4 was reportedly trained for between 90 and 100 days, Meta trained Llama 405B over 90 days, and xAI may have trained Grok-3 for around 120 days.[28] Among recent large-scale models, the longest disclosed training run was for Falcon-180B, at 180 days.

However, training durations cannot grow forever, especially at a rate of 25–50% per year, due to the rapid pace of innovation and competitive pressures.

Sevilla et al. (2022) found that training runs should not last longer than 14–15 months, because the rate of AI hardware and algorithm improvements means that a training run that lasts longer than 15 months will be outcompeted by a training run that starts later but ends at the same time. Starting from a baseline of 100-day training run durations, a 26% annual growth rate will lead to 400-day training runs in six years, approaching 14 months.

However, this is just a theoretical upper limit, and there are strong incentives for even shorter training runs so developers stay ahead of their competitors. AI companies such as Meta and Google reportedly feel great urgency to not fall behind OpenAI, and have released new models every couple of months to keep up.[29] The extrapolative model used here uses growth rates of **10% and 20% per year** over the next five years as reference scenarios. A 15% growth rate over five years would lead to training runs of 200 days by 2030.

As noted previously, training duration and training compute are not causally independent. At the high end of our power growth forecast, a reduction in training duration growth to 10% fails to reduce overall training compute growth and causes an uptick in the annual rate of power growth. But it is debatable whether maintaining 4x/year compute growth in this scenario is a likely outcome. In the other scenario where durations grow at 20% per year, this slowdown in duration growth is less likely to put the 4x growth rate in doubt.[30]

[28] xAI's Memphis cluster was online by September 2, 2023 and Grok-3 finished pretraining in early January.

[29] These include but are not limited to: Claude 3 (March '24), GPT-4o (May '24), Claude 3.5 (June '24), o1-preview (Sept. '24), o1 (Dec, '24), DeepSeek-R1 (Jan. '25), and Grok-3 and GPT-4.5 (Feb. '25).

[30] Among all of the notable models in the Epoch AI data, training compute has grown at 4.6x per year, slightly faster than the growth rate of the largest training runs, while training run durations have grown at 1.26x per year, as previously noted.

1.4 Forecasting Growth in Frontier Model Training Power Demand Through 2030

Consolidating the analysis herein of training compute growth, hardware efficiency improvements, and training duration growth generates future scenarios for frontier model power demand.

Figure 8 illustrates this forecast through 2030, projecting forward from the historic trend in the peak power draw of the largest training runs.[31] The forecast shows that power demand for the largest training runs will likely grow at a rate between **2.2x** and **2.9x** per year from a baseline of around 100 MW in early 2025, which would imply 1–2 GW by 2028, and 4–16 GW by 2030 (though growth rates in the higher end of the range are less likely to be achievable).

This figure highlights projections from two methods:

Training compute-based forecast: Forecasted power demand growth is based on the training compute of AI models, as explained in Section 1.3. The forecast assumes training compute growth continues at a rate of 4.2x/year, with a confidence interval of 3.6x to 4.9x, hardware efficiency improves at a rate of between 33% and 52% per year (combining projected hardware improvements with gains from a shift to lower numeric precision), and training run durations increase by between 10% and 20% annually. Next, a Monte Carlo forecast is used to randomly sample from the uncertainty intervals of these three growth rates, converting them into an overall forecast. This produces a mainline estimate of **2.6x** growth per year, and an uncertainty range (10th to 90th percentiles) of between 2.2x and 2.9x.[32]

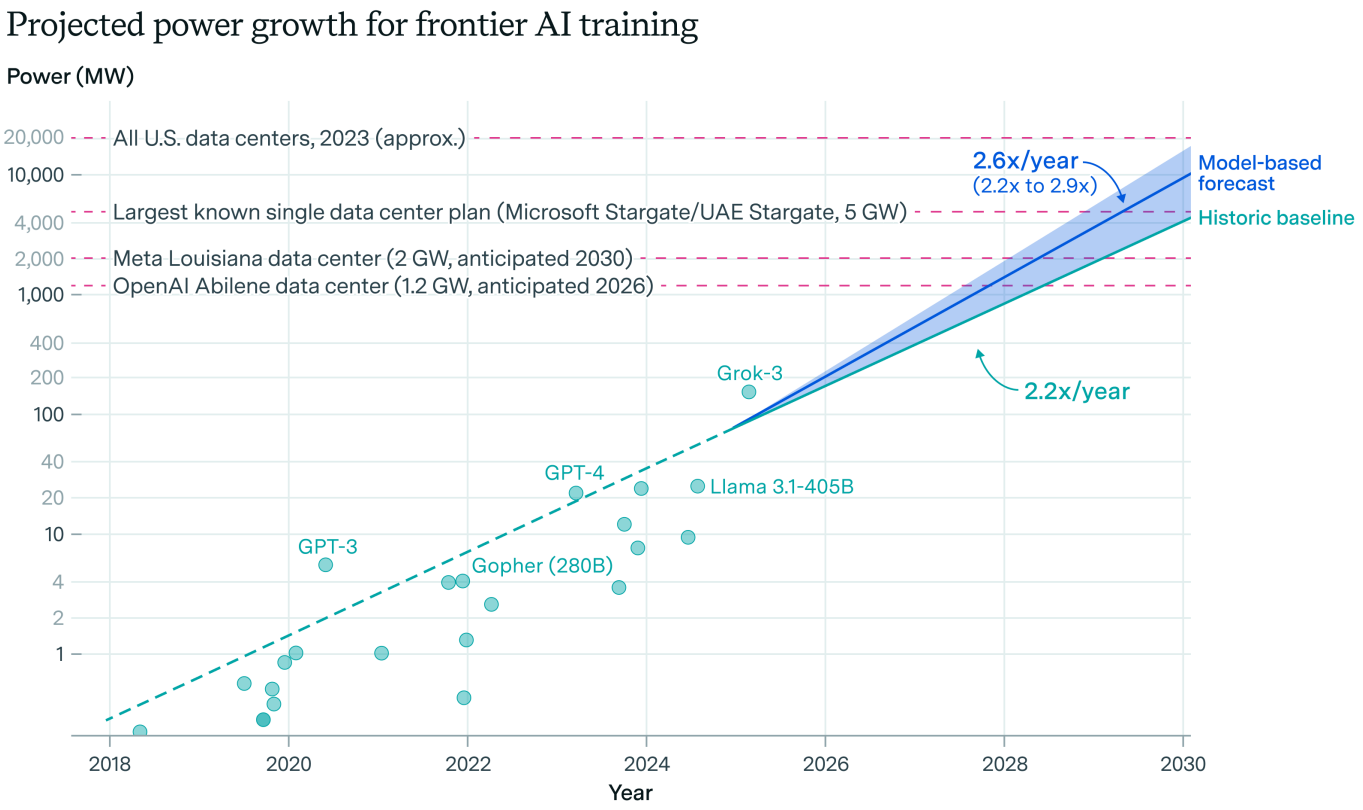


Figure 8. *Forecast for the peak power demand required to train the largest frontier models, with historic frontier AI power growth and historic training runs highlighted for context. Shaded interval represents the 10th and 90th percentiles of a Monte Carlo estimate of growth rates, given uncertainty over the hardware efficiency and training duration trends. Historic baseline represents a more conservative forecast if compute scaling continues but compute growth slows due to duration limits. Dashed red lines provide context for the projected power growth, showing plans for large-scale AI training infrastructure and other large power draws. Graph prepared by Epoch AI.*

[31] The trendline is a regression of the running Top 5 AI models by training compute for which there is sufficient data to estimate power draw but adjusted upwards to simulate the trend in the top training runs over time. See Appendix C for more details.

[32] See Appendix C for more details.

The main reason this forecast generally predicts higher growth than the historic trend is our assumption that growth in the durations of training runs will slow. One major uncertainty with this forecast is that changes in duration growth trends could have a knock-on effect of slowing down compute growth.

Historic baseline: The lower line is a projection of the historic baseline. This forecast abstracts away from questions about the trends that underpin training compute, instead predicting that the general motivation to pursue scaling will lead to a continuation of the trend observed since 2018 in the scaling of training clusters. This baseline forecast is also roughly consistent with estimates of power growth based on trends in AI supercomputers (See Appendix D).

To provide context, we also show in Figure 8 the scale of known plans for very large AI data centers. Appendix A provides a more extensive list of planned large-scale AI data centers. The prediction of multi-gigawatt-scale training runs by 2030 appears credible, given these plans.

1.4.1 Geographically Distributed Training Could Mitigate Local Power Constraints

Historically, training runs have been located in individual data centers, and the plans of several companies to construct very large individual data center campuses of 1 GW or more (see Appendix A) supports the projected compute growth. However, a shift to geographically distributed training could affect the growth of very large-scale localized power demands for training runs.

Training an AI model across large distances is technically complicated, but it appears to be feasible. An AI model needs to be updated frequently throughout training, so distributed training would require frequent communication between data centers to synchronize a shared set of model weights. In principle, network latencies are low enough that it is possible to synchronize between data centers that are thousands of miles apart.

Large-scale multi-data center training is already practiced today. Google DeepMind trained its Gemini models across multiple data centers in separate metro areas. However, it is not clear how far apart these metro areas were.[33] Google has reportedly also built distributed training clusters spread across distances of 15 to 50 miles. The distances between data centers in a distributed cluster would affect which infrastructure bottlenecks can be bypassed (site-level grid connection, utility, grid region, etc.). There are also multiple efforts to train AI models using highly distributed computers located around the world, but at a much smaller scale than frontier models.

Since distributed training appears to be possible, this raises the question of when it will become the norm and why there are any plans for very large individual data centers. This could be because the sites with adequate grid connections and infrastructure have not yet been fully exhausted. Distributed training also involves technical and engineering challenges: latency becomes more significant as distances between data centers grow, which reduces training efficiency, and the data centers need to be connected with extensive fiber optic buildouts.

Geographic distribution is a key uncertainty in understanding AI's local power impacts. It could also be necessary for continued scaling of very large training runs. Acquiring more than a few gigawatts in one site may be extremely difficult, but this is more feasible if spread across a larger area. But the high and growing cost of the AI chips themselves means that distributed training by itself is not enough to enable indefinite scaling.

1.4.2 Understanding Limitations of This Forecast

There are several limitations of this forecast, including:

- **The use of historic data on *published* AI models to establish the current trend.** Because the start dates of training runs are generally not public information, the data lag the *beginning* of current training runs by at least several months.
- **Future uncertainty intervals based on uncertainty observed from historic data.** The uncertainty interval shown in Figure 8 comes from the uncertainty on the overall growth rate among frontier training runs. It is possible that some training runs will lie outside the

[33] From the Gemini 1.5 technical report (Georgiev et al.): "Gemini 1.5 models are trained on multiple 4096-chip pods of Google's TPUv4 accelerators, distributed across multiple datacenters." In a recent podcast, Gemini lead Jeff Dean clarified that this training run was distributed across multiple metro areas.

uncertainty range shown on the figure. But if rapid exponential growth in power demand continues, the shape of the trendline will be more important than the variation in individual data points.

- **Not explicitly accounting for supply constraints on delivering power.** Multiple AI supercomputers larger than 1 GW are planned for the coming years. It is not clear if 10 GW training runs are feasible by 2030 if AI companies wish to pursue them, even with geographically distributed training.

Forecasts based on extrapolating trends are inherently less certain over time, even given our analysis demonstrating their durability. All scenarios imply extreme high power demands by 2030; the highest end projection for a single training cluster is over half the average power consumption of all U.S. data centers in 2023, and over 1% of total U.S. power generation capacity.[34]

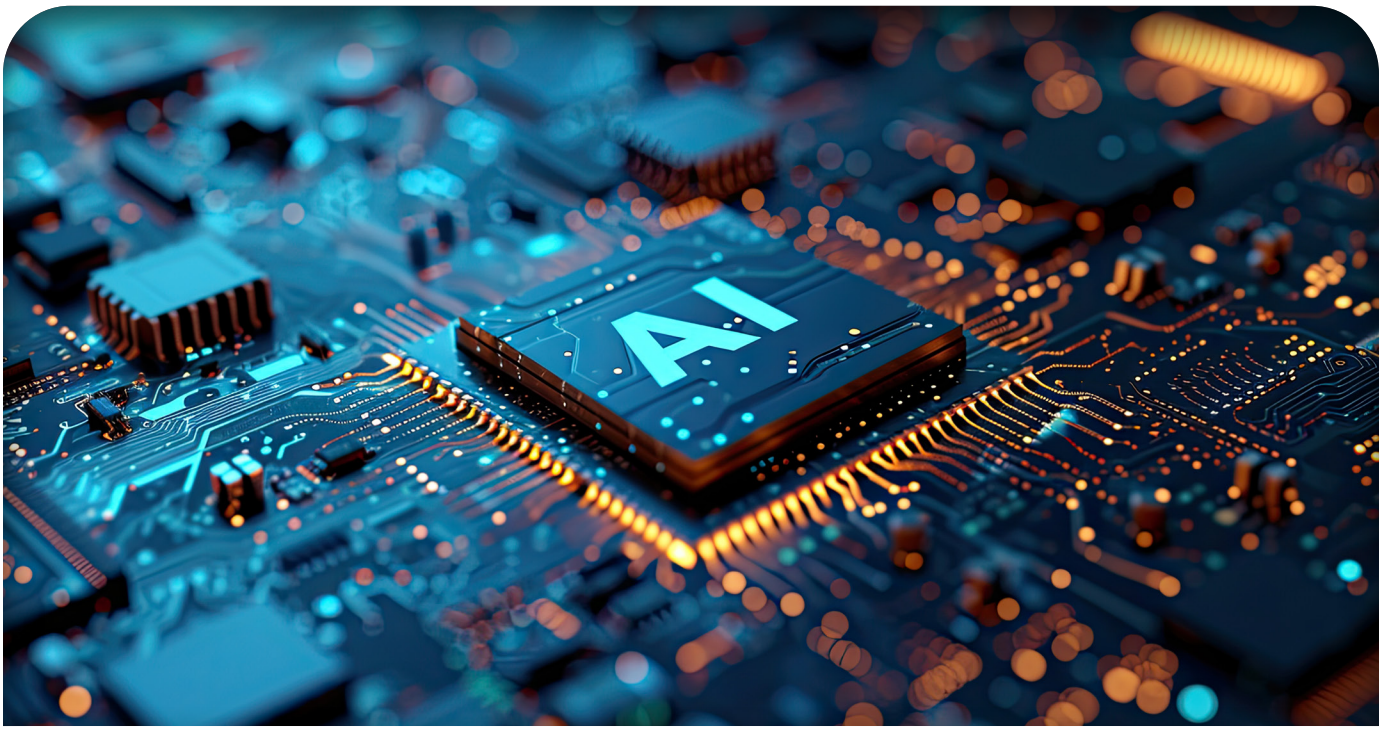
1.4.3 Implications of this Forecast for Training Data Center Power Demand

The forecast above is for power demand for individual AI *training runs*. Here we review the implications from the perspective of the power demand for AI training data centers, which is of special interest to the energy sector due to their large, localized power requirements.

There is not a one-to-one relationship between AI data centers and AI training runs. Frontier training is already motivating large, dedicated AI data centers such as xAI's Colossus. But a training run may not occupy the entire data center where it takes place, and companies may end up building oversized data centers due to the option value. For example, Amazon and Anthropic's planned 2 GW data center campus will reportedly be able to support a single, massive, training run, but Amazon may use it for other workloads such as training smaller models or inference if training runs at this scale prove unnecessary or as hardware at the campus becomes dated.

In addition, while a given training run only lasts for a certain number of months, the power demand from large AI compute clusters will be relatively continuous. Training clusters can be repurposed for other workloads, and they are not likely to sit idle due to their high capital costs.

In terms of overall power demand from AI training, the analysis so far has focused on the power demand of the largest *individual* training runs. However,there may be multiple frontier training runs of a similar scale at different facilities in a given year, in addition to demand from smaller training runs. Section 2 broadens the analysis to consider overall demand from AI data centers.



[34] U.S. data centers consumed 176 terawatt-hours (TWh) in 2023, which is around 20 GW of average consumption over one year. Note that the projections of training power draw are of power capacity, not average consumption. Total U.S. generation capacity is around 1,300 GW.

2. TRENDS IN TOTAL AI POWER DEMAND

Having looked at trends and projections in the power demand of *individual* training runs, Section 2 turns to discussing power trends in the overall U.S. AI industry in order to examine AI's overall power demand (which is a subset of overall data center power demand) and to put predictions of large-scale individual training runs into context. In addition, in the event that AI training runs become highly decentralized geographically, the scale of individual training clusters will be less significant, leaving overall power demand as the key question.

2.1 Current Level of AI Power Demand

The current state and growth trend in overall AI power demand can be estimated by using data on operating and planned AI data centers or, alternatively, using historic projected future shipments of computing hardware.

Based on an analysis of *industry data*, RAND (2025) estimates that that the accelerated servers capable of **AI workloads made up 15% of electricity consumed by data centers globally in 2024**. The International Energy Agency, which used *chip shipment data*, arrived at a similar finding—that AI data centers made up 15% of data center energy consumption globally. RAND went on to say that total data center *consumption* was 415 TWh, which represents an average hourly consumption of 47 GW; 15% of this—which represents AI workloads—is therefore **7 GW**, which implies and IT demand of around 10 GW (assuming 70% utilization). RAND separately estimates total power *capacity*, at 10.7 GW as of the beginning of 2025.

Epoch AI also has made an estimate of AI capacity based on the total quantity of AI chips.[35] As of mid-2024, the total stocks of Nvidia GPUs and Google TPUs, which make up the vast majority of all AI chips, were estimated to provide the equivalent of around 4 million H100ss in computing power, which draw about 1500 W of capacity each.[36] This suggests a corresponding AI data center power capacity of around 6 GW for the limited time frame of chip shipments tracked, 2022 to mid-2024. The total installed AI computing power has doubled about every 10 months in recent years, so a simple extrapolation of the 6 GW figure to 2025 would suggest around 9 GW, fairly close to RAND's 10.7 GW estimate.

Importantly, both of these estimates are based on global data. The exact proportion that is located in the United States is not clear, but there are several indicators that the United States contains the plurality of AI power capacity if not the majority:

- Pilz et al. collected a dataset on AI supercomputers (i.e., public information on large AI clusters and data centers) finding that the United States contained 75% of the total, weighted by compute performance.[37]
- The IEA estimated that 45% of all data center energy consumption worldwide was in the United States.[38] The share of AI data centers in the United States is likely even higher, since most of the leading AI developers and hyperscalers are headquartered in the country,
- The United States has enacted multiple rounds of export controls to restrict China's access to leading AI chips.

For the purposes of forecasting for this report, a global estimate of 10 GW of current AI computing capacity will be assumed for 2025 with 5 GW located in the U.S.

2.2 How Quickly will Total AI Power Capacity Needs Grow?

AI power demand growth can be estimated via several approaches including, utilizing projections of AI chip production, examining investment plans by leading AI companies, or relying on assessments by data center and power industry specialists. Assuming 10 GW of global AI power demand today, several approaches yield annual growth rates of 60–70% through 2030, resulting in over 100 GW in global AI power capacity by 2030. Assuming 50% of this capacity is located in the United States, that implies U.S. AI capacity of over 50 GW by 2030 (see Figure 9 for pathways and Table 1 for the underlying assumptions).

[35] Note that while "AI chips" are widely used for training or running LLMs and other forms of generative AI, some may also be used for other applications such as scientific computing or social media algorithms. Conversely, some AI inference runs on CPUs.

[36] This data is based on sales dating back to 2022 and includes some older, less efficient chips, which would increase the power capacity estimate. However, the large majority are H100s or similarly efficient Google TPUs, so any such effect should be small.

[37] Pilz et al. data is based on public information and includes only large AI clusters, so their coverage of overall AI compute may be limited, especially in China.

[38] IEA (2025), Energy and AI, https://www.iea.org/reports/energy-and-ai, p.14.

Forecasted total capacity of U.S. AI data centers

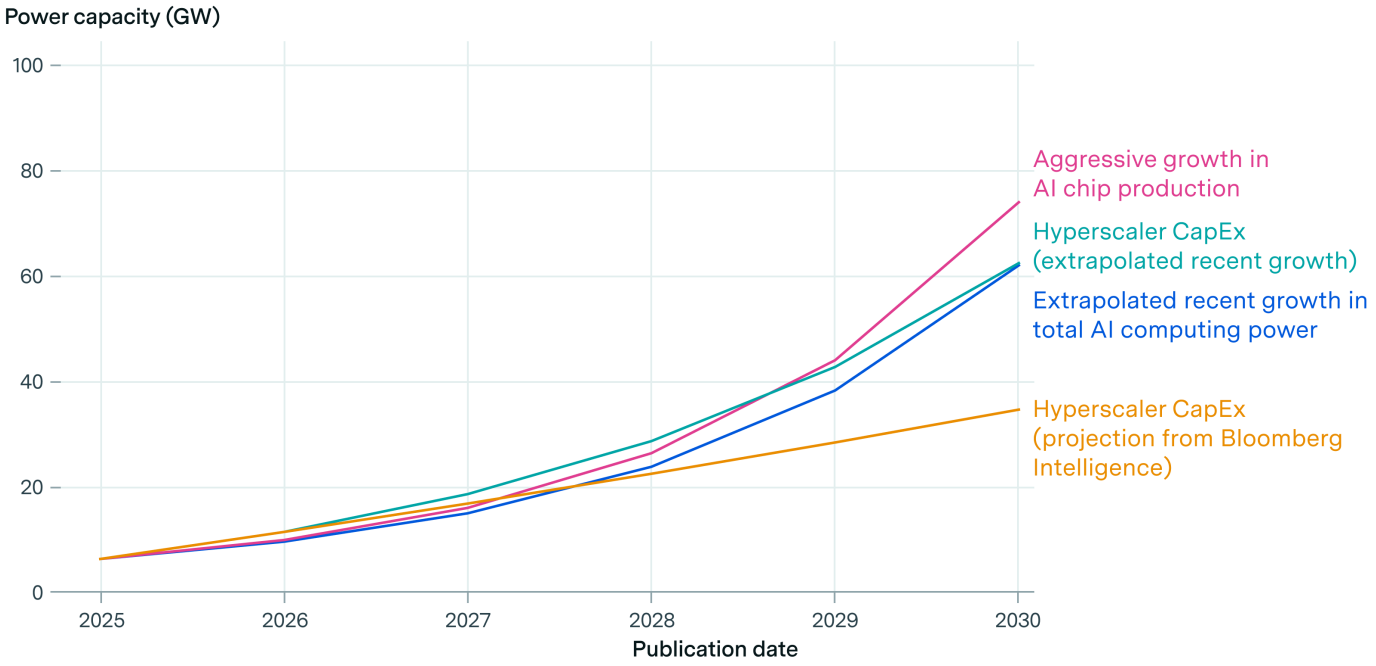


Figure 9. *Projections of growth in total U.S. AI data center capacity based on several estimates or extrapolations from current trends, assuming the United States maintains a 50% share of worldwide AI capacity. There is significant uncertainty in both the current baseline and estimated growth rates. Graph prepared by Epoch AI.*

Table 1. *Assumptions underlying AI power demand growth rates*

Alternative strategies for projecting AI data center power demand		
Approach	**Annual growth rate**	**AI capacity projection for 2030**
Recent trend in installed AI compute The total power requirements in installed Nvidia compute is growing at a rate of 2.3x per year. Adjusting by 1.4x/year efficiency improvements yields a power growth rate of around 1.64x per year.[39]	64%	120 GW worldwide 60 GW U.S.
Estimates of aggressive AI chip production scale-up Sevilla et al. 2024 estimate that annual AI chip production could grow by up to 70%/year.[40] RAND used this growth rate to project the overall growth in AI power demand.	70%	140 GW worldwide 70 GW U.S
Hyperscaler capex, assuming 2025 growth rate persists Bloomberg reports that hyperscalers will invest up to $371 billion in AI data centers in 2025, up 44% from 2024. This is corroborated by the four largest hyperscalers—Microsoft, Meta, Alphabet, and Amazon—planning over $320 billion in combined capex for 2025 (CNBC), up 40% vs. $230 billion in 2024, with most spending going towards AI.[41] It is uncertain whether U.S. hyperscalers will maintain this level of growth. Note that power draw per dollar invested in AI should stay similar over time.[42] Therefore, asf hyperscalers grow AI capex by 40%/year, total AI capacity will grow at 40%/year.[43] Actual growth may be lower due to the cost to replace retired AI chips.	40%	100 GW worldwide[44] 50 GW U.S.
Hyperscaler capex, Bloomberg Intelligence projected growth rate Bloomberg Intelligence forecasts that AI capex will grow from $371 billion in 2025 to $525 billion in 2032, or a 5% annual growth rate. This is based on proprietary estimates and is more modest than extrapolating recent growth. Assuming power per dollar stays the same, this investment can support 10 GW of capacity per year (see previous row), ramping up to ~13 GW by 2030.	5%	60 GW worldwide 30 GW U.S.

39 In practice, efficiency gains happen with each generation, not every year. Recent GPU generations have occurred every two years, so the approximation should be close.

40 Specifically, GPU production could increase by 30% to 100% per year, with 70% as a medium case. This is a model of production capacity, and actual growth will also depend on demand.

41 Microsoft will invest $80 billion in AI data centers in fiscal year 2025 (through June), Meta is planning $60–65 billion in total capex which will mostly go to AI data centers, and Alphabet and Amazon will invest $75 billion and $100 billion in total capex, respectively. Microsoft has confirmed $80 billion in AI capex; the majority of the capex from the other three will be AI-related, but the exact proportion is unclear.

42 While AI chips are becoming more energy efficient in terms of compute power per watt, they are also becoming more efficient in *compute power per dollar*. These are growing at similar rates (40% and 30% respectively), so they roughly cancel out, and power draw per dollar will stay similar over time.

43 However, this underestimates short-run growth, because $371 billion could fund up to 10 GW of additional capacity, compared to 10 GW in existing capacity.$371 billion in AI capital investment could buy around 10 GW in AI capacity: 50% of AI capex goes to chips; B200 GPUs cost $30,000 to $40,000 each and draw close to 2000 W in power. So $371 billion in AI capital investment could mean ~$200 billion going to ~5 million B200s, requiring 10 GW of power. However, total AI power capacity probably will not double in the next year: there is a time lag between 2025 investments and data centers coming online. Also, the existing 10 GW of operating capacity does not fully reflect the rapid increase in AI data center investments over the past couple of years that will soon be coming online.

44 Calculated using a baseline of 10 GW, and an additional 10 * 1.4 = 14 GW in new capacity in 2025, and so on through the end of 2029.

Another approach would be to estimate broader data center growth in the United States as a proxy for growth in AI data centers. One downside is that not all new data center capacity would go towards AI.

We review several selected estimates in Table 2. These projections, based on overall data center growth, lead to estimates of 45–90 GW of total data center capacity (AI and non-AI) in the United States by 2030. On the high end, this would more than double U.S. data center capacity, implying a roughly 50 GW increase.

The AI-focused growth rates imply that data center power demand in the United States could increase by tens of gigawatts by 2030, perhaps by roughly **50 GW**, which would be significant compared to the United States' total generation capacity of 1300 GW. Projections of overall data center growth are not much higher, suggesting that AI data centers are likely the dominant drivers of overall growth.[45]

It's important to note that these individual projections are not robust evidence about future growth, especially towards the end of the decade, since they require high growth rates in AI investments. However, it appears that, if trends continue, AI will be a significant part of the U.S. energy sector by 2030. A more conservative assumption of *linear* power capacity increases, rather than compounding growth over time, still implies multiple gigawatts of additional AI capacity per year in the United States.

2.3 How Will AI Power Capacity be Allocated?

Given an estimate of current (10 GW) and future (50 GW) power demand for the U.S. AI industry, how might this power demand be distributed across various AI tasks? Answering this question puts earlier forecasts (see Section 1.4) of the largest individual training runs in context. It also has implications for how power demand will be concentrated

Table 2. *Assumptions underlying total data center growth*

Projections of total data center growth (AI and non-AI)		
Approach	**Annual growth**	**Projection**
Recent power growth projections in U.S. data center hubs Sevilla et al. start from a baseline of 40 GW total data center capacity in the United States. Using utility company and analyst projections of 15% as aggressive scenarios for annual growth, this leads to a 50 GW increase in total capacity.	15%	90 GW total U.S. data centers, or an increase of 50 GW by 2030
Semiconductor shipment projections (December 2024) Berkeley Lab estimated that all U.S. data centers (AI and non-AI) will consume around 6.7–12% of U.S. electricity in 2028, compared to 4.4% in 2023. This corresponds to 45–90 GW of installed power capacity.	13–27%	74—132 GW in 2028
EPRI's forecasts of data center growth (May 2024) Starting from a baseline of current U.S. data centers, EPRI projected up to 15% annual growth. Data centers could consume up to 9.1% of total U.S. electricity by 2030, up from 4% today. Since this includes non-AI data centers, the growth coming from AI data centers will presumably consume under 5% of U.S. electricity.	Up to 15%	Up to ~70 GW of total U.S. data center capacity
Forecasts from the International Energy Agency (IEA) The IEA forecasts that electricity from global data centers will double by 2030, implying a growth rate of around 15% per year. They additionally estimate that demand from AI-optimized data centers could quadruple over this time period.	15% (global data centers)	60 GW worldwide 30 GW U.S.

[45] With the rapid pace of change in data centers, it is important to recognize when projections were released. For example, the EPRI projections were released before 2024 and 2025 evidence on AI uptake by customers and investments in infrastructure.

into specific locations because AI inference does not require large, networked clusters. If training becomes sufficiently geographically distributed, this would reduce the need for large, localized power demands, even if training retains a high share of overall capacity.

Moreover, a shift towards inference may have implications for how flexible AI data centers are in terms of real-time power draw. One Department of Energy report found that LLM inference could support more real-time flexibility than training, stating "LLM inference (i.e., creating responses to user requests) is amenable to real-time, geographic distribution of individual queries according to local grid load and renewables penetration, with limited negative impacts for user experience when response latency is not critical." This flexibility could reduce the burden that data centers put on electric grids and potentially unlock more data center capacity.

There are three main use cases for compute within the AI industry: *training* AI models, *inference* to deploy models, and *research experiments* to improve future models. Compute is split between multiple companies and use cases, so the largest individual training runs use a small fraction of total AI compute. The largest known training clusters, such as xAI's, contain around 100,000 to 200,000 leading AI chips, compared to the millions of AI chips in existence. Unless the structure of the AI industry changes dramatically, this means that if individual frontier training runs reach a multi-gigawatt scale, the AI industry will require at least tens of gigawatts in total capacity.

There is some public evidence suggesting that training and inference use similar amounts of compute at leading AI developers. Per one report, OpenAI's projected compute spending in 2024 consisted of $3 billion to train models, $2 billion for inference, and $1 billion on "research compute amortization." Excluding research, OpenAI's compute was split 60/40 between training and inference. Research compute is significant, but the amortization makes the actual amount unclear.

Google and Facebook also disclosed how they allocate power or energy for AI, but these reports are relatively outdated. In 2022, Patterson et al. noted that at Google, energy used for machine learning over the past three years was split 60/40 between inference and training. A 2022 paper from Facebook AI (Wu et al.) noted an AI power capacity breakdown of 10/20/70 between experiments, training, and inference (that is, inference is over 3x training). Inference demand exceeded training demand for both companies, with a much higher ratio for Facebook. However, this may be before Facebook (now Meta) began pursuing very large-scale LLMs.[46]

The bottom line is that the power demands for training and inference are unclear as the industry races towards 2030. The split will likely differ by company and how AI is integrated into their products. For some domain-specific AI systems, which are based upon relatively stable information sources, inference could dominate. For others, that reflect continually changing information, training will be an ongoing effort.

2.3.1 How Might the Training and Inference Split Evolve?

According to a theoretical analysis by Erdil (2024), the amount of compute allocated to training and inference should be roughly similar because investments in training compute can enhance inference, and vice versa. This doesn't prove that training and inference demand should be split 50/50, but the broader argument cuts against a major shift in the near term. Development and adoption of AI applications and inference-heavy innovations like reasoning models, introduced in Section 1.2.3, could cause a surge in demand for inference and shift the compute allocation towards inference. However, this effect could be mitigated by the fact that higher AI demand will motivate more investment into training to develop even more capable models or to make inference more cost-efficient.

One important factor is whether training compute scaling hits limits while usage continues to grow. In the current regime of rapid growth, training runs for next-generation models tend to be much larger than for current-generation models. If training growth slows, then demand for training will presumably shrink relative to inference compute.

The scale of frontier training runs also has implications for how AI power capacity is allocated. Realistically, there will be multiple competing large training runs (at up to 5 GW each) at any given time in addition to training demand for many smaller models.[47] Therefore, projections of individual multi-gigawatt training runs imply that training will retain a substantial share of overall AI power demand.

[46] A visualization of large-scale AI models grouped by company is available at https://epoch.ai/data/large-scale-ai-models

[47] For example, Cottier et al. forecast that total training compute for all models could be around 5x greater than training compute for the largest single training run in 2030.

3. CONCLUSION

The power required for the largest frontier AI training runs will likely grow 2.2x to 2.9x annually, potentially reaching 1–2 GW by 2028 and 4–16 GW by 2030. This demand would be highly significant, with the high end for a single model approaching 1% of total U.S. power capacity. Total power demand for AI in the United States, including training for multiple companies as well as inference, could exceed 50 GW.

It is unclear whether this demand can actually be met. While hyperscaler investments suggest rapid growth in power demand will continue in the near term, constraints in building generation capacity and transmission lines could limit this growth. Advancements in multi-data center training may help distribute power usage geographically, potentially easing these constraints.

There are also broader societal implications. If AI demand hits supply constraints for power, policies to unlock energy growth may be needed to enable continued growth, which may disrupt traditional planning processes and have environmental consequences. Power growth also poses challenges for tech companies that have already committed to using clean energy.

The analysis suggests that power demand will continue its rapid growth in the near term, likely reaching gigawatt-scale training runs by 2028. Beyond this, continued scaling is less certain. Further growth will depend on investment trends and available power capacity, with important consequences for both the U.S. energy sector, economy, and society.

RESOURCES

Glossary

Compute: Shorthand for computation. Note that in common usage "compute" can sometimes mean either a *quantity* (i.e. a total number of computations) or a *rate* (how many computations can be performed per second). This report generally uses compute to refer to compute quantities.

Floating point operation (FLOP): Compute is often measured in terms of floating-point operations (FLOP), which are basic math operations performed by computers. Floating-point is a way to represent decimal numbers in a computer. The acronym "FLOPS" is sometimes used to refer to floating-point operations *per second*, which can be a source of confusion because FLOPS can easily be misread to simply mean "floating-point operations" (in the plural). To prevent this ambiguity, this report uses "FLOP" to mean "floating-point operations,, and "FLOP/s" to mean floating-point operations per second.

AI model: The key components of AI systems; in modern AI systems, AI models are usually composed of very large neural networks. AI models are trained using large datasets to accomplish tasks such as generating text and describing or generating images, video, and audio.

Frontier model: A term commonly used to describe industry-leading AI models in terms of amount of training compute or in terms of capabilities. In Epoch AI's research, "frontier models" are sometimes provisionally defined as those that were in the Top 10 in training compute when they were released, but this is not a universal definition.

Hyperscaler: Informal term for the small number of companies that own and operate compute at very large scales, such as Microsoft, Meta, Amazon, and Google. Most AI developers access compute through a hyperscaler, but some such as xAI are acquiring large-scale compute of their own.

Large language model (LLM): Large AI models that are trained on massive amounts of language data and capable of generating natural language text or computer code, though many LLMs have vision and audio capabilities (known as "multimodality"). There is no universal threshold for "large," but most modern language models used for chatbots or other generative tasks are considered LLMs. Within the AI industry, LLMs have received the most compute investment beginning around 2020.

Thermal design power (TDP): A computer processor's power rating and technically, the amount of heat that a chip's cooling system is designed to manage. TDP is not necessarily equal to peak power consumption, which can be higher, but it is a proxy for a chip's maximum sustainable power consumption when run at high capacity (e.g. when training an AI model).

Appendix A: *Planned Large-Scale Data Centers and Training Clusters*

This appendix reviews hyperscalers' and AI developers' plans to scale up their data centers and training clusters, providing independent evidence of power scaling that supports the trends-based forecast in Sections 1 and 2 above.

Table 3 provides a select, non-comprehensive list of planned AI data centers highlighting the largest-scale concrete plans today. For more comprehensive information on both planned and existing large-scale AI data centers and clusters, see the AI supercomputers hub from Epoch AI.

These company plans suggest that AI data centers could reach or exceed 1 GW in 2026–2027 and grow to 2–5 GW by 2030. This would be consistent with a short-term acceleration from the historic trend of training power demand doubling every year with a possible slowdown after 2027. Early phases of some of these sites are operational or under construction; other sites are still in early planning.

Note that there is not a one-to-one relationship between data centers and training runs; training can be geographically distributed across multiple data centers, and a single data center can be used for multiple workloads. There is also a delay between the data center becoming active and the public release of the first model trained there.

Table 3. *Examples of Large AI Datacenters in Operation, Construction or Planning*

Frontier Model Company	Data Center	Peak Power Capacity	Timeline
xAI	**Colossus Memphis Phase 2** xAI expanded their Memphis data center to 200,000 Hopper GPUs, up from the 100,000 used to train their Grok-3 model. Phase 2 was reportedly operational with a capacity of 300 MW as of May 2025.	300 MW	Operational as of May 2025
	xAI GW-scale Memphis data center xAI is planning a separate gigawatt-scale data center, containing one million Nvidia chips.	>1 GW	Unknown
Anthropic	**Project Rainier** Amazon–Anthropic collaboration on an Indiana data center campus, initially with 400,000 Amazon Trainium 2 chips. This will eventually expand to 2.2 GW, with an unspecified timeline.	450 MW (initial phase) to 2.2 GW	2025 for initial 450 MW phase
	5 GW distributed cluster Anthropic recently predicted that by 2027, frontier training runs would require networked (distributed) training clusters drawing 5 GW, but did not provide a concrete plan for building such a cluster. This would be much faster than the historical trend.	5 GW (distributed)	2027
Meta	**Hyperion** Meta is planning a $10 billion, 2+ GW data center in Louisiana to be supported by three new gas plants. Construction is planned through 2030.	2+ GW (up to 5 GW)	1.5 GW by the end of 2027 2 GW by 2030, with unclear timeline for 5 GW
OpenAI	**Stargate Abilene (w/ Oracle)** Planned 1.2-GW data center campus in Abilene, Texas that will be part of OpenAI's broader "Stargate" project.	1.2 GW	2025: initial 200 MW phase 2026: 1.2 GW.
	UAE Stargate OpenAI announced a 1-GW cluster in the United Arab Emirates with 200 MW expected to come online by 2026. This may eventually expand to a 5 GW campus. (Note: this is the only non-US project on this list)	1-5 GW	Initial 200 MW by 2026. Unknown timeline for full 1-5 GW.
	Wisconsin (w/ Microsoft) Microsoft is planning a Wisconsin data center campus for OpenAI with a capacity of 1.5 GW by 2027, according to an industry analyst. This project will tentatively open in 2026. [In January 2025, Microsoft paused construction on the later phases of this project, perhaps temporarily.]	1.5 GW	Expansion to full scale in 2027, possibly canceled or delayed
	Stargate (w/ Oracle, SoftBank) This was a planned $100 billion OpenAI-Microsoft collaboration for a data center campus that would open in 2028 and scale up to 5 GW by 2030. Microsoft may have since pulled out of this project, and the "Stargate" name has been adopted by a $500 billion collaboration between OpenAI, Oracle, and SoftBank involving numerous data centers.	5 GW	(tentative, possibly canceled) 2028: initial phase 2030: scaled up to 5 GW

Appendix B: *Hardware Efficiency*

Long-Term Limits to Energy Efficiency

The energy efficiency of computer chips has been improving exponentially over many decades. This trend is described by Koomey's Law, which found that the energy efficiency of computing hardware consistently doubled every 1.57 years (equal to an annual growth rate of 55%) between 1950 and 2000 in a trend that closely resembles the more famous Moore's law. However, this trend has slowed to doubling every 2.6 years since 2000 (or an annual growth rate of 30%). The latter is similar to the 26% growth rate that Hobbhahn et al. found using a more expansive dataset. The length and durability of this trend means that it is very likely that computer chips will continue improving in efficiency at a similar rate over the next few years, though it is possible that these improvements will slow in the long run.

In 2023, Ho et al., investigated theoretical upper bounds for the current paradigm of microprocessors, estimating that processors could become more efficient by a factor of 50 to 1000, with a median estimate of 200. Taking the lower end of this range, a 50% annual growth rate in efficiency would take 10 years to reach this limit, and a 30% growth rate would take 15 years. So, this upper limit is not relevant for the next five years, but efficiency improvements could slow before stopping, as has happened before with the slowing of the growth rate described by Koomey's law.

Server-Level Overhead

AI chips in training clusters are configured in systems called servers, which include multiple AI chips (such as GPUs or TPUs), additional computer processors such as CPUs to coordinate the system, and cooling and networking equipment.

The server-level power overhead is significant, roughly doubling power consumption versus the power rating of individual chips. For example, while a single H100 GPU has a thermal design power of 700 W, a common server configuration with 8 H100s draws up to 10.2 kW in total, or 1275 W per GPU. Larger clusters have additional overheads, requiring around 1500 W of power per H100.[48] Whether this server-level overhead will improve with future generations of chips is not clear and would be a useful area for further research. The potential efficiency gains from reducing this overhead are bounded by a factor of 2, so even eliminating this overhead would be equivalent to two to three years of normal chip-level efficiency growth.

Data Center Overhead

Meanwhile, there is a smaller amount of overhead cost at the data center level. This is measured by power usage effectiveness (PUE), which is the ratio between a data center's total energy consumption and the energy consumption of the IT equipment within the data center. PUE has fallen over time. According to SemiAnalysis, industrywide data center PUEs fell from 2.2 to 1.55 between 2010 and 2022. Hyperscale AI data centers are even more efficient, with PUEs typically under 1.3, and Google claims an industry-leading average PUE of 1.1 for its data centers.

PUE could continue to fall, but because PUE has a floor of 1 by definition, these efficiency gains have almost been exhausted. Reduced PUE cannot reduce power consumption by more than around 9% for Google, and by more than around 20% for the less efficient hyperscalers. As such, gains from reduced PUE are not modeled in the projections in this report.

Compute Utilization

In addition to hardware's inherent efficiency at peak performance, another factor to consider is compute *utilization*, which refers to the percentage of AI hardware's theoretical computing power that is actually achieved. In practice, most large training runs tend to have 30–40% utilization rates. Improving utilization would effectively improve energy efficiency by reducing the number of chips to complete a training run. While increased compute utilization could also increase power consumption, the effect would likely be minor: one source estimates that average power consumption for AI servers is already around 70% of rated capacity. And increasing average compute utilization would likely have little or no effect on peak power consumption.

However, there is no clear trend of compute utilization during training changing over time.[49] For this reason, any effect from changes in utilization rates are not modeled in this report. In principle, continued innovation to optimize GPU clusters could improve utilization over time. This is counterbalanced by the growing engineering challenges required to scale up training runs: larger clusters and distributed clusters may make memory and communication bottlenecks more significant. Utilization is capped at 100%, so it can improve by at most a factor of around 2.5.

[48] This is consistent with the fact that xAI's data center in Memphis draws 150 MW of power for 100,000 H100 GPUs.

[49] This analysis found a not-statistically-significant upward trend over time.

Appendix C: *Frontier Power Demand Forecast Methodology*

The code used for the analysis below can be found here.

The forecast of power demand for training frontier models here has two components:. First an estimate of the historic trend in power demand for training frontier models to establish a baseline for 2025, then an estimate of the growth rate in this power demand moving forward.

Historic Baseline

To estimate the historic trend, data on AI models collected by Epoch AI were used. This presents a choice in terms of which models to include. For example, the single largest training run over time will tend to run ahead of the average trend among large-scale models in general, especially because there is limited data on the power demands of many recent frontier models. However, simply measuring the trend in the largest training runs over time (e.g. GPT-4 or Grok-3) would use far fewer datapoints and yield a less robust result.

To balance these considerations, we run a regression on power demand of running top five AI models by training compute for which there is sufficient data to estimate power draw. These models date back to 2012, though the trendline is truncated to 2018 in our main visualization.

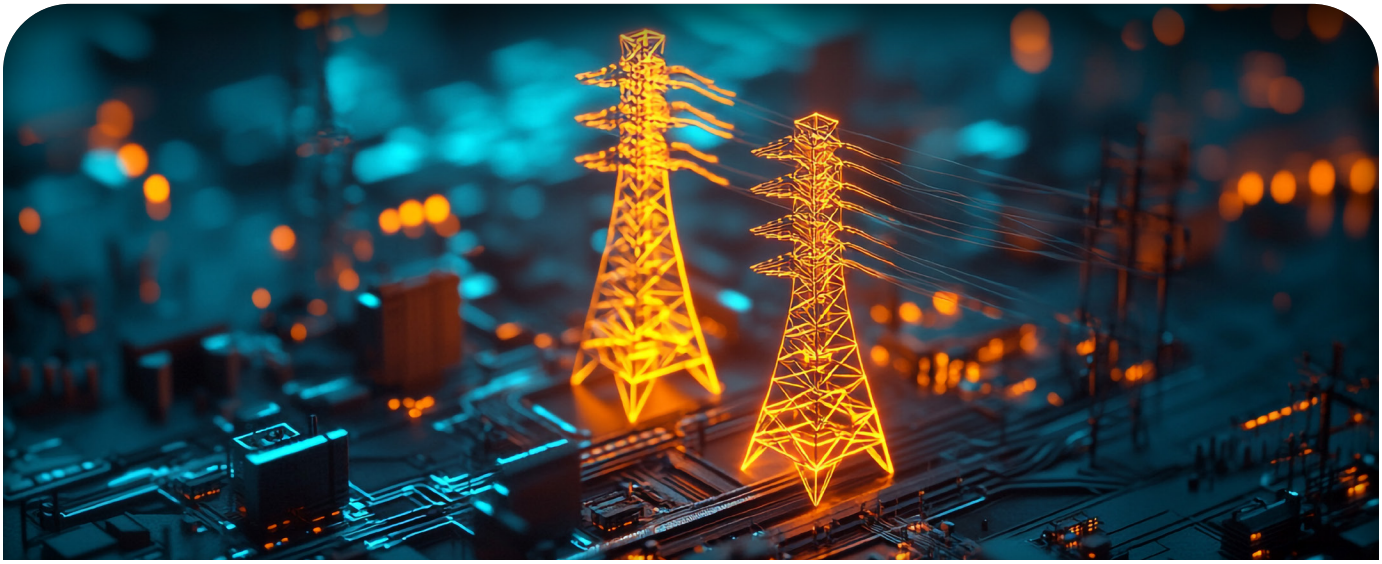
We then adjust the resulting trend upwards to simulate the trend in the most power-intensive training run over time. This was done by multiplying the trendline by the standard deviation of the regression's residuals, multiplied by the z-score for the 80th percentile (0.84). This approach uses the variation in the trend to simulate the trend in the largest individual training run over time rather than the average of the leading five models. This results in a trendline with a growth rate of 2.2x per year, and that reaches ~80 MW by the beginning of 2025, which is somewhat below the actual largest training run to date (150 MW for Grok-3).

Forecasting Power Demand Growth

The main forecast for the growth rate in power demand in Sections 1 and 2 above is based on the training compute-based decomposition explained here: growth in power demand is equal to the growth rate in training compute, divided by the growth rates in hardware efficiency and training run duration.

These growth factors are estimated and forecast separately and combined into an overall power demand forecast using a Monte Carlo simulation. This entails randomly sampling three values from our uncertainty intervals for the growth rates in training compute, hardware efficiency, and duration growth, and multiplying the results together. The median result of these samples was a growth rate of 2.6x per year, with the 10th and 90th percentiles of 2.2x and 2.9x growth per year. The Monte Carlo simulation assumes that uncertainty for all three factors is distributed log-normally, which is a standard assumption for estimating the multiplicative product of several variables.

The historic trendline is highlighted separately as a more conservative forecast that does not assume that the trends in training compute, training run duration, and hardware efficiency can vary independently over time.



Appendix D: *Power Demand Trend in AI Supercomputers*

An alternative method for estimating the power growth of frontier training runs is to directly measure the growth of large-scale AI computing clusters, also known as AI supercomputers. Pilz et al. compiled a dataset of large-scale AI supercomputers after a comprehensive review of public information.

There are two ways to measure power growth based on this data:

- The power capacity of leading AI supercomputers, measured directly using hardware specifications, has been growing at a rate of **1.95x** per year.
- Computing power of the largest AI supercomputers, measured in 16-bit floating point operations per second, has improved at a rate of 2.5x per year (90% confidence interval of 2.4x to 2.7x). Projecting this trend forward while dividing it by the growth rate of ML hardware efficiency improvements of 1.33x (not including efficiency gains from changing number formats) yields a growth rate of **1.88x** per year.

These growth rates are somewhat lower than the low end of the forecast based on frontier model training runs, which was around 2.2x per year.

In terms of projecting this growth rate forward, the same arguments for why the industry will continue scaling model training runs apply to AI supercomputers: scaling requires continued growth in the size and computing power of training clusters. However, the model-based extrapolation differs by including the possibility that AI labs may target compute scaling that is faster than they have historically expanded their AI training clusters due to constraints on training run durations.

One advantage of estimating frontier model power growth using the AI supercomputers dataset is that it measures the growth rate of supercomputer power capacity directly rather than inferring the growth rate of AI training clusters using a more complex method of dividing training compute growth by training run duration growth rates.

However, a downside is that this dataset is less focused on the specific trend—the scaling of frontier AI models—that will be an important driver of AI power demand going forward. For example, some supercomputers in this dataset were primarily used for scientific computing for research rather than to develop commercial AI models.

THE FOLLOWING ORGANIZATION(S), UNDER CONTRACT TO EPRI, PREPARED THIS REPORT:

EPOCH ARTIFICIAL INTELLIGENCE, INC.

About EPRI

Founded in 1972, EPRI is the world's preeminent independent, non-profit energy research and development organization, with offices around the world. EPRI's trusted experts collaborate with more than 450 companies in 45 countries, driving innovation to ensure the public has clean, safe, reliable, and affordable access to electricity across the globe. Together…Shaping the future of energy.®

About Epoch AI

Epoch AI is a multidisciplinary nonprofit research institute investigating the trajectory of Artificial Intelligence. Epoch AI curates data and conducts high-quality research into some of the most significant trends in AI in order to help society improve its understanding of AI and make better decisions.

PRINCIPAL INVESTIGATORS

JOSHUA YOU, *Data Analyst*
josh@epoch.ai

DAVID OWEN, *Senior Researcher*
david.owen@epoch.ai

EPRI CONTACTS

DAVID PORTER, *VP, Electrification & Sustainable Energy Strategies*
(704) 595-2700, dporter@epri.com

TOM WILSON, *Principal Technical Executive, Integrated Grid and Energy Systems*
(650) 855-7928, twilson@epri.com

For more information, contact:

EPRI Customer Assistance Center
800.313.3774 • askepri@epri.com

3002033669 **August 2025**

EPRI
3420 Hillview Avenue, Palo Alto, California 94304-1338 USA • 650.855.2121 • www.epri.com