Will AI's growth create an explosion of energy consumption?

Further improvements in hardware and software efficiencies may counteract an expected surge in demand for electricity needed to power the new large language models.

When OpenAI's ChatGPT was introduced in November 2022, it quickly captured the public's imagination, surpassing 100 million active users within two months. Since then, analysts have been issuing eye-popping projections for the growth of generative artificial intelligence (AI). Bloomberg Intelligence in March forecast a $1.3 trillion market by 2032, from a market size of $40 billion in 2022—a compounded annual growth rate of 43%.

It's been widely reported that new AI models are power hungry and that they will add to the consumption of electricity by data centers. But no consensus has emerged as to the size that jump will be.

The data-center builder Schneider Electric's 2023 white paper *The AI Disruption: Challenges and Guidance for Data Center Design* estimated that AI workloads accounted for 8% of the estimated 54 GW of electricity used by data centers last year. Schneider expects that share to rise to 15–20% by 2028, when data-center demand is forecast to reach over 90 GW. Of total AI usage, 20% is currently used for the training of models, with the rest going to inference—the individual instances with which the the model is tasked. The ratio is expected to evolve to 15:85 by 2028.

Since the introduction over the past year and a half of ChatGPT and other so-called large language models (LLMs), such as Microsoft's Copilot and Google's Bard, some researchers have issued—and news outlets have reported—predictions of skyrocketing electricity demand. One such forecaster was Alex de Vries, a data analyst and PhD candidate at the Free University of Amsterdam School of Business and Economics. He says that global electricity demand for AI could grow larger than the entire consumption of Argentina by 2027. "With AI, the whole principle is that bigger is better," de Vries says. "Bigger models are more robust and perform better. But they require more computational resources and more power."

Overblown forecasts?

In a 2023 commentary in *Joule*, de Vries pointed to estimates by Alphabet chairman John Hennessy that the cost of Google searches might increase by a factor of 10 if an LLM were used in every interaction. (See the figure.) That isn't likely to happen in the near term, he acknowledges. For starters, it would require the prompt delivery of more than 500 000 high-end servers, at a cost to Google of around $500 billion.

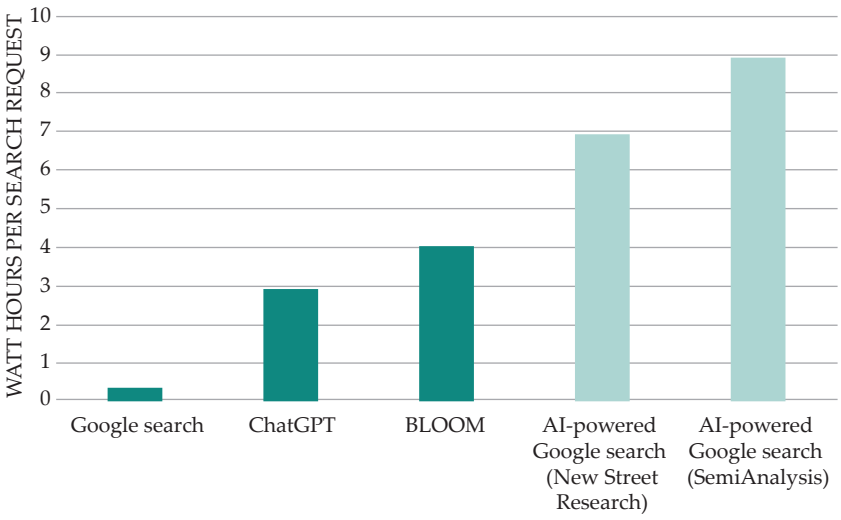
Yet past forecasts of exponential increases in electricity consumption from other transformative developments in IT have proven wildly inaccurate, says Daniel Castro, who authored a January white paper on AI energy use for the Information Technology and Innovation Foundation (ITIF) in Washington, DC. He points to a French think tank's widely reported estimate in 2019 that the carbon footprint of streaming Netflix for 30 minutes is equivalent to driving a car 6.4 km. The correct comparison is 9–91 m, which the think tank later acknowledged. The Center for AI and Digital Policy said in 2022 that "AI-enabled systems require exponentially rising computing power . . . despite overwhelming evidence showing that the assertion was misleading and overblown," the ITIF report states.

Vast efficiency improvements in data-center operations have occurred as centers have proliferated. One widely cited 2020 study by Eric Masanet and colleagues, between 2010 and 2018, states that compute instances at data centers surged by a factor of more than five and storage capacity by a factor of 25, while total energy use rose just 6%.

But de Vries and others argue that AI will grow so rapidly that it will overwhelm any further increases in efficiency that can be wrung out of computing hardware, which they say is already reaching the limits of Moore's law for chip feature sizes. "There's a fear of missing out," says Roberto Verdecchia, an assistant professor at the University of Florence in Italy. "There is a race to make a better model, and a new model comes out every other day. There isn't time to improve energy efficiency."

Developer awareness

A 2023 report by Google and Boston Consulting Group notes that AI model design is an evolving field, and new releases and versions consistently demonstrate improved energy efficiency while maintaining performance. Improvements in software and algorithmic optimization are likely to significantly enhance efficiency and decrease computational requirements, the report says. For example, 18 months after the release of GPT-3, the AI model used by ChatGPT, Google pro-



GOOGLE SEARCHES would require more than 10 times the electricity if artificial-intelligence functionality were added, according to two different analyses. (Adapted from A. de Vries, *Joule* **7**, 2191, 2023.)

duced an LLM nearly seven times as large. That model, GLaM, outperformed GPT-3 and required one-third the energy to train, according to the ITIF report.

A spokesperson for OpenAI says the company recognizes that training large models can be energy intensive and that it's "constantly working to improve efficiencies." The company gives "considerable thought about the best use of our computing power and support efforts with our partners to meet their sustainability goals," she says.

OpenAI's model-training runs, while individually very energy intensive, often enable customers to skip having to train their own models from scratch, the spokesperson says. "We also believe that large language models can be helpful in accelerating scientific collaboration and discovery of climate solutions."

Microsoft, which introduced its LLM chatbot Copilot (formerly known as Bing Chat) in 2023, is investing in R&D to better measure the energy usage and carbon intensity of AI and find ways to make large models more efficient to train and use, a spokesperson says.

There are practical limits to the size and improvements that can be made to LLMs—and hence limits on their energy consumption. Neil Thompson, director of the FutureTech research project at MIT's Computer Science and Artificial Intelligence Lab, says that halving the errors produced by LLMs requires 1.9 million times the amount of computing.

Jonathan Koomey, who is part of a team of researchers preparing a congressionally mandated report on electricity use in data centers, says that the surge of interest and investment in AI recalls previous IT hype cycles, such as the dot-com craze and the overbuilding of fiber-optic networks in the 1990s. "There are real uses that will come out of AI, but I'm very skeptical that this will be a thing that takes over the whole economy." He says that he isn't sure that AI would improve the accuracy of Google searches. "If you can automate the testing of accuracy, great. But the world is messy and full of bad actors trying to cause trouble, and there are gray areas and ambiguities. You may never be able to solve the accuracy problem."

ITIF's Castro foresees "plenty of places where people will try and fail to use AI or they'll do it because of the novelty where there's no value." Developers, he says, are looking for ways to create smaller models that are just as effective. Caching answers to frequently asked prompts instead of generating entirely new responses for each instance is an example, he says.

David Kramer

AS BITCOIN'S PRICE reaches record levels, the cryptocurrency is expected to attract more power-hungry mining operations.

CO_2-emitting sources—more green energy than is used by almost any other industry sector. The UN report and the CCAF both estimate that about 38% of bitcoin mining's energy in 2021–22 came from clean sources. A July report from Greenpeace identified a similar proportion, while noting that some coal-fired plants that had been slated to close were kept open or even reopened to fill demand from bitcoin mining. The fraction of green power has likely fallen since China's largely hydroelectric-powered miners mostly disappeared, says CCAF's Neumüller.

A 2022 report by the Sierra Club and Earthjustice accused some companies of "greenwashing" by locating their plants in proximity to wind or solar farms. Unless a company has a power purchase agreement or a direct connection to a renewable supplier, the proportion of renewables they use will be the same as that of the grid from which they draw. Apart from a few publicly traded bitcoin mining companies, few self-report their energy consumption source, Neumüller says.

When bitcoin was launched in 2009, the maximum supply was set at 21 million to keep the currency scarce and prevent inflation. Bitcoins currently number around 19 million. The reward for mining a block is set to periodically halve to reduce the rate at which new bitcoins can be minted. A halving is set for this year.

David Kramer PT