The tricky task of calculating AI’ s energy use

ProQuest document link

ABSTRACT (ENGLISH)

The International Energy Agency, a forecaster, estimates that global data-centre power demand could increase by between 128% and 203% by 2030, mostly because of AI-related energy consumption. Brent Thill of Jefferies, an analyst, estimates that this stage accounts for 96% of the overall energy consumed in data centres used by the AI industry. A raft of improvements meant that the final training run was more than ten times faster than that of Meta’ s Llama 3.3 model just a few weeks earlier, with a roughly proportionate reduction in power used. Lynn Kaack, who leads the AI and Climate Technology Policy Group at the Hertie School in Berlin, worries that, by increasing efficiency and reducing costs in areas like shipping, AI will incentivise companies to increase their activity.

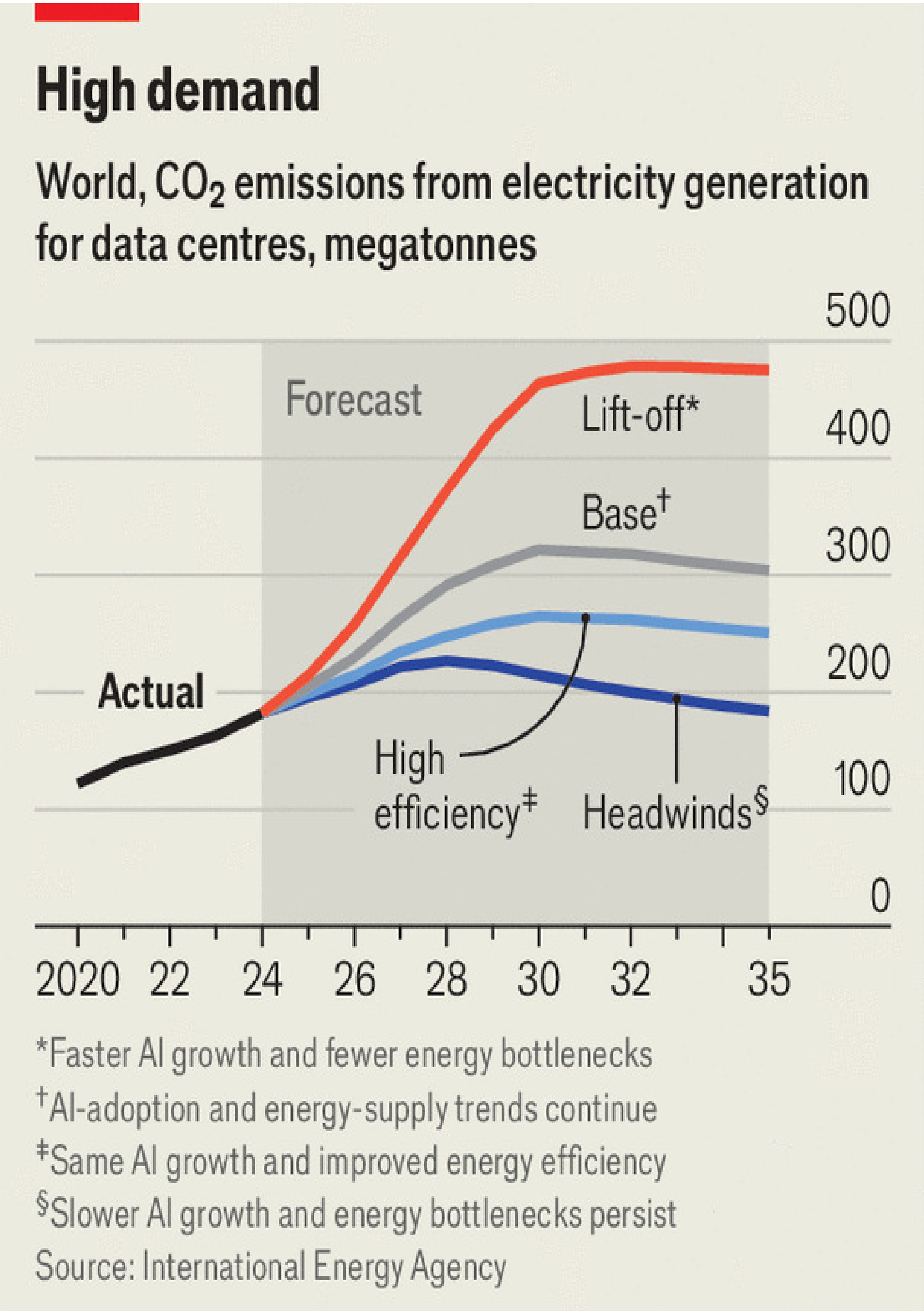
FULL TEXT

A fifth of all electricity used in Ireland is spent powering the country’ s data centres, more than is used by its urban homes. With one data centre for every 42,000-odd people, Ireland has one of the highest per-person concentrations of computing power in the world. Loudoun County, just outside Washington, DC, beats it: its 443,000 residents rub shoulders with scores of data centres—more than the next six biggest clusters in America combined. In 2022 their peak energy usage was almost 3 gigawatts (GW), a power draw that, if maintained year round, would approach Ireland’ s total annual consumption.

Around 1.5% of global electricity is spent on powering data centres. Most of that is for storing and processing data for everything from streaming video to financial transactions. But artificial intelligence (AI) will make up much of future data-centre demand. By 2038 Dominion, a power company, expects the data centres in Loudoun County alone to need more than 13GW. The International Energy Agency, a forecaster, estimates that global data-centre power demand could increase by between 128% and 203% by 2030, mostly because of AI-related energy consumption.

Big tech is confident that the environmental benefits justify the costs. “AI is going to be one of the main drivers of solutions to the climate situation,” says Demis Hassabis, the boss of Google DeepMind. Others disagree. This week’ s special section explores the arguments in detail. It examines the ways in which AI can help clean up some of the most polluting industries, including energy production and heavy industry, and discusses the possibility of moving data centres off Earth altogether. It will also examine why AI’ s energy footprint is so hard to quantify, and what its true environmental impact might be.

Tech firms are generally unwilling to share information about their AI models. One indirect way to estimate the environmental impact of building and deploying AI models, therefore, is to look at the firms’ self-reported carbon emissions. Google’ s greenhouse-gas emissions rose by almost half between 2019 and 2023, according to the search giant, primarily because of increases in the energy consumption of data centres and supply-chain emissions. Microsoft’ s emissions jumped by roughly a third in 2023, compared with three years earlier, partly due to its own focus on AI.



Enlarge this image.

Another approach to estimating AI’ s environmental footprint is to add up the energy use of the infrastructure used to build the models themselves. Meta’ s Llama 3.1, a large language model (LLM), for example, was trained using chips from Nvidia which can draw 700 watts of power each, around half that of a fancy kettle, and it ran those chips for a cumulative 39.3m hours. The resulting energy used, 27.5 gigawatt-hours (GWh), is enough to supply 7,500 homes with a year's worth of power.

Tech companies, perhaps unsurprisingly, are keen to argue that this energy bill is not nearly as outlandish as it might appear. The immediate climate impact of the final Llama 3.3 training run, Meta estimates, is emissions worth 11,390 tonnes of CO_2—about the same as 60 fully loaded return flights between London and New York. Those are the emissions, at least, of the power grid that supplied the company's data centre. But Meta argues that, since electrons are fungible, if enough renewable energy is bought on the opposite side of the country—or even at another time altogether—the true emissions fall to zero.

Put to good use

Focusing on the energy impact of training models, however, may be a distraction. Boris Gamazaychikov, who is in charge of AI sustainability at Salesforce, a software company, compares it to trying to estimate the carbon footprint of a flight by including the impact of building the plane itself. Not only is that construction cost tiny compared with the fuel used over a typical lifetime in service, it's also impossible to calculate the per-passenger impact until the aircraft is finally retired.

Instead, he says, it is best to focus on the energy impact of using AI, a process called inference. Brent Thill of Jefferies, an analyst, estimates that this stage accounts for 96% of the overall energy consumed in data centres used by the AI industry. Mr Gamazaychikov is trying to put hard numbers on that side of the industry, working with HuggingFace, an AI cloud provider, to systematically test the efficiency of hundreds of AI models. The results show the difficulty of generalising: the difference between the most and least power-hungry models is more than 60,000-fold.

Some of that difference arises from the AI models' varying purposes. The most efficient model tested, called BERT-tiny, draws just 0.06 watt-hours (Wh) per task—about a second's worth on an exercise bike—but is useful only for simple text-manipulation tasks. Even the least power-hungry image-generation model tested, by contrast, requires 3,000 times as much electricity to produce a single image.

All the same, says Sasha Luccioni of HuggingFace, concrete figures are not always available. Her company could test only the models it could download and run on its own hardware. "OpenAI has not released a single metric about ChatGPT," Ms Luccioni says, even though such data exist.

Another difficulty in calculating energy use is the fact that AI models are rapidly evolving. The release of DeepSeek V3 in December, a top-tier AI model made by a lab spun off from a Chinese hedge fund, initially looked like good news for those concerned about the industry's energy use. A raft of improvements meant that the final training run was more than ten times faster than that of Meta's Llama 3.3 model just a few weeks earlier, with a roughly proportionate reduction in power used. Inference also became less power-hungry.

In January, as the implications of that improvement became clear, the stock prices of chipmakers crashed. But Satya Nadella, the boss of Microsoft, predicted the upset would be brief, citing the Jevons paradox, a 19th-century observation that the rising efficiency of steam engines opened up new economic uses for the technology and thereby raised demand for coal.

For AI, the rebound effect arrived in the form of "reasoning" models, including DeepSeek's follow-up model, R1. If normal chatbots exhibit what Daniel Kahneman, a psychologist and Nobel economics laureate, called "type one" thinking—prioritising speedy responses—reasoning models display "type two": structured replies that attempt to break a problem into its constituent parts, solve it with a variety of approaches, and check their answer is correct before settling on it as the final response.

Training a reasoning model is not much harder than training a normal AI system, especially if you have pre-existing models to learn from. But running it requires significantly more power, since the "reasoning" step, in which the problem is thought through before a final answer is reached, takes longer. The efficiency improvements DeepSeek pioneered in V3 were more than eaten up by the extra thinking time used by R1 a couple of months later.

If models become more efficient still, there are yet more uses to which they can be put. In recent months, several AI labs have launched "Deep Research" tools, combining reasoning models with the ability to search the web for information and set themselves follow-up tasks. The tools are one of the first mainstream examples of what the AI industry calls “agents”, quasi-autonomous AI systems that can carry out many tasks sequentially. And because it takes them between five and 30 minutes to give a response, running such an agent uses more energy than asking a simple query.

Such efficiency gains leave some wary of the Jevons paradox popping up in other industries. Lynn Kaack, who leads the AI and Climate Technology Policy Group at the Hertie School in Berlin, worries that, by increasing efficiency and reducing costs in areas like shipping, AI will incentivise companies to increase their activity.

Those concerned about the trajectory of AI’s environmental costs are looking for ways to alter it. Mr Gamazaychikov, for instance, hopes that his effort to rank various AI models will allow users and businesses to find the most efficient one for any given task, rather than always using the “best”.

But the closed nature of the biggest labs complicate things. OpenAI, for instance, gives away access to its top-tier models below cost, according to Sam Altman, its boss; Google and Amazon charge less for access to their own AI systems than the cost of the electricity alone, insiders claim. That means users have less motivation to hunt for the most efficient model than they would if they had to pay the true cost of their use. And greater transparency around efficiency and emissions may not result in meaningful behavioural change: after all, there is little evidence to show that growing awareness of the carbon cost of flying has stopped people taking flights.

Coming clean

Many observers think that the best way forward is through tighter regulation, both of AI itself and of the energy it consumes. The first has had limited success in Europe—from the summer of 2026, developers of “high risk” AI will need to tell regulators about the energy it consumes—and is struggling to get off the ground almost everywhere else. In America the Trump administration’s bonfire of red tape means voluntary efficiency drives are more likely than new regulations.

That said, trying to regulate the development of AI specifically is not the only option: broader policies meant to motivate emissions cuts, such as carbon pricing, can help too. Arguably the most important change will come from speeding up the transition to clean energy, and boosting the amount available so that demand for greener AI does not gobble up the low-carbon electricity also needed to decarbonise other sectors, from transportation to construction. Figuring out how to do that shouldn’t require Deep Research.

Curious about the world? To enjoy our mind-expanding science coverage, sign up to Simply Science, our weekly subscriber-only newsletter.

DETAILS

Subject:	Computer centers; Artificial intelligence; Electricity; Estimates; Emissions; Carbon; Electric power; Construction costs; Energy efficiency; Environmental impact; Energy consumption
Business indexing term:	Subject: Computer centers Artificial intelligence Construction costs
Location:	Ireland
Company / organization:	Name: Hangzhou DeepSeek Artificial Intelligence Co Ltd; NAICS: 541715; Name: OpenAI; NAICS: 541715
Publication title:	The Economist (Online); London
Publication year:	2025

Publication date:	Apr 9, 2025
Section:	Science &technology
Publisher:	The Economist Intelligence Unit N.A., Incorporated
Place of publication:	London
Country of publication:	United Kingdom
Publication subject:	Business And Economics
Source type:	Magazine
Language of publication:	English
Document type:	News
ProQuest document ID:	3188131199
Document URL:	https://www2.lib.ku.edu/login?url=https://www.proquest.com/magazines/tricky-task-calculating-ai-s-energy-use/docview/3188131199/se-2?accountid=14556
Copyright:	Copyright The Economist Intelligence Unit N.A., Incorporated Apr 9, 2025
Last updated:	2025-04-10
Database:	ProQuest One Academic

LINKS

Linking Service