

How much electricity does AI consume?

It's common knowledge that machine learning consumes a *lot* of energy. All those AI models powering email summaries, regicidal chatbots, and videos of Homer Simpson singing nu-metal are racking up a hefty server bill measured in megawatts per hour. But no one, it seems – not even the companies behind the tech – can say exactly what the cost is.

Estimates do exist, but experts say those figures are partial and contingent, offering only a glimpse of AI's total energy usage. This is because machine learning models are incredibly variable, able to be configured in ways that dramatically alter their power consumption. Moreover, the organizations best placed to produce a bill – companies like Meta, Microsoft, and OpenAI – simply aren't sharing the relevant information. (Judy Priest, CTO for cloud operations and innovations at Microsoft said in an e-mail that the company is currently "investing in developing methodologies to quantify the energy use and carbon impact of AI while working on ways to make large systems more efficient, in both training and application." OpenAI and Meta did not respond to requests for comment.)

One important factor we can identify is the difference between training a model for the first time and deploying it to users. Training, in particular, is extremely energy intensive, consuming much more electricity than traditional data center activities. Training a large language model like GPT-3, for example, is estimated to use just under 1,300 megawatt hours (MWh) of electricity; about as much power as consumed annually by 130 US homes. To put that in context, streaming an hour of

Netflix requires around 0.8 kWh (0.0008 MWh) of electricity. That means you'd have to watch 1,625,000 hours to consume the same amount of power it takes to train GPT-3.

But it's difficult to say how a figure like this applies to current state-of-the-art systems. The energy consumption could be bigger, because AI models have been steadily trending upward in size for years and bigger models require more energy. On the other hand, companies might be using some of the proven methods to make these systems more energy efficient – which would dampen the upward trend of energy costs.

The challenge of making up-to-date estimates, says Sasha Luccioni, a researcher at French-American AI firm Hugging Face, is that companies have become more secretive as AI has become profitable. Go back just a few years and firms like OpenAI would publish details of their training regimes – what hardware and for how long. But the same information simply doesn't exist for the latest models, like ChatGPT and GPT-4, says Luccioni.

“With ChatGPT we don't know how big it is, we don't know how many parameters the underlying model has, we don't know where it's running ... It could be three raccoons in a trench coat because you just don't know what's under the hood.”

“It could be three raccoons in a trench coat because you just don't know what's under the hood.”

Luccioni, who's authored several papers examining AI energy usage, suggests this secrecy is partly due to competition between companies but is also an attempt to divert criticism. Energy use statistics for AI – especially its most frivolous use cases – naturally invite comparisons to the wastefulness of cryptocurrency. “There's a growing awareness that all this doesn't come for free,” she says.

Training a model is only part of the picture. After a system is created, it's rolled out to consumers who use it to

generate output, a process known as “inference.” Last December, Luccioni and colleagues from Hugging Face and Carnegie Mellon University published a paper (currently awaiting peer review) that contained the first estimates of inference energy usage of various AI models.

Luccioni and her colleagues ran tests on 88 different models spanning a range of use cases, from answering questions to identifying objects and generating images. In each case, they ran the task 1,000 times and estimated the energy cost. Most tasks they tested use a small amount of energy, like 0.002 kWh to classify written samples and 0.047 kWh to generate text. If we use our hour of Netflix streaming as a comparison, these are equivalent to the energy consumed watching nine seconds or 3.5 minutes, respectively. (Remember: that’s the cost to perform each task 1,000 times.) The figures were notably larger for image-generation models, which used on average 2.907 kWh per 1,000 inferences. As the paper notes, the average smartphone uses 0.012 kWh to charge – so generating one image using AI can use almost as much energy as charging your smartphone.

The emphasis, though, is on “can,” as these figures do not necessarily generalize across all use cases. Luccioni and her colleagues tested ten different systems, from small models producing tiny 64 x 64 pixel pictures to larger ones generating 4K images, and this resulted in a huge spread of values. The researchers also standardized the hardware used in order to better compare different AI models. This doesn’t necessarily reflect real-world deployment, where software and hardware are often optimized for energy efficiency.

“Definitely this is not representative of everyone’s use case, but now at least we have some numbers,” says Luccioni. “I wanted to put a flag in the ground, saying ‘Let’s start from here.’”

“The generative AI revolution comes with a planetary cost that

is completely unknown to us.”

The study provides useful relative data, then, though not absolute figures. It shows, for example, that AI models require more power to generate output than they do when classifying input. It also shows that anything involving imagery is more energy intensive than text. Luccioni says that although the contingent nature of this data can be frustrating, this tells a story in itself. “The generative AI revolution comes with a planetary cost that is completely unknown to us and the spread for me is particularly indicative,” she says. “The tl;dr is we just don’t know.”

So trying to nail down the energy cost of generating a single Balenciaga pope is tricky because of the morass of variables. But if we want to better understand the planetary cost, there are other tacks to take. What if, instead of focusing on model inference, we zoom out?

This is the approach of Alex de Vries, a PhD candidate at VU Amsterdam who cut his teeth calculating the energy expenditure of Bitcoin for his blog *Digiconomist*, and who has used Nvidia GPUs – the gold standard of AI hardware – to estimate the sector’s global energy usage. As de Vries explains in commentary published in *Joule* last year, Nvidia accounts for roughly 95 percent of sales in the AI market. The company also releases energy specs for its hardware and sales projections.

By combining this data, de Vries calculates that by 2027 the AI sector could consume between 85 to 134 terawatt hours each year. That’s about the same as the annual energy demand of de Vries’ home country, the Netherlands.

“You’re talking about AI electricity consumption potentially being half a percent of global electricity consumption by 2027,” de Vries tells *The Verge*. “I think that’s a pretty significant number.”

A recent report by the International Energy Agency offered

similar estimates, suggesting that electricity usage by data centers will increase significantly in the near future thanks to the demands of AI and cryptocurrency. The agency says current data center energy usage stands at around 460 terawatt hours in 2022 and could increase to between 620 and 1,050 TWh in 2026 – equivalent to the energy demands of Sweden or Germany, respectively.

But de Vries says putting these figures in context is important. He notes that between 2010 and 2018, data center energy usage has been fairly stable, accounting for around 1 to 2 percent of global consumption. (And when we say “data centers” here we mean everything that makes up “the internet”: from the internal servers of corporations to all the apps you can’t use offline on your smartphone.) Demand certainly went up over this period, says de Vries, but the hardware got more efficient, thus offsetting the increase.

His fear is that things might be different for AI precisely because of the trend for companies to simply throw bigger models and more data at any task. “That is a really deadly dynamic for efficiency,” says de Vries. “Because it creates a natural incentive for people to just keep adding more computational resources, and as soon as models or hardware becomes more efficient, people will make those models even bigger than before.”

The question of whether efficiency gains will offset rising demand and usage is impossible to answer. Like Luccioni, de Vries bemoans the lack of available data but says the world can’t just ignore the situation. “It’s been a bit of a hack to work out which direction this is going and it’s certainly not a perfect number,” he says. “But it’s enough foundation to give a bit of a warning.”

Some companies involved in AI claim the technology itself could help with these problems. Priest, speaking for Microsoft, said AI “will be a powerful tool for advancing

sustainability solutions,” and emphasized that Microsoft was working to reach “sustainability goals of being carbon negative, water positive and zero waste by 2030.”

But the goals of one company can never encompass the full industry-wide demand. Other approaches may be needed.

Luccioni says that she’d like to see companies introduce energy star ratings for AI models, allowing consumers to compare energy efficiency the same way they might for appliances. For de Vries, our approach should be more fundamental: do we even need to use AI for particular tasks at all? “Because considering all the limitations AI has, it’s probably not going to be the right solution in a lot of places, and we’re going to be wasting a lot of time and resources figuring that out the hard way,” he says.

<https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption>