Like any industrial activity, the creation and use of AI has an environmental cost. The two concerns that come up most often are greenhouse gas (GHG) emissions and water usage.[1] As of 2025, these impacts are small at the individual level and moderate at the global level, but the global impacts are expected to increase in the future.

What counts as "energy used by AI"?

AI is a broad field that includes LLMs and other types of generative AI but also prediction and recommender systems,[2] web search, ad targeting, and computer vision.[3] When talking about "energy use from AI," sources may implicitly include all of these or exclude some, which makes it hard to compare numbers or understand which of these subcategories are more energy-intensive. Large companies that train AI generally publish their total energy use, but rarely provide a detailed breakdown of how much energy went to AI. Similarly, they do not distinguish between training the models and "inference," i.e., running the models.[4]

Furthermore, different kinds of generative AI vary greatly in their energy usage. For instance, image generation is about as energy intensive as text generation, but video generation uses substantially more energy.

All that is to say that there is no single number for "how much energy does AI use," and the data we have is full of holes. We concentrate here on best-guess estimates of "what are the individual and global environmental impacts of using an LLM like ChatGPT to generate text?"[5] since this is a question that has received a lot of media attention, but there are other ways to frame the question.

Calculating the impact

The vast majority of AIs are run in datacenters, which require electricity and water to operate.

Datacenters emit very few GHGs on site, and the majority of emissions come from energy generation. While these emissions can vary substantially by country and even by state, we'll use here the American average of 0.4 g of CO_2 per Wh.

In the same way, some water is used to cool datacenters, but the majority of water usage comes from energy generation. The US Department of Energy found that in total about 5 mL of water are evaporated per Wh of energy used by datacenters.[6]

Since power use is proportional to both water use and GHG emissions, most of the research on the environmental impacts of AI focuses on estimating power use and, from that number, calculates water use and GHG emissions.

---

[1] Other concerns include air pollution (which correlates with GHG emissions), land use, the impacts of mining for materials, and electronic waste.

[2] Dylan Patel of SemiAnalysis estimates that 60% of Meta's GPUs are used for recommender systems.

[3] We don't know how much of the energy use for AI is for LLMs; Andy Masley expects it to be under 3%, but there has been little research on this subject.

[4] There are some efforts to make this data more transparent.

[5] ChatGPT is a frontend to many models; we consider here the standard model as of July 2025, GPT-4o. The impact of reasoning models such as o3 is likely substantially higher, but is not well-studied.

[6] The report bases this figure on 100 billion liters of direct water consumption plus 800 billion liters of indirect water consumption (from energy generation), divided by 176 TWh of energy use.

Although other estimates have been made, for the rest of this article, we'll use these estimates of 0.4 g of CO2 and 5 mL of water for every Wh used in a US datacenter.

Impact of individual LLM queries

Let's start small and check the energy use of a single LLM query. For instance, should you worry that using ChatGPT to help you write a single email will have an unreasonable impact?

For popular models, the total inference cost across all users exceeds training compute,[7] [8] so we'll ignore the training compute here and concentrate on inference.[9] Estimates of how much energy is needed to run one query vary, but most are under 3 Wh, often by a lot. Still, we can use this figure as an upper bound[10] [11] of about 15 mL of water use[12] and 1 g of CO_2-equivalent.[13] [14] [15]

Is that a lot? Not really. The request to GPT-4o uses about one hundredth of the energy needed to boil a cup of water, and emits a similar amount of GHGs to driving 0.5 meters.

We often underestimate how much water is used in everyday activities and products we buy. The 15 mL of water used to power a single ChatGPT query is thousands of times less than growing a head of lettuce, and about one million times less than producing a kg of beef or chocolate. It is also dwarfed by the water needed to produce everyday items, such as the 2700 L needed to produce a t-shirt. The claim that ChatGPT uses 10x the energy of a traditional Google search is likely true, but Google searches have a minuscule impact compared to other personal choices, so 10x that is still tiny.[16]

Adding 1000 extra queries to ChatGPT per day would increase the average American's total carbon footprint and water usage by no more than 5% and 0.5% respectively. One

---

[7] The training of frontier models only uses a small percentage of all the GPUs these companies have access to. This suggests that there might be a substantial amount of compute that is used in experiments that are not ultimately used in published models. Or the extra compute may be intended for some other purpose, e.g., preparation for training larger AIs in future.

[8] Epoch has found that training and inference compute are comparable, whereas others find that inference is 80-90% of compute.

[9] We also ignore emissions required to build and eventually dispose of the hardware.

[10] The energy use per query seems to be going *down*, which may seem surprising given that the size of models has been growing, but a combination of better training algorithms, more efficient chips, model distillation, and inference efficiency improvements such as mixture-of-experts has been sufficient to push the inference costs down.

[11] Julien Delavande of Hugging Face built a tool to check the real-time use of models, using the open-weight model Qwen2.5-7B-Instruct.

[12] The widely-cited 500 mL of water per query misrepresents the relevant paper: the real value with these numbers varies based on the datacenter used, but was always much smaller than 500 mL.

[13] We can do a sanity check on these small numbers by observing that LLM providers must pay for the electricity and water they use, and they transfer these costs to the customers using their services. If price per token is a good indicator of energy use, which it seems to be, the fact that the cost per token has been sharply dropping fits with more recent analysis pointing to lower inference costs. Since it costs users less than 1 cent to generate a page of text, this suggests that the total monetary cost of both the water and the energy for such a request does not exceed 1 cent. It is in fact much smaller for most models.

[14] Mistral published a somewhat detailed environmental report on the lifecycle of one of its models and found slightly higher numbers.

[15] Google found 0.24 Wh per query to Gemini.

[16] It's a moot point anyway, since Google now serves AI overviews powered by Gemini in response to most searches.

would need to query ChatGPT every 10 seconds to match the impact of watching TV.[17] If you want to reduce your personal water usage, you can either do 100,000 fewer queries to ChatGPT or avoid eating a single beef hamburger or chocolate bar.

LLM use through APIs

Companies that offer LLMs through a web interface or app such as ChatGPT often also serve their products in less visible ways through APIs to enterprise customers. These can power applications such as web searches, customer support bots, helper bots in software, and coding agents.[18] We don't know exactly how the number of such queries compares to ChatGPT queries,[19] but these queries accounted for about 20% of OpenAI's 2024 revenue and 70-75% of Anthropic's Q2 2025 revenue.

There is a lack of published data on this subject, but it looks like as of 2025 API use is comparable to more visible, ChatGPT-like queries, and increasing quickly.

Broader impact

Taking a broader view, what's the total environmental impact of AI today?[20] We must first look at the impact of datacenters more broadly.

The number of datacenters in the US[21] has been increasing recently, in part due to increased use of AI. In 2023 it was estimated that datacenters were responsible for 4% of total electricity consumption[22] and 2% of GHG emissions in the US. Regarding water, the Lawrence Berkeley National Laboratory calculates that datacenters in the US consumed about 800 billion liters in 2023, which is about 0.2% of total US water draw.

But even now, most of the resource use of datacenters is not used to power AI. Estimates of what percentage of datacenters' energy is used for AI vary a lot,[23] but it was probably well under 10% in 2024. If the impact of AI is an order of magnitude less than the impact of all datacenters, that means it's responsible for 0.2% of US GHG emissions and 0.02% of US water draw.

Again, is that a lot? It depends on what benefits this use of AI generates. Most people who use AI do so because they get something out of it, and almost every aspect of using the internet uses AI in some way, so it's hard to be online without using AI at least indirectly. Globally, the impact of AI is comparable to that of some aspects of modernity that many of us take for granted, such as yard work, clothes dryers, or Christmas lights.

Impact on communities

The environmental cost may be manageable both at the global and individual level, but the concentrated impact on local communities can be severe. Datacenters, especially

---

[17] Triple this rate if you are streaming 4k video.

[18] Examples include Cursor, Windsurf, Github Copilot, and Claude Code.

[19] We include here all other chatbots served through a dedicated interface, such as Claude, Gemini, Grok, etc.

[20] We concentrate on US data, but efficiency may vary substantially throughout the world, with, e.g., France producing 5–10x less carbon per Wh than the rest of the world.

[21] The US has about half of worldwide datacenters (including hyperscalers), and these datacenters are used worldwide, so this number over-represents the impact of Americans.

[22] The total yearly US electricity consumption has been hovering around $4 * 10^{15}$ Wh for over a decade.

[23] Alex de Vries estimated 2.5% in 2023, Uptime Institute estimated 2% in Q1 2024, and the Q4 2024 report by the LBNL cited above found ~25%.

the ones that are used to train AI, are very power-dense in that they require a lot of power facilities in one location, which often stresses local energy grids. Similarly, computer chip fabrication facilities use up a lot of water and can stress local water supplies. The number and size of these kinds of facilities is expected to grow, and impacted communities will have to adapt their infrastructure or simply ban the building of such resource-draining facilities.

Future AI use

Now that we've looked at the present, let's look at the future. We can estimate future electricity use by multiplying the energy cost of such queries by how many are made.

For any given level of capacity, the energy cost has been decreasing quite rapidly[24] due to hardware improvements and algorithmic progress. On the flip side, models are getting bigger and 2024 has seen the rise of reasoning models that use substantially more compute at inference time, so the trajectory for average query cost could go either way.

The trajectory for the number of queries is clearer: it is going up. As models get cheaper (see also the Jevons Paradox) and smarter, they are being integrated into more and more products. If eventually an LLM gets queried every time we take an action online (or if AI agents become ubiquitous), the number of queries sent per person could greatly increase and the total impact, both per-person and globally, could be substantial.

Even with the uncertainty regarding efficiency, the trajectory is clear: investment in datacenters for AI is increasing and there are plans to expand the power grid[25] to power these new datacenters.[26] This suggests that the overall environmental impact of AI will increase.

We don't know exactly what the future will bring. New techniques could be developed that could increase or decrease the energy use, and more fundamentally, if AI ends up transforming society, the picture might change in unpredictable ways.

Related

- If we go extinct due to misaligned AI, at least nature will continue, right?
- Isn't the real concern with AI something else?

Scratchpad

[24] Assuming financial costs roughly track energy use.

[25] Historically, the majority of datacenters were powered by renewable sources. However, renewable energy sources have long lead times between deciding to increase power production and actually having it available. This has led some operators to switch to gas power plants, which can be available quickly, to accommodate the rapid building of new datacenters for AI. One review found that in the US, the carbon intensity of electricity used in datacenters was 48% higher than the national average.

[26] Simultaneously, the increasing prevalence of smaller distilled models could drive the cost of inference down.