The Cloud Is Drinking the River

The hidden water and power bill of always-on AI, and the towns paying it so you can have an answer.

Kymata Labs Research · An independent research institution studying how AI systems actually behave in production.

~12 min read

Tags · AI & Energy · Data Centers · Sustainability · Water

Every prompt has a body. It runs on electricity, and the machine that answers it must be kept cool, which in much of the world means water. At the scale of billions of queries a day, that physical cost stops being a rounding error. Drawing on the IEA's 2025 *Energy and AI* report, the same report's own honesty anchor, named community cases in West Des Moines and The Dalles, and carefully-bounded per-query water and energy estimates, this paper argues the "weightless" cloud is one of the thirstiest machines ever built. The real thesis is not aggregate dominance but local concentration and socialised cost: the people paying for the answer are mostly not the people asking the question.

The argument, in moves

1. **Every prompt has a physical cost.** It runs on electricity, and the machine that answers it must be kept cool, which in much of the world means evaporating fresh water. At billions of queries a day, that cost lands somewhere specific.
2. **The aggregate is not the story; the concentration is.** Data centres were only about 1.5% of the world's electricity in 2024 and account for roughly 10% of demand growth to 2030. This is not the profile of a force breaking the global grid.
3. **The load lands unevenly.** Nearly half of US data-centre capacity sits in just five regional clusters, and the cost is paid in specific gallons by specific towns that mostly never agreed to the trade.
4. **Efficiency does not rescue us; it accelerates us.** Per-query energy keeps falling while total demand keeps climbing, because each cheaper query invites a great many more. This is the Jevons paradox, named by Microsoft's own CEO.
5. **The bill rarely arrives at the address of the person who asked the question.**

The one-liner: the cloud was sold as weightless, but its weight is being set down quietly on a handful of specific places, their grids, their aquifers, their rates.

The strongest evidence: a town, not an average

Aggregate statistics are easy to ignore in a way that towns are not. The clearest evidence for what "socialised cost" means is not a global projection; it is a named utility, a named month, and a named share of a real community's supply.

In July 2022, the month before OpenAI finished training GPT-4 in Microsoft's Iowa data-centre cluster, Microsoft pumped roughly **11.5 million gallons of water** into that cluster in West Des Moines, Iowa. According to the local water utility and the Associated Press, that came to about **6% of the district's water for the month** [6]. One company drew six percent of a community's monthly water to cool the machine that trained the model now answering your prompts.

In The Dalles, Oregon, the story is as much about secrecy as about water. Google fought in court to keep its data centres' water use classified as a **trade secret**, and when the figure was finally disclosed it showed the company had used about **274.5 million gallons in 2021, roughly a quarter of all the water used by the town** [7]. The community hosting the facility had to sue to learn how much of its own river the cloud was drinking. These are not claims that AI is boiling the oceans. They are precise, local, and documented, and the precision is the argument.

The authoritative block, and the honesty anchor

The authoritative numbers come from the International Energy Agency. In its 2025 *Energy and AI* report, the IEA put global data-centre electricity at roughly **415 TWh in 2024, about 1.5% of the world's electricity**, and projected it to **more than double to around 945 TWh by 2030** [1], a figure comparable to the entire electricity consumption of Japan today. The United States alone accounted for **45% of global data-centre electricity** in 2024, and in advanced economies data centres are set to drive more than 20% of electricity-demand growth to 2030, nearly half of it in the US.

Here is the figure that makes the paper credible rather than alarmist. The same IEA report notes that data centres account for only about **10% of global electricity-demand growth** to 2030 [1]. At the planetary level the larger pressures on the grid are electrification, industry, and cooling, not AI. The thesis of this paper is therefore not aggregate dominance but **local concentration plus socialised cost**, and the proof of it is geographic: nearly half of US data-centre capacity sits in just **five regional clusters**. A gentle global average can hide a brutal local one, and in those clusters it does.

What the companies say, and what the independent work finds

It would be easy to write this as a horror story of runaway aggregate figures. The honest numbers are sharper than the scary ones, and they survive the caveats. The contested estimates below are presented exactly as ranges, because on this subject the authority lives in the qualifiers.

Claim	Figure (with its caveat)	Source
Per-query water (vendor)	~0.000085 gal, "one fifteenth of a teaspoon" — no published methodology	Altman, 2025
Per-query water (independent)	~500 mL per 20–50 queries — GPT-3-era, incl. power-plant water	UC Riverside, 2023
Training water	Training GPT-3 ~700,000 litres of freshwater	UC Riverside, 2023
Per-query energy (older)	Google search ~0.3 Wh vs ChatGPT ~2.9 Wh	EPRI, 2024
Per-query energy (newer)	GPT-4o query ~0.3 Wh, ~10× lower as models grew efficient	Epoch AI, 2025

In June 2025, OpenAI's Sam Altman offered his own numbers: an average ChatGPT query, he wrote, uses about **0.34 watt-hours** of electricity and roughly **0.000085 gallons of water, "about one fifteenth of a**

teaspoon" [2]. The figure is tiny and reassuring, and the part that matters most is that it was **published with no methodology**, with no way for anyone outside to check it. Held against it, the most-cited independent estimate, UC Riverside's 2023 *Making AI Less Thirsty*, found roughly **500 mL of fresh water per 20 to 50 ChatGPT queries** and estimated training GPT-3 alone consumed on the order of **700,000 litres** of freshwater [3]. The qualifiers deserve as much attention as the headline: that is a GPT-3-era estimate, per 20–50 queries, including indirect power-plant water, not 500 mL for a single query, and some critics put the figure nearer 5 mL per query. What is in dispute is the magnitude; what is not in dispute is that the water cost is real. Energy per query is the same kind of moving target: EPRI's 2024 analysis put a Google search at about 0.3 Wh and a ChatGPT query at about 2.9 Wh [4], while Epoch AI a year later estimates a modern GPT-4o query at around 0.3 Wh, about ten times lower [5]. The sensible way to hold per-query energy is as a range still in motion.

We called it "the cloud" so we'd stop asking where it was

The metaphor did its job. "The cloud" took a continent of concrete, copper, and coolant and turned it into something that sounds like weather: everywhere and nowhere, weightless and free. For most of the internet era that fiction was harmless, because the footprint was small enough that the abstraction cost nothing. AI changed the arithmetic. Training a frontier model and serving it to hundreds of millions of people is a far more physical act than serving a web page, and the IEA's projected doubling of data-centre electricity by 2030 is the moment the abstraction becomes visible, the weather turning out to have a water bill.

The deeper trap is that efficiency does not rescue us; it accelerates us. Per-query energy keeps falling while total demand keeps climbing, because each cheaper query invites a great many more. Microsoft's own CEO, Satya Nadella, named the mechanism in January 2025 by invoking the **Jevons paradox**: "As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of." [8] Every efficiency gain is reinvested as more usage, and the footprint grows on the back of the very improvements meant to shrink it. The draw is already rising: in their own disclosures, Google reported water use up about **20% to 5.6 billion gallons in 2022**, and Microsoft reported a **34% rise from 2021 to 2022** [7].

What to do: stop treating the cloud as weightless

This is not an argument against AI, nor a guilt trip about asking a chatbot a question. It is an argument about who pays and whether they got any say, and the honest leverage is collective rather than personal, because skipping a single prompt does nothing to a county's water table.

- **Individuals:** stop treating the cloud as weightless. When a data centre is proposed near you, ask the questions that decide the deal, whose grid, whose water, whose rates, and reward the companies that disclose over the ones that litigate to hide.
- **Employers:** if you build on AI, part of the footprint is yours. Choose providers and regions on water stress and grid mix, not price and latency alone; measure and report real consumption rather than repeating a vendor's un-methodologised teaspoon figure; and plan for the rebound, because as your AI gets cheaper per query your total usage will climb.
- **Policymakers:** mandate water and energy disclosure **before** a facility is approved, decide explicitly who pays for the grid upgrades new load requires, and govern siting by local water stress rather than available land and tax incentives alone. The cost concentrates in five clusters, and the scrutiny should concentrate there too.

The answer came to you free. Somewhere, it was paid for, in specific gallons, by a specific place, at a street address whose residents mostly never got to vote on the question you just asked.

Frequently asked questions

So is AI actually wrecking the grid, or is this overblown?

Both claims are wrong, and the honest answer sits between them. At the global level the IEA is blunt: data centres were only about 1.5% of the world's electricity in 2024, and they account for roughly 10% of the growth in global electricity demand to 2030, which is not the profile of a dominant driver. The aggregate average is where the real story hides, because the load is so concentrated. The IEA notes that nearly half of US data-centre capacity sits in just five regional clusters, and that in advanced economies data centres drive more than 20% of electricity-demand growth to 2030, nearly half of it in the US. The grid is not melting everywhere. It is straining in particular places, and those places carry the cost for everyone else.

How much water does one ChatGPT query really use?

Nobody can give you a clean number, and you should distrust anyone who claims to. OpenAI's own figure is roughly 0.000085 gallons, about one fifteenth of a teaspoon, but it arrives with no published methodology. The most-cited independent estimate, from UC Riverside in 2023, works out to roughly 500 mL of freshwater per 20–50 queries, and even that figure carries heavy qualifiers: it is GPT-3-era, it counts the water evaporated at the power plants supplying the electricity, and newer, more efficient models almost certainly use less, with some critics putting it closer to 5 mL per query. The exact number is less important than what every version of it confirms, which is that a single query has a real physical water cost, and at billions of queries a day that cost lands somewhere specific.

If the companies are reporting their water use, isn't that transparency working?

It is working now, but it had to be dragged into the light first. In The Dalles, Oregon, Google argued in court that its data centres' water consumption was a trade secret, and only after a legal fight did the number come out: about 274.5 million gallons in 2021, roughly a quarter of the whole town's water. The disclosures that followed are genuine progress, yet the starting posture was secrecy, and the communities hosting these facilities often learned the scale of the draw only after the contracts had been signed.

Won't AI just get efficient enough that this stops mattering?

That is the comforting assumption, and the evidence cuts against it. Per-query energy has indeed fallen, with Epoch AI estimating a typical GPT-4o query at around 0.3 Wh, roughly ten times lower than older numbers. The trouble is that efficiency has not historically shrunk total demand; it has grown it. Microsoft's own CEO invoked the Jevons paradox in January 2025, saying that as AI gets cheaper and more accessible its use "skyrockets" into a commodity "we just can't get enough of." When each query gets cheaper, we simply run far more of them. The IEA's projected doubling of data-centre electricity by 2030 already bakes in the efficiency gains, and the footprint grows regardless.

What can I actually do about it?

Honestly, very little moves the needle at the individual level, and pretending otherwise would be its own kind of greenwashing. This is not a recycling problem you solve at the household sink. The leverage is collective, and it sits with how utilities approve new data-centre load, who pays for the grid upgrades that load requires, what water a facility is permitted to draw in a drought-prone county, and whether companies must disclose consumption before the deal is done rather than after a lawsuit. The one genuinely useful individual act is to stop treating the cloud as weightless, so that when a data centre comes to your county you think to ask exactly whose power and whose water it plans to use.

References

1 International Energy Agency (2025). "Energy and AI": Executive Summary. (Data centres ~415 TWh / ~1.5% of global electricity in 2024; projected to more than double to ~945 TWh by 2030; US = 45% of global data-centre electricity in 2024; data centres ~10% of global electricity-demand growth.) https://www.iea.org/reports/energy-and-ai/executive-summary

2 Altman, S. (2025). “The Gentle Singularity.” (OpenAI’s own framing: average ChatGPT query ~0.34 Wh and ~0.000085 gallons of water, no published methodology.) https://blog.samaltman.com/the-gentle-singularity

3 Li, P., Yang, J., Islam, M. A. & Ren, S. (2023). “Making AI Less Thirsty: Uncovering and Addressing the Secret Water Footprint of AI Models.” arXiv:2304.03271. (GPT-3-era estimate: ~500 mL per 20–50 queries, including indirect power-plant water; training GPT-3 ~700,000 litres of freshwater.) https://arxiv.org/abs/2304.03271

4 Electric Power Research Institute (2024). “Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.” (~0.3 Wh per Google search vs ~2.9 Wh per ChatGPT query.) https://www.epri.com/research/products/3002028905

5 Epoch AI (2025). “How much energy does ChatGPT use?” (A typical GPT-4o query likely ~0.3 Wh, roughly 10× lower than older estimates, due to efficiency gains.) https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

6 O’Brien, M. & Fingerhut, H. (2023). “Artificial intelligence technology behind ChatGPT was built in Iowa, with a lot of water.” Associated Press. (West Des Moines: Microsoft pumped ~11.5M gallons into its Iowa cluster in July 2022, ~6% of district water that month.) https://apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-f551fde98083d17a7e8d904f8be822c4

7 Solomon, C. (2023). “Data centers and their impact on Oregon’s water and power.” Reynolds Center / The Oregonian. (The Dalles: Google fought to keep its water use secret; disclosed ~274.5M gallons in 2021, ~a quarter of the town’s water. Corporate disclosures: Google’s water use rose ~20% to 5.6B gallons in 2022; Microsoft’s rose 34% from 2021 to 2022.) https://businessjournalism.org/2023/11/oregonian-data-centers/

8 Kurtz, A. & Selyukh, A. (2025). “DeepSeek, AI and the Jevons paradox, explained.” NPR Planet Money. (Microsoft CEO Satya Nadella, Jan 27, 2025, on AI and the Jevons paradox.) https://www.npr.org/sections/planet-money/2025/02/04/g-s1-46018/ai-deepseek-economics-jevons-paradox