ChatGPT

AI Energy and Water Use in the Data Center Boom

Executive synthesis

The environmental footprint of modern AI is, in practice, the footprint of the data center industry that trains, hosts, and serves AI models, plus the electricity systems and water systems that support those facilities. [1]

Most frontier AI workloads today run on infrastructure owned and operated by hyperscalers and a smaller number of specialized partners, rather than on “AI company owned” facilities. OpenAI’s stateless APIs and first-party products are hosted on Azure, while OpenAI also commits additional compute through large-scale initiatives such as Stargate. [2] Anthropic has publicly positioned AWS as its primary cloud and training partner, including large dedicated clusters spread across multiple U.S. data centers. [3] Google, Microsoft, and Meta operate extensive data center fleets of their own and publish more detailed operational sustainability programs than most AI-only labs. [4]

On energy, the most defensible “big picture” is that data centers are still a single-digit share of electricity use globally, but they are growing fast enough to matter for grid planning, and they can dominate demand growth in certain regions. [5] On water, the key truth is that impacts are hyper-local and highly design-dependent: the same amount of compute can have dramatically different water consumption depending on cooling technology, climate, and the water intensity of the electricity supply. [6]

Large operators are actively trying to reduce water strain through a mix of measures, including closed-loop liquid cooling, “zero water evaporation” designs, recycled or reclaimed water, and watershed replenishment projects. [7] At the same time, the buildout is so large, and so geographically concentrated, that even “best practice” designs can still create legitimate community pushback around local water supplies, local electricity prices, and air pollution from backup or on-site generation. [8]

How AI compute translates into energy and water

A useful starting point is what an AI-serving data center actually consumes. Servers and accelerators do the computing, but total facility electricity also includes cooling, power conversion, networking, and reliability infrastructure like UPS and backup generation. In modern data centers, servers are often the majority of the electricity load, while cooling can range from roughly 7% in very efficient hyperscale facilities to over 30% in less efficient enterprise sites. [9]

Data center water use has two major components:

Direct, on-site water consumption, primarily for cooling (especially evaporative systems) and, in some facilities, humidification. [10]
Indirect water consumption, embedded in electricity generation, which varies widely with the grid’s generation mix and cooling technology at power plants. [10]

The central engineering tradeoff that drives public confusion is that reducing water can increase energy, and reducing energy can increase water. Evaporative cooling often saves electricity while consuming more water, while air-based or dry-cooler approaches can reduce on-site water consumption but may require more electricity depending on climate and design. [11]

Google's own public framing captures this tradeoff well: in many contexts water cooling can be more energy-efficient than air cooling, and Google has stated that water-cooled data centers can use about 10% less energy than many air-cooled data centers, reducing associated emissions, but that responsible water sourcing and site-by-site decisions are necessary. [12]

Where AI workloads run today

Data center usage patterns for frontier AI companies are often misunderstood because "AI companies" and "data center operators" are frequently different entities. The cleanest mental model is: AI labs rent or contract compute, hyperscalers and specialized infrastructure firms build and operate the facilities, and local utilities and water districts supply the "real world" inputs. [13]

OpenAI primarily runs on Microsoft Azure [14] for stateless APIs and first-party products, while also pursuing additional compute elsewhere through large-scale projects and partnerships. [15] OpenAI's public Stargate materials describe major capacity under development with Oracle [16] and SoftBank Group Corp. [17], including a flagship site near Abilene [18] and multiple additional U.S. sites. [19]

Anthropic has stated that Amazon Web Services [20] is its primary cloud and training partner, and it highlights deep collaboration to optimize training on Trainium accelerators. [21] AWS describes "Project Rainier" as a very large Trainium2-based cluster deployed across multiple U.S. data centers with Anthropic already running workloads. [22]

For Google, the relevant "AI data centers" are largely Google's own global data center fleet. Google has acknowledged significant water use for cooling at many campuses, has published water-use metrics for U.S. sites (starting with 2021), and reports using reclaimed or non-potable water at more than 25% of its data center campuses. [23] Google also reports that its average data center consumed roughly 450,000 gallons of water per day in 2021, and that its global fleet consumed about 4.3 billion gallons in that year. [12]

Microsoft similarly operates a vast data center portfolio and is pushing water-focused design changes explicitly framed as necessary for AI-era power density. Microsoft reports a next-generation design that avoids water evaporation for cooling (while still using small amounts for domestic uses), and it reports a recent fleet-average data center WUE of 0.30 L/kWh, improved versus its earlier reported figures. [24]

Meta operates a global data center fleet and positions water stewardship as a core design and operations issue for scaling AI. Meta describes a "typical" design using direct-to-chip, closed-loop liquid cooling with dry coolers, which can result in no operational water use in the cooling system and minimal site water use beyond domestic and safety needs, depending on local conditions. [25]

xAI is structurally different from most AI labs because it is also acting like a fast-moving data center developer and, in some cases, a power plant developer. Reporting describes xAI's "Colossus" development in Memphis [26] and subsequent expansion near Southaven [27], with the company facing sustained scrutiny over emissions from gas turbines used for on-site power. [28] At the same time, local reporting describes a plan to cool a Memphis xAI facility with treated wastewater, aiming to reduce withdrawals from local drinking water supplies. [29]

What is being built for future growth

The next wave of AI infrastructure is characterized by scale, power density, and speed. For context, the International Energy Agency [30] projects global data center electricity consumption at roughly 415 TWh in 2024 and expects it to roughly double to around 945 TWh by 2030 in its base case. [31] It also emphasizes that grid buildout timelines can be a binding constraint, and estimates that a meaningful share of planned projects could face delays if grid risks are not addressed. [32]

OpenAI's Stargate announcements describe multi-gigawatt AI data center capacity under development. OpenAI states that a 4.5 GW agreement with Oracle would bring Stargate to over 5 GW under development and "over 2 million chips," and it later announced five additional U.S. Stargate sites that it says bring planned capacity to nearly 7 GW. [19] Separately, OpenAI describes dedicated inference and training capacity arrangements with NVIDIA [33] and references deployment across multiple infrastructure partners. [34]

Hyperscalers are also building larger "AI-ready" facilities and are publishing more explicit commitments around community impacts. Microsoft, for example, positions "zero water evaporation" cooling designs as a default for new data center builds, acknowledges potential PUE impacts when replacing evaporative systems, and argues its chip-level cooling approach can mitigate energy penalties. [24]

Meta's buildout similarly includes very large, AI-linked facilities. A Reuters [35] report describes Meta's plan for a new data center in El Paso [36] that could scale to a 1 GW site, with the company stating it will be matched with 100% renewable energy and use a closed-loop liquid cooling system that recycles water. [37]

Google's public posture is that it must expand power procurement alongside data center growth. Google reports signing a corporate agreement to purchase power from multiple SMRs developed by Kairos Power [38], targeting up to 500 MW by 2035, and it reiterates its 24/7 carbon-free energy goal for 2030. [39] On the water side, Google describes ongoing development of "low-water" cooling alternatives intended to reduce water use while preserving energy efficiency. [12]

Across the sector, the energy supply mix for this growth is not expected to be purely renewable in the near term. The IEA's outlook anticipates renewables supplying a large share of incremental demand, while dispatchable sources led by natural gas also expand materially, and nuclear contributes as well, with first small modular reactors projected to come online around 2030. [32]

Policy pressure is also rising. The The White House [40] "Ratepayer Protection Pledge" frames the data center boom as a potential driver of higher electricity prices and calls for companies to build, bring, or buy new generation resources, pay for delivery upgrades, and negotiate rate structures that prevent shifting costs to households. [41] The White House fact sheet explicitly lists signers including Amazon, Google, Meta, Microsoft, OpenAI, Oracle, and xAI. [42]

What the evidence says and where claims go wrong

The most "true" statement about AI, data centers, energy, and water is that the system is measurable at the data center and grid level, but AI's share inside that system is often not explicitly disclosed, which forces researchers into estimates and scenarios. The IEA stresses uncertainty about both today's data center consumption and future AI uptake, and it uses multiple cases to illustrate how efficiency and adoption rates could swing outcomes. [31]

On electricity, there is strong consensus across major analytical institutions that demand is rising rapidly:

Globally, the IEA estimates data center electricity consumption at roughly 415 TWh in 2024 (about 1.5% of global electricity) and projects about 945 TWh by 2030 (just under 3%) in its base case. [9]
In the U.S., Lawrence Berkeley National Laboratory [43] estimates data centers consumed about 176 TWh in 2023, around 4.4% of total U.S. electricity consumption, and ties the post-2017 growth in part to GPU-accelerated servers for AI becoming a significant portion of the server stock. [44]

On water, the same U.S. Department of Energy [45] -supported Berkeley Lab work provides unusually concrete, system-level estimates:

Direct, on-site data center water consumption in the U.S. grew from about 21.2 billion liters (2014) to about 66 billion liters (2023), with hyperscale and colocation accounting for the large majority of the 2023 total. [44]
Indirect water consumption embedded in electricity use can be larger than direct cooling water. Berkeley Lab estimates an indirect water footprint of nearly 800 billion liters for U.S. data centers in 2023, and it reports a national-average indirect water intensity on the order of a few liters per kWh (while noting that PPAs and behind-the-meter generation are not incorporated in that method). [44]

This split between direct and indirect water explains why "water per AI prompt" claims are so fragile. The underlying reality is not one fixed number, it is a product of at least three variables:

How much electricity a given prompt actually triggers (which varies with model size, hardware, batching, latency targets, and workload mix). [46]
How the data center rejects heat (evaporative, hybrid, dry coolers, direct-to-chip loops, immersion), and the local climate. [47]
How water-intensive the marginal electricity supply is at the time and location the compute runs. [6]

Peer-reviewed work reinforces that variability can be extreme. A 2025 review of workload-level water use finds variations exceeding orders of magnitude, driven by differences in grid water intensity and server/workload efficiency among other factors, which implies that "one number for all AI" will often be misleading. [48]

It is also true that some widely-shared AI water estimates originate in scenario modeling rather than direct operator disclosures. The influential "Making AI Less 'Thirsty'" work proposes a methodology and illustrates that training large models can consume millions of liters of water depending on where and when computation occurs, but the paper's framing is explicitly about estimates under specific assumptions, not universal metering of every AI system. [49] This is one reason Brookings warns that "clickbait-y" framing around individual chatbot use is a distraction from the more structural issues, like fragmented water governance and the need for regional planning capacity. [50]

Finally, a practical truth about "responsibility" is that many mitigation options are real, but they are not free:

Water-saving cooling can raise electricity demand in hot conditions, so the climate impact depends on the cleanliness of the power supply. [51]
Energy-saving evaporative cooling can shift burden to local water systems, especially in water-stressed basins. [52]
Efficiency gains can reduce impact per unit of compute, but rapid demand growth can still increase absolute energy and water use. [53]

Responsibility grades

Grades below reflect a bounded rubric based on public evidence: transparency (do they publish water and energy metrics), operational commitments (clear targets like water positive or 24/7 carbon-free energy), demonstrated engineering actions (cooling design changes, recycled water), and community and policy posture (how they address local cost and health impacts). Scores are necessarily imperfect because AI workload shares are rarely disclosed cleanly inside broader cloud footprints. [54]

Google: A-
Google publishes a relatively detailed narrative of cooling tradeoffs, provides fleet-scale water figures, and commits to both 24/7 carbon-free energy by 2030 and explicit water stewardship, including non-potable water use at a meaningful share of campuses and development of lower-water cooling alternatives. [55] The main constraint on a higher score is that absolute demand is rising quickly and water impacts remain locally contentious in some regions, but on disclosure plus engineering direction, Google is among the leaders. [56]

Microsoft: A-
Microsoft is unusually explicit about operational WUE, describes a "zero water evaporation" design shift for new facilities optimized for AI workloads, and directly acknowledges the water-energy tradeoff while claiming mitigation via chip-level cooling and higher-temperature economization. [24] Microsoft also maintains public 2030 commitments on carbon and water, and it reports major progress on renewable electricity matching milestones. [57] The key risk is that rapid growth in energy use can outpace decarbonization in the near term, increasing reliance on carbon-free procurement quality and carbon removals strategies, which is a legitimate point of debate in external assessments. [58]

Meta: B+
Meta's data center-facing sustainability posture is comparatively strong: it frames water stewardship as core to data center design, reports closed-loop dry-cooler configurations that can eliminate operational cooling water in some facilities, and commits to a "water positive in 2030" goal backed by restoration projects and public reporting. [25] The reason this is not an "A" is the sheer scale of planned growth, including very large campuses, where "best practice" still requires high-confidence execution in water-stressed regions and credible renewable matching at the right times and places. [59]

OpenAI: C+

OpenAI is taking material steps to scale infrastructure, including multi-gigawatt plans through Stargate and expanded compute partnerships, but it has comparatively limited direct public accounting of environmental metrics for its own operations, largely because much of its footprint sits inside partners' data center fleets. [60] OpenAI's participation in the White House pledge signals awareness of electricity system impacts and cost allocation, but from a transparency and operational-footprint reporting standpoint, it trails the hyperscalers that actually operate the facilities. [61]

Anthropic: C

Anthropic is relatively transparent about its dependence on AWS at very large scale, but it does not, in public materials associated with its main compute announcements, provide the kind of water and energy commitments or reporting that would allow an independent reader to assess footprint management directly. [3] In practice, Anthropic's responsibility posture for power and water is highly coupled to AWS's data center standards and water-positive program, which are substantial, but that dependency is also why Anthropic's own score cannot match operators that disclose facility-level strategies and metrics. [62]

xAI: D

xAI has at least one potentially meaningful water mitigation strategy, cooling with treated wastewater rather than drawing from drinking water supplies. [29] However, multiple credible reports describe serious governance and environmental compliance concerns, including extensive reliance on gas turbines for on-site power and legal and regulatory conflict around Clean Air Act permitting and associated community health burdens. [63] Relative to peers, the combination of limited transparency, high local externalities, and the apparent normalization of on-site fossil generation during rapid expansion drives the low grade. [64]

---

[1] [5] [9] [13] [27] [31] [46] [53] [56] https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

[2] [15] [60] https://openai.com/index/continuing-microsoft-partnership/
https://openai.com/index/continuing-microsoft-partnership/

[3] [21] https://www.anthropic.com/news/anthropic-amazon-trainium
https://www.anthropic.com/news/anthropic-amazon-trainium

[4] [17] [39] https://sustainability.google/operations/
https://sustainability.google/operations/

[6] [10] [11] [16] [30] [43] [45] [47] [51] [52] https://www.nature.com/articles/s41545-021-00101-w
https://www.nature.com/articles/s41545-021-00101-w

[7] [24] [26] https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/12/09/sustainable-by-design-next-generation-datacenters-consume-zero-water-for-cooling/
https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/12/09/sustainable-by-design-next-generation-datacenters-consume-zero-water-for-cooling/

[8] [50] https://www.brookings.edu/articles/ai-data-centers-and-water/
https://www.brookings.edu/articles/ai-data-centers-and-water/

[12] [23] [33] [36] [55] https://blog.google/company-news/outreach-and-initiatives/sustainability/our-commitment-to-climate-conscious-data-center-cooling/
https://blog.google/company-news/outreach-and-initiatives/sustainability/our-commitment-to-climate-conscious-data-center-cooling/

[14] [49] https://arxiv.org/abs/2304.03271
https://arxiv.org/abs/2304.03271

[18] [63] [64] https://www.eenews.net/articles/xai-gets-air-permit-for-unauthorized-gas-turbines/
https://www.eenews.net/articles/xai-gets-air-permit-for-unauthorized-gas-turbines/

[19] https://openai.com/index/stargate-advances-with-partnership-with-oracle/
https://openai.com/index/stargate-advances-with-partnership-with-oracle/

[20] [29] https://www.governing.com/resilience/wastewater-will-cool-this-memphis-data-center
https://www.governing.com/resilience/wastewater-will-cool-this-memphis-data-center

[22] https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster
https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster

[25] https://about.fb.com/news/2025/12/advancing-water-stewardship-in-metas-data-center-communities/
https://about.fb.com/news/2025/12/advancing-water-stewardship-in-metas-data-center-communities/

[28] [40] https://apnews.com/article/571c16950259b382f9eae61bd59260ef
https://apnews.com/article/571c16950259b382f9eae61bd59260ef

[32] [35] [54] https://www.iea.org/reports/energy-and-ai/executive-summary
https://www.iea.org/reports/energy-and-ai/executive-summary

[34] https://openai.com/index/scaling-ai-for-everyone/
https://openai.com/index/scaling-ai-for-everyone/

[37] [59] https://www.reuters.com/business/meta-commits-15-billion-ai-data-center-texas-2025-10-15/
https://www.reuters.com/business/meta-commits-15-billion-ai-data-center-texas-2025-10-15/

[38] [44] https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf?utm_medium=email&utm_source=substack
https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf?utm_medium=email&utm_source=substack

[41] https://www.whitehouse.gov/articles/2026/03/ratepayer-protection-pledge/
https://www.whitehouse.gov/articles/2026/03/ratepayer-protection-pledge/

[42] [61] https://www.whitehouse.gov/fact-sheets/2026/03/fact-sheet-president-donald-j-trump-advances-energy-affordability-with-the-ratepayer-protection-pledge/
https://www.whitehouse.gov/fact-sheets/2026/03/fact-sheet-president-donald-j-trump-advances-energy-affordability-with-the-ratepayer-protection-pledge/

[48] https://www.sciencedirect.com/science/article/abs/pii/S0921344925001892
https://www.sciencedirect.com/science/article/abs/pii/S0921344925001892

[57] https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report/
https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report/

[58] https://trellis.net/article/microsoft-2025-sustainability-report-not-backing-off-climate-goal/
https://trellis.net/article/microsoft-2025-sustainability-report-not-backing-off-climate-goal/

[62] https://aws.amazon.com/sustainability/data-centers/
https://aws.amazon.com/sustainability/data-centers/