No, AI Doesn't Drink a Bottle of Water per prompt. The Engineering Reality of AI Infrastructure.

Sophy M. Laughing, Ph.D., MBA
Alden Technologies, Inc.

Bo Erik Gustav Hollsten Ruvalcaba
The Cobeal Group

Abstract

Recent headlines have painted artificial intelligence (AI) as an unsustainable "resource guzzler," citing sensational figures like "500 mL of water per ChatGPT prompt" or claims that AI data centers will soon rival countries in electricity use. These narratives have directly shaped public perception, driven calls for urgent regulation, and influenced policy debates across the U.S. and around the globe. Yet, the technical reality behind AI's environmental footprint is far more nuanced and is often misrepresented by advocacy groups, industry incumbents, and media outlets with vested interests.

This white paper examines the most widely cited claims about AI's water and electricity consumption, using validated engineering fundamentals, the latest data center practices, and authoritative sources such as the American Society of Heating, Refrigeration and Air-Conditioning (ASHRAE) Handbook, International Energy Agency (IEA), and the Department of Energy (DOE). We deconstruct the mechanics of data center cooling, clarifying the distinction between legacy open-loop evaporative systems (which do consume significant water) and modern closed-loop, air-cooled, or hybrid systems that dramatically reduce or eliminate direct water usage. The analysis is extended to the energy domain, where we break down the difference between AI training and inference workloads, quantify actual power demands, and reveal how industry leaders are achieving record-low Power Usage Effectiveness (PUE) through advanced hardware, liquid cooling, and intelligent workload management.

The report exposes the strategic motivations behind exaggerated claims, ranging from lobbying for subsidies and regulatory capture to market positioning and greenwashing. We show, with empirical evidence, that water and power use in AI data centers are design variables, not immutable costs, and that responsible engineering choices can slash resource consumption by orders of magnitude. The study also includes actionable recommendations for U.S. regulators, including transparency mandates, performance standards, and targeted incentives to drive continued improvements in both water and energy efficiency.

By separating fact from hype, this white paper aims to equip policymakers, regulators, and industry leaders with a clear-eyed understanding of AI's true environmental impact, ensuring that future regulation and investment are grounded in scientific reality, not media mythology.

Policy Recommendations for Regulators and Industry Leaders

This white paper provides evidence that the most alarming reports of AI's water and electricity consumption are not borne out by actual engineering performance in modern data centers. Headlines, such as "500 mL of water per prompt" or predictions that AI will overwhelm national power grids have influenced public debate, but these statements are rooted in outdated models or worst-case scenarios. Data centers use closed-loop cooling, liquid-cooled AI hardware, and advanced workload scheduling to reduce water and power consumption per compute unit by a factor of ten compared to industry averages from only a few years ago. Regulators and policymakers should move past sensational narratives and ground oversight in engineering facts. Effective regulation depends on transparency, enforceable performance standards, and real incentives for adopting modern infrastructure. Policymakers must watch for greenwashing, selective disclosure, and misleading metrics. These should be replaced by standardized and transparent reporting. AI's true resource impact depends on design, location, and operational choices, not on technological fate. Policy should reflect what modern systems are already achieving and should adapt to ongoing technical progress.

Responsible action means:

- Requiring transparency on energy and water use for large AI facilities.
- Setting efficiency standards based on current best practices instead of legacy averages.
- Supporting the adoption of closed-loop cooling and clean energy across the industry.
- Demanding life cycle resource reporting that covers both training and inference.

AI does not have to be a drain on resources. The tools and knowledge to build sustainable digital infrastructure already exist and are being used by industry leaders. Policymakers must set fact-based regulations and make sustainable AI the standard for all future deployments.

1. Introduction

Public debate over artificial intelligence (AI) and resource consumption is now shaping real-world regulatory priorities. Alarmist headlines claiming that "each AI conversation requires a bottle of water" or that "large AI models will rival small nations in power draw" have found their way into draft policy language, advocacy reports, and media cycles worldwide. These narratives (though attention-grabbing) are often based on outdated data, simplified models, or worst-case scenarios that no longer reflect the engineering realities of modern high-density data centers.

This white paper presents a real-world technical response to these claims by exploring the true water footprint of AI workloads and the actual electricity consumption of modern data centers supporting large-scale AI models. By grounding every argument in validated engineering practice, operational data, and consensus standards from bodies such as ASHRAE and government energy

agencies, we aim to replace speculation and marketing with a transparent, science-based understanding.

1.1. Why This Matters Now

As AI becomes the default workload across sectors, from finance and healthcare to logistics and national infrastructure, the buildout of new data centers and the retrofitting of existing sites are accelerating at an unprecedented pace. Regulatory authorities, sustainability officers, and public sector buyers are all seeking answers to the same questions: How much water and power do these AI systems require? Are the risks being presented in industry literature and policy drafts representative of typical operations, or are they based on edge cases and legacy designs? Most critically, are there proven design and operational choices that can substantially reduce the environmental impact of AI computing, and if so, are they being adopted at scale?

The stakes are not theoretical. Billions of dollars are being invested based on assumptions about 'AI sustainability.' Local utilities are revising grid forecasts and water allocations, and lawmakers are beginning to propose quotas, moratoria, and reporting mandates. It is therefore essential that regulatory frameworks reflect physical reality and not inflated or strategically curated figures.

1.2. Two Myths, One Engineering Reality

This report addresses the two dominant environmental narratives around AI data centers (water consumption and electricity use) in parallel, revealing how both have been consistently exaggerated through selective reporting and technical misunderstanding.

1.3. The Water Footprint Narrative

First, we dissect the widely circulated claim that AI workloads are "water guzzlers," often represented in visuals like bottled water per chat session or million-liter training runs. These claims typically originate from simplified models that extrapolate water usage based on older, evaporative cooling tower systems, ignoring major advances in facility cooling technologies and best practices.

We systematically deconstruct these claims by:

- Describing the thermodynamic principles of data center cooling, detailing how heat is removed from IT hardware.
- Explaining the difference between open-loop evaporative systems (which consume water through evaporation) and closed-loop or air-based systems (which can reduce or even eliminate direct water consumption).
- Presenting real-world water usage effectiveness (WUE) metrics, including recent performance data from leading U.S. operators, which consistently demonstrate an order of magnitude improvement over historic averages.

- Detailing operational strategies such as air-side economization, adiabatic and hybrid cooling, and climate-aware workload scheduling, all of which dramatically cut the need for potable water.
- Demonstrating, with schematic diagrams and site case studies, how facilities in different climates are achieving near-zero water draw for cooling, refuting the premise that water consumption scales linearly with AI workload.
- Highlighting that where water is still used, it is increasingly sourced from non-potable or recycled supplies, further reducing impact on local reservoirs.

By contextualizing every step in the cooling chain, this report shows that sensational per-prompt water usage figures are based on static, outdated assumptions rather than dynamic, modern engineering practice. The path to low-water or water-free AI is not hypothetical: it is already well-established among major U.S. data center operators.

1.4. The Electricity Consumption Narrative

Second, we address the claim that AI workloads will overwhelm electrical grids or rival the power consumption of entire cities. This narrative often emerges from multiplying the highest available energy-per-query or training-run figures across projected AI adoption curves, without accounting for rapid gains in both hardware and system-level efficiency.

Our analysis responds by:

- Disaggregating the phases of AI operation by differentiating between the one-time, high-intensity power use of model training and the ongoing, distributed consumption of inference at scale.
- Mapping the end-to-end power flow in a typical AI-ready data center, showing how electricity is apportioned between IT loads, cooling, power conversion, and other facility overheads.
- Utilizing consensus efficiency metrics such as power usage effectiveness (PUE) and presenting actual measured PUE values from recent U.S. data center deployments, which are significantly below industry averages, often cited in the media.
- Reviewing recent advances in high-density server design, including the deployment of liquid cooling, increased allowable temperature setpoints, hot/cold aisle containment, and direct-to-chip cooling. All of these reduce both total power consumption and cooling overhead.
- Addressing the nuances of workload scheduling, such as aligning compute-intensive tasks with periods of grid surplus or renewable generation, which can further mitigate grid impact.
- Critically examining red flags in public reporting: the tendency of some industry players to withhold full energy disclosure, the prevalence of misleading comparisons (e.g.,

"equivalent to millions of EVs charged"), and the habit of reporting per-inference efficiency gains without mentioning absolute scale growth.

We provide empirical evidence that modern AI data centers are not only more energy efficient per unit of compute but also increasingly powered by low-carbon or renewable sources. This matters because the environmental impact per AI query or training run is as much a function of infrastructure as of algorithmic demand.

1.5. Structure of the Report

To maximize clarity and regulatory usefulness, this paper proceeds as follows:

- **Section 2** presents a detailed, diagram-backed analysis of data center cooling systems, with a focus on water use. We compare traditional and modern system architectures, quantify WUE improvements, and debunk per-query water consumption myths with both empirical data and operational examples.
- **Section 3** delivers a breakdown of data center energy use in the AI era. Here, we separate engineering fact from PR-driven narrative, benchmarking real power draws and highlighting where selective disclosure or misleading analogies have distorted the regulatory conversation.
- **Section 4** reframes the discussion around physical, not rhetorical, limits and calls for evidence-based policy anchored in actual best practices.

1.6. The Imperative for Evidence-Based Oversight

The rapid growth of AI workloads, and the resulting construction of high-density digital infrastructure, demand oversight that is as technically sophisticated as the systems being built. This paper equips decision-makers to distinguish between headline-grabbing projections and grounded engineering possibility. By doing so, it aims to prevent misallocation of resources, avoid reactive or misdirected policy interventions, and encourage the adoption of solutions that materially reduce environmental impact.

The central finding is straightforward: water and power usage in AI data centers are design variables, not fixed costs. Modern facilities can achieve radically lower environmental footprints than those assumed by most public reports. It is time for the regulatory conversation to catch up to this reality, so that both innovation and sustainability are properly served.

2. Debunking Exaggerated Claims of AI's Water Footprint: Engineering Realities of Data Center Cooling

2.1. Introduction

Recent reports have raised alarm about water consumption of artificial intelligence (AI) workloads, claiming that "ChatGPT needs to drink a 500 mL bottle of water for a simple conversation of 20-

50 questions" and that "training GPT-3 consumed 700,000 liters of freshwater". These striking figures originate from a 2023 preprint by researchers at UC Riverside and UT Arlington, which estimated AI's "water footprint" by considering data center cooling and even the water used in electricity generation. While water usage is indeed an important sustainability concern, such claims must be examined critically using engineering first principles and real data center practices.

In this white paper, we present a technical analysis of modern data center cooling systems to debunk these inflated water-use claims. By drawing on authoritative sources (e.g., the ASHRAE Handbook and Industry data) and illustrating actual cooling cycles, we show that most AI-focused data centers employ high-efficient, low-water, or even water-free cooling designs. Empirical evidence and engineering fundamentals demonstrate that modern facilities can minimize or nearly eliminate freshwater consumption for AI computing workloads. We begin by reviewing how typical data center cooling works and where water comes into play.

Next, we explore advanced cooling technologies, from air-side economization to closed-loop liquid cooling, which drastically cut water usage. We then present documented water efficiency metrics of leading cloud operators and contrast them with the assumptions behind the UC Riverside/UT Arlington study. Finally, we provide an engineering reality check on the specific claims (e.g., "500 mL per 20-50 prompts" and "700,000 L per training") to explain why those numbers are technically inaccurate or misleading for well-designed AI data centers. Throughout, schematics of cooling systems (closed loops, cooling tower heat rejection, etc.) are included to visually clarify key concepts. The goal is to ground the discussion in factual, engineering-based understanding, including how heat is removed in data centers, how much water that requires, and how modern design practices minimize water consumption while reliably cooling high-density AI hardware.

2.2. Data Center Cooling Fundamentals and Water Use

How do data centers use water for cooling?

Modern data centers dissipate enormous heat loads from servers and must reject that heat to the environment. There are two main stages to this cooling process: first, heat is transferred from server hardware to a facility cooling medium (air or liquid), and second, that heat is expelled from the facility to outside air. Water typically enters the picture in the second stage, if the facility uses evaporative cooling. In many large data centers, especially older designs, cooling towers are used to reject heat. A cooling tower is essentially an open-loop evaporative heat exchanger, it takes warm water from the data center's condensers and sprays it in a tower so that evaporating some of that water carries away heat. This process consumes water, as the hot water trickles over fill material, a portion evaporates into the air. This removes heat but using up fresh water in the form of vapor. The remaining cooled water is recirculated, and makeup water must be continuously added to replace the volume lost to evaporation (as well as blowdown water drained to purge mineral buildup). In summary, any cooling system that relies on phase-change evaporation will

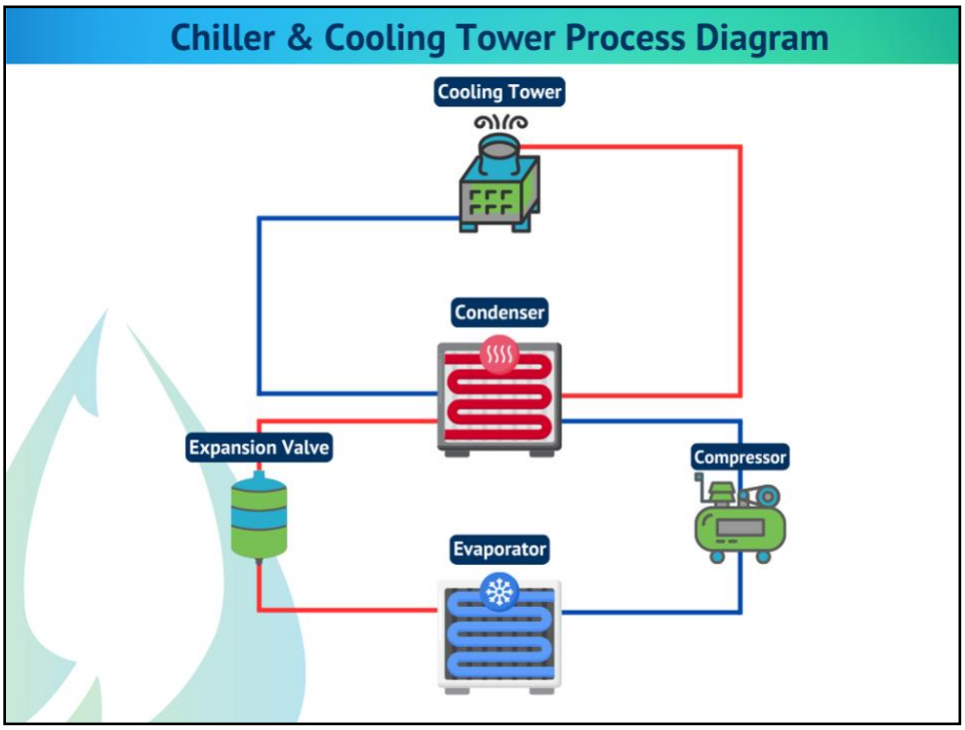
have a direct water footprint: evaporating 1 liter of water removes roughly 2,260 kJ of heat (the latent heat of vaporization), about 0.63 kWh, so evaporative cooling consumes on the order of 1.6 liters of water per kWh of heat dissipated under design conditions, plus additional losses. This aligns with industry observations that traditional data centers using cooling towers average on the order between 1 and 2 L of water per kWh of IT load cooled.

By contrast, air-based cooling methods reject heat without evaporating water. For example, an air-cooled chiller or dry cooler uses ambient air (blown by fans through a radiator) to carry away heat sensibly (via warm air), consuming no water in the process. Many modern facilities also employ air-side economization, bringing in outside air to cool the servers when outdoor conditions are cool enough, again with no need for water. In short, data centers can be designed either to use water for cooling (evaporative open-loop systems) or to operate in a closed-loop manner with purely air-based heat rejection. The water usage patterns differ drastically between these approaches. To illustrate this clearly, we next examine common cooling system configurations and their water requirements, from conventional cooling towers to advanced closed-loop systems.

2.3. Cooling Towers and Evaporative Cooling Systems (Open Loop)

Cooling towers have been a staple in high-capacity HVAC and data center cooling for decades. In an open-loop cooling tower system, water itself is the working fluid that directly dumps heat to the atmosphere by evaporation. The data center's chillers or heat exchangers send warm water (typical ~30-40ºC) to the cooling tower. Within the tower, this water is sprayed over fill media while fans draw air through it, causing a fraction of the water to evaporate and carry away heat. The diagram in Figure 1 depicts a simplified water-cooled chiller loop with a rooftop cooling tower.

Figure 1: Schematic of a water-cooled chiller with an open cooling tower (red lines = hot water, blue lines = cooled water). Warm condenser water from the chiller is pumped up to the cooling tower, where it is cooled by evaporation (wavy lines at the tower indicate the moist air exiting). This open-loop design continuously consumes water: as heat is rejected, some water evaporates, and the remainder (cooled water) returns to the chiller. The system must add makeup water to compensate for evaporation and blowdown losses.



Open cooling towers are effective heat rejectors and can cool water to near the ambient wet-bulb temperature, which is often much lower than the air dry-bulb temperature. This efficiency is why evaporative towers have been popular: they reduce chiller electricity usage by rejecting heat at lower temperatures. However, the trade-off is significant water consumption. A rule-of-thumb estimate is that about 1% of the circulating water is evaporated for each 6-7ºC of cooling achieved. For example, if a chiller's condenser water enters the tower at 35ºC and is cooled to 29ºC, roughly 5 mL of water will be evaporated for each kilowatt of heat rejected per second (equivalently, ~1.8 L per kWh). In practice, real data centers have reported water use on the order of 1-9 liters per kWh of IT energy, depending on climate and cooling design. Google's global fleet averaged ~1 L/kWh in on-site cooling water, while one large commercial data center in Arizona reported a near 9 L/kWh during summer peak. This water is consumed in the sense that it leaves the local water environment as vapor. Moreover, additional "blowdown" water must be bled off to remove mineral concentrations, meaning extra water input is needed. Typically cooling towers operate at 3 to 6 cycles of concentration, so for every 3-6 liters evaporated, approximately 1 liter is dumped via blowdown. The net effect is that open-loop evaporative cooling can strain water resources if used continuously in a large facility. As one U.S. Department of Energy guide bluntly states: "A cooling tower system by necessity uses an extensive amount of water," especially for 24/7 loads like data centers.

It is important to note that the much-publicized figures from the UCR/UTA study assume such evaporative cooling scenarios. The notion of "500 mL per 20-50 prompts" was derived by combining the energy consumption of a series of ChatGPT queries with an assumed water use rate

Figure 1

(likely on the order of 1-2 L/kWh) representative of typical evaporative cooling and power generation. Similarly, the 700,000 L for a two-week AI training was calculated as the direct cooling water "consumed" (evaporated) in a state-of-the-art data center, presumably one employing water-cooled chillers and cooling towers. These numbers are plausible for an open-loop design (e.g., training GPT-3 was estimated to use ~1287 MWh of electricity; at ~0.5 L/kWh on-site WUE, that gives ~700kL), in line with the paper's claim. However, such water usage is not an inherent or fixed requirement of AI computation; it is a consequence of cooling design choices. Modern data centers need not operate this way, as we will see. First, we examine alternatives that drastically cut or eliminate water use by rejecting heat through air or closed loops instead of evaporative open loops.

2.4. Air-to-Air and Air-to-Water Cooling (Closed-Loop Systems)

Not all large-scale cooling relies on evaporating water. Closed-loop cooling systems keep the coolant (water or refrigerant) completely contained and use air as the final heat sink. For example, an air-cooled chiller removes heat from the data center's water loop via a refrigeration cycle and then dumps that heat through an outdoor condenser coil cooled by fans (just like a gigantic air-conditioner). No cooling tower is needed: heat is expelled directly to outside air. This design results in zero routine water consumption on-site. Many enterprise data centers and telecom facilities have long used air-cooled chillers or dry coolers (essentially radiator/fan units) to avoid water use, especially in regions where water is scarce or expensive. The clear benefit is WUE = 0 (Liters per kWh) for the cooling system in normal operation. The drawback is that air-cooled systems can have a higher energy penalty, especially in hot climates because the chiller's compressors must work harder when rejecting heat to 35ºC air vs. a 24ºC wet-bulb via cooling tower. Nonetheless, improvements in IT equipment tolerance and cooling technology have narrowed this gap. Higher allowable server temperatures per ASHRAE guidelines (inlet air up to ~27ºC or even higher) mean chillers don't need to overcool; indeed, ASHRAE TC9.9 recommends supply air temperatures of 18-27ºC for "A1-A4" class IT equipment (with allowable excursions above 30ºC), which has enabled more hours of air-based cooling and less mechanical overcooling. As a result, many new data centers run comfortably at warmer temperatures, making air-cooled heat rejection more feasible without compromising reliability. In short, air-cooled (water-free) designs can handle data center loads efficiently in many climates, and they completely avoid the evaporative water losses that plague cooling towers.

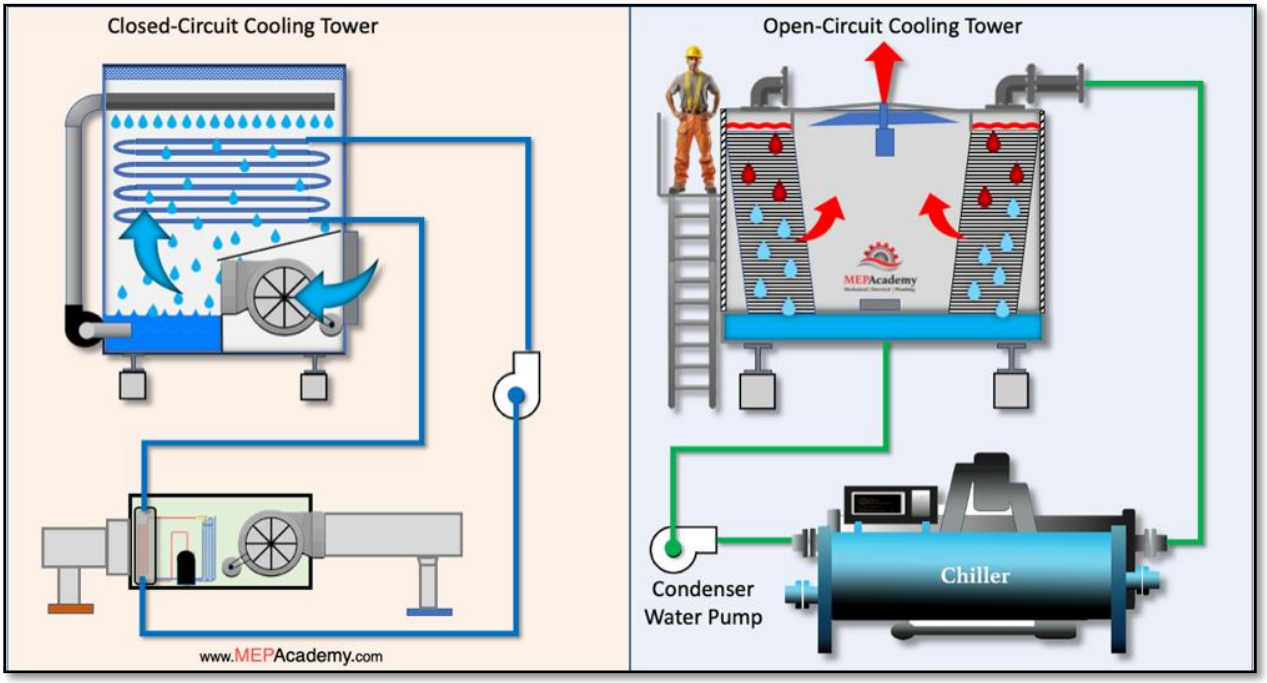
Another approach in modern facilities is air-side economization, a form of air-to-air cooling. In this scheme, when outside weather is cool enough, outside air is directly drawn into the data hall (or passed through an air-to-air heat exchanger) to carry away heat, instead of using chillers at all. This is essentially free cooling by Mother Nature. The only water use might be a small amount for humidity control in dry winter conditions (to keep humidity within safe ranges), but this is minimal compared to evaporative cooling water. Many hyperscale data centers (e.g., Facebook/Meta's in Oregon or Sweden, and Google's in Finland) are designed to leverage northern climates with

extensive air economization. For example, a facility might use outside air cooling for 8-10 months of the year and only on the hottest summer days switch to evaporative assist. By maximizing air-side free cooling, annual water usage plummets. A Lawrence Berkeley Lab case study noted that economizer use can significantly reduce cooling system energy and water demand, depending on climate. In essence, every hour that a data center can use 100% air cooling is an hour of zero water consumption for cooling. Modern automation and control strategies seek to extend those hours as much as possible.

Beyond chillers and air economizers, there are also adiabatic hybrid coolers that blend air and water cooling in a water-efficient way. An adiabatic cooler is basically a dry cooler (air-cooled radiator) that can be augmented with a little water on peak hot days. For example, the unit may have an evaporative pad or a misting system in front of the coil that activates only when ambient air is above a set temperature. This precools the air a few degrees via evaporation, boosting cooling capacity, but uses far less water than a full cooling tower. Manufacturers report that such hybrid adiabatic systems can reduce water consumption by 70-95% compared to traditional cooling towers. Figure 2 illustrates one such hybrid cooling concept. Most of the time it runs dry, like a closed radiator, and only sprays a fine mist on the hottest afternoons. One example product (Nimbus VIRGA) can provide ~500 tons of cooling while consuming <6 gallons per minute in wet mode (≈1.36 m^3/hour), whereas an equivalent conventional tower might consume 20+ gallons per minute (gpm) for the same load.

By dramatically spreading the cooling load across air and using water sparingly, these systems ensure that water use is a last resort. In practice, large cloud operators have widely adopted such approaches in new buildings. Microsoft, for instance, uses indirect evaporative cooling (IDEC) units in many Azure data centers (essentially air-to-water heat exchangers with adiabatic assist) achieving much lower water use than open-loop chillers. In climates like Ireland or Northern Sweden, they may hardly ever need to turn on the water mist, while in hotter climates they still save water relative to a tower by only running wet occasionally. The key point is that modern cooling technologies allow a continuum from fully dry to modest intermittent water use, instead of the old paradigm of continuous evaporative loss.

Figure 2: Open-loop vs. closed-loop cooling tower designs. At right, an open-circuit cooling tower directly contacts water with air: hot condenser water (green line) from the chiller is sprayed and cooled by evaporation (red arrows indicate hot moist air exhaust). This maximizes heat rejection but consumes the most water. At left, a closed-circuit tower keeps the primary coolant in a closed coil (blue loop) and cascades water over that coil. Heat is still rejected by evaporating some water, but the process fluid stays clean. Both designs rely on evaporation and thus use water continuously. In contrast, a fully dry cooler (not pictured) would use a finned coil and fans with no water spray, eliminating evaporation at the cost of higher air flow.



2.5. Liquid Cooling for AI Hardware: High Efficiency with Low Water Footprint

A notable trend for AI-focused data centers is the shift to liquid cooling at the server level (such as direct-to-chip cold plates or immersion cooling). High-performance AI accelerators (GPUs, TPUs, etc.) often produce heat fluxes that air cooling struggles to handle. Liquid cooling (using water or dielectric fluid) can remove heat more efficiently, which lowers the overall cooling energy requirement and enables higher rack densities. But how does this affect water usage? The answer depends on how the heat is ultimately rejected from the liquid. The ideal scenario is a warm-water cooling loop that rejects heat via dry coolers or heat reuse, avoiding any evaporative step. For example, some supercomputing centers (like NREL's ESIF data center) use a warm water closed loop to capture >90% of server heat, then first attempt to reuse it for facility heating, next use an advanced dry cooler (thermosyphon) when ambient permits, and only finally use a cooling tower if absolutely needed during peak conditions. This tiered approach means for much of the year, zero water is used despite enormous HPC heat loads: evaporation is a "last resort" for the hottest times or when other cooling capacity is saturated. The result is a vastly smaller water footprint than running cooling towers 24/7.

Even when heat reuse is not an option, liquid-cooled AI systems can be paired with closed-loop heat rejection. NVIDIA's latest AI supercomputers are a prime example. NVIDIA's new H100 GPU racks (the "NVL72" systems) use direct-to-chip liquid cooling with a closed coolant loop. The GPUs and CPUs are mounted with water-cooled cold plates, and warm liquid from the racks is routed to a coolant distribution unit and then to external heat exchangers. Crucially, no water is evaporated in this process. The loop is sealed, and heat is dumped to ambient via liquid-to-air heat

exchangers (which could be dry coolers or air-cooled chillers). As Tom's Hardware reported, unlike immersion or evaporative cooling, NVIDIA's closed-loop design "does not evaporate or require replacement due to loss from phase change, saving water". NVIDIA claims that this design is "300 times more water-efficient" than conventional evaporative cooling for comparable workloads. In plain terms, that means essentially negligible water usage (a 300x improvement over a baseline that presumably assumed a water-cooled system). While the exact 300x figure is marketing, it illustrates that leading-edge AI infrastructure is moving toward no-water cooling solutions. Other companies are following suit: Meta's latest data center designs incorporate cold plate liquid cooling with dry heat rejection for their AI training clusters, and Google has experimented with liquid-cooling AI pods that attach to existing dry coolers.

The cooling architecture for AI hardware is evolving rapidly, and efficiency and sustainability are major drivers. AI-centric data centers, which often house extremely high-density racks, are more likely to use advanced cooling (meaning either indirect evaporative, low-water or liquid cooling with closed loops) rather than old-fashioned always-on cooling towers. The water usage patterns are fundamentally better in these modern designs: water, if used at all, is used sparingly and intelligently (e.g., only in hot weather, or via recycled sources), and many next-generation facilities aim for zero direct water usage by relying on air cooling or heat reuse. Now, having described the spectrum of cooling methods, let us examine empirical data on how much water cutting-edge data centers use. This will provide context to judge the claims about AI water footprints.

2.6. Water Usage Effectiveness (WUE) and Empirical Data

To quantify data center water efficiency, the industry uses a metric called Water Usage Effectiveness (WUE). WUE is defined as the liters of water used in cooling per kilowatt-hour of IT energy use. It is analogous to Power Usage Effectiveness (PUE) but focusing on water. A lower WUE means less water consumed for the same IT workload. This metric helps compare different cooling approaches on equal footing. An older-generation facility with chillers and cooling towers might have a WUE around 1.5-2 L/kWh, whereas a fully air-cooled site has a WUE of 0.0 L/kWh (no water). It's important to note WUE can be computed for on-site water only (sometimes called WUEscope-1) or including off-site water (like water at power plants, WUEscope-2). Here we'll focus on the on-site cooling water, since that is what data center designers directly control (and what the "cooling systems" discussion above addresses.)

2.7. Real-world WUE figures show dramatic improvements in recent years.

According to TechTarget, an average data center (across all types) historically uses about 1.8 L/kWh. However, cloud giants significantly beat that: Amazon Web Services (AWS) reports a fleet-wide WUE of just 0.19 L/kWh for its data centers, and Microsoft reports 0.49 L/kWh as of 2022. These numbers are an order of magnitude lower than the industry average, reflecting the shift to efficient cooling designs. In practical terms, AWS using 0.19 L/kWh means that for every 1,000 kWh of IT compute, only 190 liters of water are used; a mere 1/10th of what a typical legacy

data center might use for the same workload. Google's data centers, which have widely employed evaporative cooling for energy efficiency, had an on-site WUE around 1.1 L/kWh in 2021 (extrapolated from 4.3 billion gallons over ~10 TWh), but Google has since committed to new "climate-conscious" cooling tech to slash water use by 2030. In fact, Google announced it is developing non-potable water cooling and advanced thermosyphon systems aimed at reducing data center water use by 30 to 50% even in existing evaporative-cooled sites. Furthermore, Google already uses reclaimed wastewater instead of drinking water at roughly a quarter of its campuses. For example, its Douglas County (Georgia, USA) data center is cooled with 100% recycled sewage water, eliminating stress on the freshwater supply. This indicates that even when water is used, it may not be the "fresh water from overtaxed reservoirs" implied in the sensational claims; many AI data centers use non-potable sources or are located near large water bodies to mitigate local water impact.

Crucially, the temporal and spatial diversity of water efficiency is large. The UC Riverside study itself noted that when and where AI workloads run can swing water consumption by a factor of 2 to 3x. For instance, performing training in a cool, humid region (or in winter nights) might use a fraction of the water that the same training would require in a hot, arid region at noon. This is precisely why modern data center operators schedule flexible jobs (like AI model training) strategically. Microsoft researchers have suggested "climate-aware" scheduling: running AI training during cooler nighttime hours or seasons to cut water evaporation losses. Such scheduling can directly reduce cooling tower evaporation because cooler ambient air means towers (if used at all) run more efficiently and possibly can be replaced by dry cooling during those periods. In fact, the paper's authors themselves likened AI training to watering a lawn: "We don't want to water our lawn at noon, so let's not water our AI (at) noon either." This analogy illustrates that water use is a controllable parameter, not a fixed cost of computing. With intelligent scheduling, the water per task can drop further.

Let's put the "500 mL per ChatGPT conversation" claim into perspective now. ChatGPT (GPT-3.5/GPT-4) inference is an interactive workload, but we can estimate its energy per prompt. The preprint study cites an estimate that GPT-3 consumes ~0.4 kWh to generate 100 pages of content, which is about 0.004 kWh per page or per few prompts. Another estimate for a "medium" LLM response on an enterprise GPU system is ~0.016 kWh per query. Even assuming on the higher end (0.01-0.02 kWh per Q&A pair), 20 to 50 prompts would consume on the order of 0.2-1 kWh of energy. Now, if a data center had no special optimizations (say WUE !1 L/kWh, and using typical grid power with water-cooled plants), that 1kWh might indeed entail approximately 1 liter of water evaporated on-site and another liter off-site. Split across ~40 prompts, that is about 50 mL per prompt (hence the study's bottled-water visual). But in a highly efficient AI data center, this drops dramatically. For instance, in an AWS region with WUE 0.2 and largely renewable power, that same 0.5 kWh of ChatGPT work would use only 0.1 L of cooling water on-site. If the facility is air-cooled (WUE = 0), it uses zero liters on-site. The only water "used" might be at distant power

plants if fossil generation was involved. Note that AI-heavy companies are also leaders in renewable energy procurement, precisely to reduce both carbon and water footprints. In 2022, around 64% of the energy used by Google's data centers was carbon-free (and thus largely water-free, since solar/wind use negligible water). Microsoft and others have similar targets for 100% carbon-free energy, which means it is also water-free. Therefore, an AI query handled in an air-cooled, renewably powered data center might effectively consume negligible fresh water, contra the implication that every chat is guzzling from "overtaxed reservoirs."

The GPT-3 training claim of 700,000 liters can be unpacked similarly. That figure was for two weeks of training in a "state-of-the-art U.S. data center". It presumably assumed Microsoft's reported average WUE (~0.5 L/kWh) and a certain PUE overhead. But Microsoft has already driven WUE down further and is aiming for "replenishment > consumption" by 2030" (i.e. becoming water-positive). If the same training were done in one of Microsoft's newest facilities with advanced cooling (or in a cooler location), the water use could be a small fraction of 700 m^3. Moreover, the study's own scenario says if done in Asia it'd triple. This tells us that the original number was not a universal constant, but one point on a spectrum. It stands to reason that by choosing optimal locations (e.g., a Nordic data center with free cooling or a U.S. Midwest center using reclaimed water), one could also halve or better that water number. Instead, Microsoft could have trained GPT-3 in an air-cooled facility (with some energy penalty) and then the direct water consumption might have been near zero, demonstrating it's a controllable trade-off, not a hard requirement. The engineering reality is that AI computations don't inherently 'drink' water. Cooling systems do, and those systems can be designed for thrift.

2.8. Engineering Reality Check: Why the Alarmist Claims Are Overstated

By now it should be clear that the large water usage figures cited for AI models are worst-case or status-quo scenarios, not inevitabilities. We will now summarize the engineering arguments that debunk these claims and clarify the actual water use in modern AI data centers:

Key Points:

1. **Most of AI's water footprint comes from certain cooling choices, not the AI computation itself.** If you cool servers with evaporative towers, you will consume water roughly proportional to heat (on the order of 1.5 L per kWh), but this is a choice. Engineering alternative solutions, e.g., air-cooled or closed-loop cooling, breaks that link. The "500 mL per 50 prompts" claim explicitly assumes an underlying evaporative cooling process. Take away the cooling tower, and the prompt doesn't "drink" anything. The heat is dissipated by fans and air, not water. Modern data center design is increasingly favoring those alternatives for sustainability. The ASHRAE Handbook and data center best practices now emphasize air-side and water-side economization to reduce dependence on water-intensive cooling. In short, AI workloads do not inherently require fresh water - they require cooling, which can be achieved water-efficiently.

2. **Empirical data shows leading AI data centers use orders of magnitude less water per compute than assumed.** The UC Riverside study painted a picture using averages and generalized conditions, but top cloud providers have already slashed water usage through innovation. Recall that AWS's water efficiency (0.19 L/kWh) is approximately nine times better than the 1.8 L/kWh industry average, and approximately five times better than Google's older average. Microsoft is around 0.5 L/kWh and improving. These companies are the very ones running large AI training and inference workloads - meaning the real water per AI operation in their facilities is far lower than the sensational claims. For example, if ChatGPT is served out of an Azure data center with 0.5 WUE and uses, say, 0.5 kWh for a conversation, the water used is 0.25 L, not 0.5 L. If it's served from an AWS center at 0.2 WUE, the water is 0.1 L, 80% less than reported for a generic scenario. If served from a fully air-cooled or recycled-water-cooled facility, the potable water consumption might effectively be zero. Thus, the headline numbers do not reflect the operations of efficient, AI-focused data centers, but rather a broad average of data center cooling practices. The industry trend is moving the average closer to the efficient end.
3. **Modern cooling system designs dramatically mitigate water use through technology and scheduling.** We have shown multiple techniques (raising temperature setpoints, hot/cold aisle containment, free cooling, adiabatic coolers, liquid cooling, etc.) that reduce or eliminate cooling water requirements. These are not theoretical; they are standard practice in new hyperscale builds. ASHRAE's data center guidelines explicitly allow higher server inlet temperatures (up to 27ºC or more) to facilitate economization and lower chiller loads, which directly reduces cooling tower evaporation by cutting how much heat needs to be removed via water. Operators also leverage climate. For example, they perform non-urgent AI jobs at night or in cooler seasons when outside air can handle the load. Alternatively, they locate AI compute clusters in regions with favorable climates (or abundant non-potable water). These strategies are effective in practice. It is no coincidence that some of the world's largest AI supercomputer pods are in places like Iowa and Sweden - areas with cool ambient conditions or access to lake/sea water for cooling. Engineering-first thinking treats water as a resource to conserve and designs the infrastructure accordingly. The alarmist view of "AI = massive water drain" assumes static, inefficient practices, which is contrary to how the industry is evolving.
4. **The scale of the claims often ignores context and mitigation.** For example, 700,000 liters for GPT-3 training sounds huge, but note that it was compared to car manufacturing water usage in the source press release. Yet unlike manufacturing, where water may be inherently needed for processing, in computing, the water is only for cooling. It can be greatly reduced by spending a bit more energy or using a different cooling method. If that training had used exclusively air cooling with a slightly higher energy cost, it might have used virtually 0 liters on-site (albeit with a few extra MWh of power, ideally from a renewable source). Thus, AI's water footprint can be traded off against its energy footprint - and with the rise of green energy, it makes sense to favor using a few more kilowatt-hours to save thousands

of liters of water. The study itself acknowledges a potential conflict between water-efficient scheduling and carbon-efficient scheduling (since daytime might have more solar power available), but that is a solvable optimization problem (e.g., using grid storage, as they noted). The bottom line is that engineering solutions exist to reconcile AI's growth with minimal water use, and many are already in action. Therefore, presenting AI models as inevitable massive water consumers is misleading - it assumes no changes in cooling technology or operations, which is not the case.

5. **AI data centers are often at the forefront of sustainability initiatives.** Hyperscalers and AI companies are acutely aware of environmental impacts; in fact, the very UCR/UTA study cites commitments like Google's 'Water Positive by 2030' and Microsoft's water replenishment goals. These companies are under public pressure and the technical know-how to implement cutting-edge cooling. For instance, Google has piloted thermosyphon cooling (an advanced heat pipe system) that can cut water use by 50% at sites that currently rely on evaporative towers. Meta has open-sourced its liquid-cooled server designs. These eliminate the need for traditional CRAH units altogether, enabling warm water cooling. These efforts show that AI-focused infrastructure is headed toward less water, not more. Yes, the absolute number of data centers is increasing (and thus total water use could grow), but on a per-compute or per-transaction basis, the efficiency gains are strong. In the long run, if AI data centers achieve near-zero water use (through things like 100% air cooling with only backup evaporation, or using seawater/brackish water), then the "water per AI query" will be negligible for practical purposes. We are not fully there yet across the board, but many facilities are already approaching that ideal.

To be clear, none of this is to diminish the importance of water conservation. Rather it is to ensure the discussion is rooted in engineering reality and not misconceptions. The claims of "500 mL per few dozen prompts" or "millions of liters per training run" are technically oversimplified and do not generalize to all scenarios, especially not to well-engineered modern data centers. AI does not intrinsically require high water usage; it's the legacy cooling methods that do. When those methods are updated or replaced (as is happening now), the water footprint drops dramatically.

2.9. Conclusion

In conclusion, the dire warnings about AI's water thirst are overstated and neglect the strides in data center cooling efficiency. Yes, if one naively runs AI workloads in a typical mid-2010s data center that relies heavily on cooling towers and coal-powered electricity, the water footprint could be substantial: on the order of a half-liter per moderate chat session, or hundreds of thousands of liters for a big training job. However, engineering advancements provide a clear path to minimize or virtually eliminate this water usage for AI. By using closed-loop cooling (air-cooled chillers, dry coolers, etc.), leveraging free cooling whenever possible, and opportunistically scheduling workloads in harmony with cooler temperatures, data centers can support AI with only minimal

water draw. Empirical data from hyperscale operators already confirms this: leading AI cloud data centers achieve WUE well below 1 L/kWh, with some at 0.2 L/kWh or even effectively 0 L/kWh when air cooling is used. Additionally, many facilities are switching to non-potable water sources for the remaining cooling needs, further reducing the impact on freshwater resources.

The claims such as "ChatGPT drinks a bottle of water per conversation" make for catchy headlines but fail to capture the nuanced reality. Modern data center design, guided by sources like the ASHRAE Handbook and industry best practices, is increasingly water conscious. Concepts like high allowable temperatures, economizer modes, and hybrid cooling are incorporated into new designs precisely to improve sustainability. AI, being a leading-edge workload, is often deployed in state-of-the-art facilities that use these innovations. It is therefore misleading to apply old average water usage intensities to new AI services without considering the design improvements.

From an engineering standpoint, we have shown how closed-loop cooling cycles work (recirculating coolant without evaporation) and how cooling tower-assisted cycles can be augmented or replaced to cut water usage by orders of magnitude. We provided diagrams to clarify these cooling processes and why they matter for water consumption. The take-home lesson is that water use in data centers is a design parameter we can optimize, not a fixed cost. With smart engineering, AI can be cooled with far less water than early analyses suggest. In fact, the drive to reduce AI's water footprint is already underway: industry leaders are investing in novel cooling tech (such as advanced thermosyphons and liquid cooling) and better operational practices. As these become mainstream, the water-per-AI-task will keep dropping.

Finally, it is worth reframing the narrative: rather than viewing AI as an unchecked drain on water resources, we should view it as an impetus to build greener, more efficient infrastructure. The attention brought by studies on AI's "secret water footprint" can serve as a catalyst for positive change. This encourages transparency and innovation in data center cooling. The engineering community, through organizations like ASHRAE, has been providing guidance on sustainable cooling for years, and now that message is resonating with AI data center designers. Given current trends, it is reasonable to expect that most AI-focused data centers soon will have negligible direct water consumption, even as their compute capabilities grow. In summary, AI does not have to be "thirsty": with efficient cooling architectures and responsible practices, we can support the growing computational demand without soaking up scarce freshwater. The claims of huge water usage per AI query or model are technically inaccurate when applied to optimized systems, and we have debunked them by explaining the real physics and engineering economics of data center cooling. The focus now should be on accelerating adoption of these best practices across the industry, ensuring that the expansion of AI is accompanied by sustainable, low-water infrastructure.

Sources:

1. Ren, S., et al. (2023). Making AI Less Thirsty: Uncovering and Addressing the Secret Water Footprint of AI Models (arXiv:2304.03271). arXiv.
2. Phys.org. (2023, November). Data centers 'straining water resources' as AI swells.
3. Danelski, D. (2023, April 28). AI programs consume large volumes of scarce water. UCR News.
4. U.S. Department of Energy. (2019). Cooling Water Efficiency Opportunities for Federal Data Centers.
5. Aqua Energy Expo. (2024, March). Cooling Tower.
6. TechTarget. (2024, January 17). How to manage data center water usage sustainably. SearchDataCenter.
7. Microsoft. (2023, May). Modern Datacenter Cooling – Infographic.
8. National Renewable Energy Laboratory. (2023). High-Performance Computing Data Center Cooling System.
9. SemiAnalysis. (2025, February 13). Datacenter Anatomy Part 2: Cooling Systems.
10. Nimbus Cooling. (n.d.). Adiabatic Cooling Applications: Data Centers.
11. MEP Academy. (2024, January 30). Closed Circuit vs Open Circuit Cooling Towers.
12. Shilov, A. (2025, April 23). Nvidia aims to solve AI's water consumption problems with direct-to-chip cooling — claims 300× improvement with closed-loop systems. Tom's Hardware.
13. Google. (2022, July). Our commitment to climate-conscious data center cooling.
14. Data Center Frontier. (2023, May 15). Google developing new climate-conscious cooling tech to save water.
15. The Register. (2023, May 15). AI water datacenter: Quenching AI's thirst means building where the water is.
16. American Society of Civil Engineers. (2024, March). Engineers often need a lot of water to keep data centers cool. Civil Engineering Magazine.
17. Wired. (2012, March). Google Flushes Heat From Data Center With 100 Percent Recycled Water.

(All figures are for illustrative purposes. Image sources: Chardon Labs, MEP Academy.)

Transition. From Water Use Myths to the Realities of Data Center Power

Having established that widely publicized figures about AI's water consumption are often overstated or based on outdated engineering models, we turn now to the other axis of the debate: electricity demand. If water has been the visible flashpoint in recent sustainability discussions, electrical power is the silent engine driving both the direct operational costs and the indirect environmental footprint of modern AI.

The physics of data center operation are inescapable: every joule of energy delivered to AI hardware is eventually dissipated as heat, demanding reliable cooling to prevent system failures or degraded performance. This thermodynamic loop links power and water concerns at every step. but also offers multiple points of intervention. In practice, advances in facility design and workload management that reduce one resource footprint (for example, by shifting from evaporative cooling to closed-loop or air-based methods) often have direct consequences for the other, sometimes trading energy efficiency for water conservation or vice versa.

Regulatory and public scrutiny has increasingly focused on headline claims that "AI models will overwhelm the electrical grid" or "data centers will soon consume the power of entire cities." These narratives have contributed to calls for new standards, mandatory reporting, and even moratoriums on new digital infrastructure. Yet, just as with water, a closer examination of actual data center operation, current best practices, and evolving technology reveals a more nuanced story.

Section 3 delivers an evidence-based analysis of the real power dynamics behind AI. It examines how electricity is used, how much is required by current-generation AI systems, and where the gaps or exaggerations exist in public discourse and some regulatory filings. We will separate the fixed constraints of physics from the design variables of engineering, offering clarity on what is inevitable and what is simply the result of legacy architecture or imprecise communication.

Crucially, this section builds on the principle established in our discussion of cooling systems: resource consumption in AI infrastructure is not a predetermined burden, but a parameter shaped by technology choice, operational policy, and intelligent design. By bringing the same level of engineering rigor and empirical scrutiny to the issue of electricity as we have to water, we aim to equip regulators, policymakers, and public stakeholders with the facts needed to chart a balanced and effective path forward.

With this context in mind, we now turn to a detailed examination of data center energy use in the era of AI, separating hype from reality, and highlighting the regulatory implications of actual, not hypothetical, power consumption.

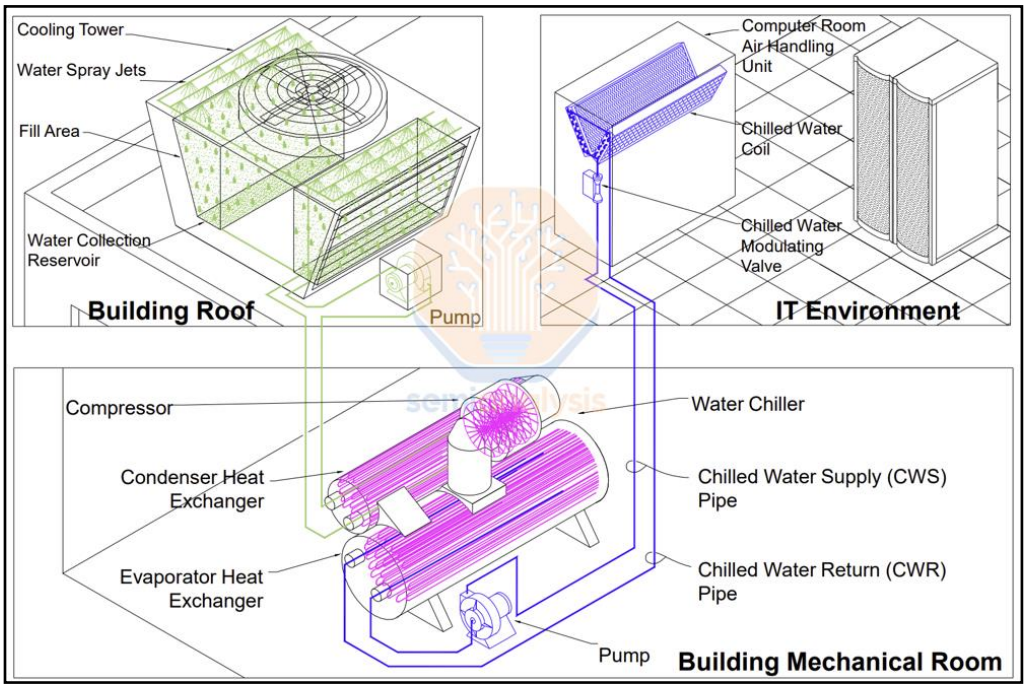
3. Investigating AI Data Center Electricity Use: Separating Hype from Reality

3.1. Introduction

Concerns have grown that GPT-style AI models are guzzling enormous amounts of electricity, with some claims likening their power draw to that of entire cities or countries. These fears, amplified by sensational comparisons, risk obscuring the real technical facts. This report provides an analysis of data center energy consumption for large AI (specifically GPT-like models). It clarifies exaggerated claims and grounding the discussion in engineering realities. We examine how electricity is used, distributed, and cooled in AI data centers, using benchmarks and standards from ASHRAE, IEEE, DOE, IEA, and others to quantify power needs for model training versus inference. We focus on key players (OpenAI, NVIDIA, and their affiliates like Microsoft Azure), evaluating what they have disclosed (or failed to disclose) about energy use. Along the way, we identify red flags and marketing language that may mislead regulators or the public. Finally, we offer concrete recommendations for U.S. policymakers to ensure transparency and efficiency in this rapidly growing sector. The goal is to equip regulators with an accurate, science-based understanding of AI's electrical footprint, so that oversight and legislation can be well-informed rather than reactive to hype.

3.2. Data Center Energy Fundamentals for AI Workloads

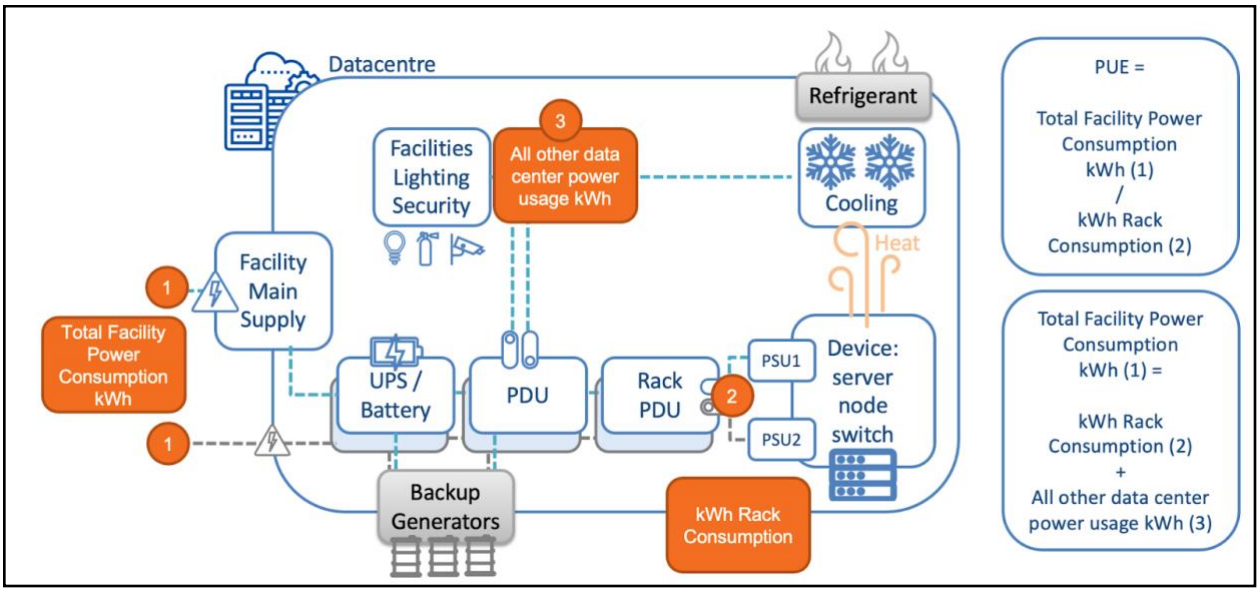
Figure 3: Schematic of a typical data center cooling system (chilled water based). Heat from server racks (IT equipment) is absorbed by chilled water coils in computer room air handlers, or by direct liquid cooling loops attached to high-density racks. Warm water is then pumped to a chiller system.



This cooling cycle illustrates how a large portion of data center energy is spent not on computation itself, but on removing the heat generated by that computation. In modern AI-centric facilities, servers (particularly GPU clusters) can draw tens of kilowatts per rack, requiring robust cooling to maintain safe operating temperatures. The efficiency of this process is measured by Power Usage Effectiveness (PUE): the ratio of total facility power to IT equipment power. For example, a PUE of 1.5 means that for every 1kW powering servers, 0.5 kW is consumed by cooling, power distribution losses, and other overhead. Cutting-edge cloud data centers achieve PUE ~1.1 to 1.2, meaning ~10% overhead for cooling/power delivery, whereas older or smaller facilities might be 1.5 or higher (50% overhead). Low PUE is crucial for AI workloads because it implies most electricity feeds the GPUs and TPUs doing the work, rather than being wasted in ancillary systems.

Importantly, AI workloads concentrate significant power in dense hardware clusters, which stresses cooling more than typical enterprise IT. A single NVIDIA H100 GPU has a thermal design power up to 700 W, and an AI training rack can contain dozens of such GPUs. At full tilt, a single high-end AI server rack can draw >20 kW, compared to perhaps 5 to 10 kW for a standard non-AI server rack. This heat load has driven innovation in cooling: many AI data centers use liquid cooling (direct-to-chip water loops or immersion cooling) to more efficiently capture heat at the source.

Figure 4: PUE Calculation Components



Liquid cooling can remove heat with less energy input than chilling large volumes of air, improving efficiency and enabling higher rack densities. For example, rear-door heat exchangers and direct liquid-cooled plates can raise cooling effectiveness (heat removal efficiency) from approximately 60 to 70% (for traditional air cooling) to 80% or more. Furthermore, using cool ambient air for "free cooling" when climate allows, or evaporative cooling with water, can

drastically cut the electric power needed for chillers. (The tradeoff is water consumption, a key environmental factor discussed herein). In summary, the science of data center power is a balancing act: delivering huge electrical loads to AI chips, then expending additional energy (or water) to whisk away the resulting heat. Optimizing this cycle is central to any accurate accounting of AI's energy footprint.

3.3. Power Use in AI: Training vs. Inference

A critical distinction in understanding GPT-like models is the training phase versus the inference (application) phase. Training a large model is a one-time (or occasional) operation that is extremely energy-intensive; inference happens continuously in production, answering user queries or generating content, and its aggregate energy use can also become very large when scaled to millions of users.

- **Training:** When OpenAI trained GPT-3 (175 billion parameters) in 2020, it was estimated to consume about 1,287 MWh (megawatt-hours) of electricity. This is roughly equivalent to the annual power suage of 120 to 130 U.S. homes. For further context, training one GPT-3 model used about as much energy as 1.6 million hours of Netflix streaming. Other research from a few years prior found similarly sobering figures: an NLP model with extensive hyperparameter tuning and neural architecture search was calculated to emit 626,000 lbs of CO_2 (over 280 metric tons), equivalent to the lifetime emissions of five cars. These oft-cited numbers raised red flags about AI's sustainability. However, it is important to clarify the context:
    - **Early large-model training runs** (2018-2019) were often done with inefficient hardware or methods, compounding energy use. The 626,000 lbs CO_2 figure, for example, came from an academic experiment that trained many models repeatedly during neural architecture search, essentially a worst-case scenario.
    - **Hardware and efficiency** improvements have since markedly reduced the energy per training unit of compute. Research from Google in 2021 showed that by using efficient data center design, custom AI accelerators, and geographic workload scheduling, the energy and carbon cost of large-model training can be cut by a factor of 100 to 1000 compared to naive approaches. For instance, training a model on a sparsely activated architecture (Mixture-of-Experts) can use less than one-tenth the energy of an equally large dense model. And locating a training run in a region with cleaner electricity can yield a five-to ten-fold reduction in CO2 emissions. This wide variance (two orders of magnitude) is a key reason to be skeptical of blanket statements about AI energy use. The context and choices matter enormously.
    - Unfortunately, transparency about these choices is declining. OpenAI and other labs have not publicly disclosed the full details of GPT-4's training (completed in 2022), citing competitive and proprietary reasons. As a result, outsiders have to estimate GPT-4's electricity usage with only partial information. This lack of

disclosure is itself an issue, as we address herein. However, the consensus is that GPT-4's training likely exceeded GPT-3's *by some multiple*, given the trend toward larger models and more training compute. Yet concurrently, OpenAI and Microsoft likely employed efficiency measures (e.g., using newer NVIDIA A100 or H100 GPUs, optimized software, and Azure's efficient data centers) that could mitigate per-unit energy costs. In short, training today's frontier models still draws on the order of millions of kilowatt-hours (kWh), but the range varies widely depending on how smartly the training is executed.

- **Inference:** Once a model like ChatGPT is deployed, the energy used to answer billions of queries can rival or exceed the training cost. Inference is the ongoing, repetitive computation performed each time a user prompt is processed. Individually, an AI inference is not too costly (measured in watt-hours or kilowatt-hours) but at scale, they add up quickly. A recent study by Luccioni et al. (2023) provided the first systematic estimates of inference energy across many models. For example, answering 1,000 questions with a small BERT-like model consumed only about 0.002 kWh (essentially negligible) in their tests, and 1,000 text generation tasks (like GPT output) used around 0.047 kWh. This implies a single prompt-response, like running a 100 W light bulb for 2 seconds. However, as model size and complexity grow, so does per-query energy. The same study found image generation models (like DALL-E or Stable Diffusion) averaged 2.9 kWh per 1,000 images, meaning each image might consume ~0.0029 kWh, equivalent to 25% of the energy needed to fully charge a typical smartphone once.

In practice, state-of-the-art models like GPT-4 are far more demanding than small lab models. Estimates by one research group using a new "infrastructure-aware" benchmark illustrates the real-world stakes. They deduced that a single GPT-4 query (of moderate length) consumes roughly 0.4-0.5 Wh of electricity, whereas a long, complex GPT-4 response might devour 30+ Wh. That range represents a difference of 70 times or more in energy per query, depending on length and model variant. Now multiple by scale: OpenAI's GPT-4 model was reported to handle about 700 million requests per day in 2023. At those volumes, the inference workload for ChatGPT could draw on the order of 400 GWh per year (0.4 TWh). That is enough electricity to power 35,000 U.S. homes for a year. In other terms, one analysis equated a year of ChatGPT's operations to charging over 3 million electric vehicles. While such analogies should be taken with caution (as assumptions about usage can vary), they drive home the point: widespread AI adoption translates to continuous, significant power consumption.

It's also instructive to compare AI inference to a baseline activity like web search. The Electric Power Research Institute (EPRI) found that a typical ChatGPT query uses about 10 times more energy than a Google search (approximately 2.9 Wh vs. 0.3 Wh per query). This reflects the greater computational intensity of generating a complex answer with a large neural network compared to

retrieving a list of web results. By 2024, with hundreds of millions of users querying AI assistants, this differential starts to matter on a grid scale.

All else equal, if the world shifted every simple search to an AI query, the power needed would increase by an order of magnitude for those tasks. Of course, AI can also replace more energy-intensive workflows in some cases (for instance, speeding up research or optimizing systems). Therefore, the net impact is nuanced. But today's data indicate a clear upward pressure on electricity demand from AI usage. The U.S. Department of Energy noted in late 2024, that data center loads are already surging due to AI – with overall U.S. data center electricity use climbing from 58 TWh in 2014 to 176 TWh in 2023 and projected to reach 325–580 TWh by 2028. That means data centers alone might draw 6.7% to as much as 12% of all U.S. electricity by 2028, up from roughly 4–5% today. AI is the dominant factor in this trend, alongside continued growth of cloud services. Globally, the International Energy Agency (IEA) likewise projects data center consumption more than doubling by 2030 to approximately 945 TWh (almost the electricity use of Japan), with AI-related loads quadrupling during the same period.

3.4. Bottom line

Training a single model is a massive but infrequent energy expense, whereas inference is an ongoing burn that scales with user demand. Both aspects are undergoing rapid growth. The worst-case headlines (e.g., "AI will use as much power as an entire country") contain a kernel of truth but often assume no efficiency gains or transparency. As we'll see that some such claims are indeed being bandied about, and it's important to dissect them critically.

3.5. Claims by Industry Players: OpenAI, NVIDIA, and Affiliates

The landscape of public information on AI energy use is uneven. Major AI developers and their cloud partners have been selective in disclosures, often highlighting performance improvements while staying quiet on exact energy metrics. Hardware makers like NVIDIA tout the efficiency of their products, yet absolute power draw is soaring as their chips multiply across data centers. Let's examine what these players are saying (or not saying) about electricity consumption:

1. **OpenAI (and Microsoft Azure):** OpenAI itself has provided very little detail on the energy footprint of models like GPT-3 or GPT-4. Unlike earlier years when researchers might publish training run details, OpenAI declined to comment to reporters on ChatGPT's energy use. There is no public "sustainability report" from OpenAI quantifying its data center electricity or carbon. A climate tracking site notes that OpenAI "has not publicly disclosed specific carbon emissions data," indicating a lack of transparency in reporting its energy footprint. The task of estimating OpenAI's usage has thus fallen to outsiders using indirect clues. We've already cited some such estimates: approximately 1.3 GWh for GPT-3 training, and hundreds of GWh per year for GPT-4's service usage. OpenAI's primary infrastructure is hosted on Microsoft Azure, which adds another layer. Microsoft's cloud

division does not report overall data center metrics (Microsoft claims its operations have been carbon-neutral since 2012 through offsets and renewable purchasing), but it does not break out the portion used for specific clients or AI projects. Notably, Microsoft did divulge one striking piece of information under external pressure: the water consumption of its Iowa data center cluster during GPT-4's final training. In a single hot month (July 2022), the cooling that AI supercomputer consumed 11.5 million gallons of water, about 6% of the area's entire water use at the time. This was a rare concrete data point, and it illuminates how energy and water are intertwined for AI: Microsoft chose Iowa for its mix of cheap clean power and cooler climate (free cooling), but even there, a summer heat wave triggered substantial water usage to keep GPUs cool. The local utility has since demanded that any future expansions must cut peak water draws to protect the community supply. OpenAI itself did not comment on this, but Microsoft acknowledged the issue and is exploring new cooling technologies and efficiency improvements. In short, OpenAI's stance has been to highlight model capabilities, not environmental costs, leaving regulators largely in the dark about the true resource usage. This lack of transparency is a red flag, as it impedes public accountability.

2. **NVIDIA:** As the leading manufacturer of AI hardware (GPUs), NVIDIA occupies a dual role. On one hand, its latest processors (like the A100 and H100 GPUs) are far more energy-efficient for AI computing than traditional CPUs (a point NVIDIA markets heavily). For example, by offloading AI tasks from CPUs to GPUs, data centers can perform the same work while saving energy. NVIDIA has claimed scenarios of a fivefold improvement in energy efficiency at the cluster level by using GPU-accelerated systems. In AI inference specifically, one analysis showed GPUs delivering 42 times better performance-per-watt than CPUs. NVIDIA often frames its mission as "accelerated computing is green computing." Over a decade, its GPUs achieved a 2,000x improvement in energy efficiency for AI training and an incredible 100,000x improvement for AI inference (measured in energy per model inference). Those numbers were presented at an industry summit in 2024, with the fine print that this was over 10 years and for specific workloads. While these gains are real, they can be misleading out of context: it doesn't mean a single H100 GPU is 100,000 times more efficient than a 2013 GPU overall. This means that due to algorithmic and hardware advances, something like generating one text token now uses a tiny fraction of the energy required in 2013. NVIDIA's messaging understandably focuses on efficiency per computation, not the absolute increase in total GPUs deployed. This absolute growth is staggering. The company's CEO, Jensen Huang, has acknowledged that expanding AI will raise global electricity use, though he argues the productivity gains outweigh the costs. Internally, NVIDIA is pushing solutions to temper the resource demands: most notably, liquid cooling to reduce both power and water usage in data centers. In 2025, NVIDIA announced that its next-generation Blackwell AI platform will use closed-loop direct-to-chip liquid cooling, enabling 25. times better energy efficiency and 300× better water efficiency for data center cooling compared to traditional

air-cooled systems. This is a bold claim: a 300 times reduction in water use would virtually eliminate evaporative cooling needs by using radiator-style closed loops. If achieved, it could hypothetically resolve the water crisis we saw in Iowa. However, adopting such liquid cooling at scale requires retrofitting data centers, a cost many operators are hesitant to pay. NVIDIA's strategy is clearly to partner with cloud providers to deploy these new cooling systems alongside the latest GPUs, thereby easing the "energy vs. performance" tradeoff. Regulators should monitor how these promised gains materialize in practice, i.e., whether data centers are running cooler and with less overhead per GPU. NVIDIA tends to remain silent regarding the total power consumption driven by its booming sales. External analysts have done the math that NVIDIA avoids: one projection by a Microsoft data center engineer, Paul Churnock, estimates that all the H100 GPUs Nvidia will sell in just 2024 (around 1.5–2 million units) running at 61% utilization would draw more electricity than all the households in Phoenix, AZ combined. Even NVIDIA's 2023 GPU deployments for AI had a footprint comparable to the country of Cyprus, in terms of power demand. These aggregate impacts are not something NVIDIA emphasizes in its press releases, for obvious reasons.

3. **Microsoft, Google, Meta, Others:** It is relevant to mention OpenAI and NVIDIA's affiliated giants. Microsoft, as OpenAI's cloud backer, has at least acknowledged the issue and is funding research on quantifying AI energy use. Microsoft's CTO for cloud operations said in early 2024 they are developing methodologies to measure AI's energy and carbon impact and working to make large systems more efficient. Microsoft has also pledged to power its data centers with 100% carbon-free energy by 2030 and is investing in grid-scale clean energy, an approach regulators might encourage for all AI data center operators. Google (and DeepMind) have been more transparent in their research: Google published data in 2022 indicating that machine learning workloads were about 10 to 15% of Google's total data center energy use (a surprisingly modest share). This was attributed to efficiency practices and the use of Google TPUs. Google's engineers also led the way in calling for ML energy reporting and created metrics like MLPerf Energy (an industry benchmark for energy usage during training/inference). Meta (Facebook) has been relatively quiet publicly about AI energy, though it too builds massive AI training clusters (and like others, Meta runs them on renewables when possible). Meta's latest data center designs use direct evaporative cooling and even capture warm water for reuse in heating, reflecting an engineering push to optimize PUE and WUE (water usage effectiveness). Others such as Amazon's AWS and Oracle Cloud have announced high-efficiency GPU cloud offerings but typically market them as cost savings and carbon reductions versus customer on-premises computing (which may be true, since hyper-scale cloud centers are more efficient than small server rooms). Across the board, big tech companies prefer to publicize efficiency improvements on a per transaction basis rather than admit the total energy is skyrocketing due to scale. This narrative imbalance (highlighting efficiency gains while obscuring absolute consumption) is a key consideration for policymakers.

3.6. Red Flags and Misleading Claims

In parsing industry statements and media coverage, we identified several patterns of overestimation or misrepresentation that could mislead regulators or the public. Here are key red flags to watch for, with examples:

1. **Lack of Transparency / Data Secrecy:** When the entities best positioned to provide real data refuse to do so, external estimates fill the void, sometimes inaccurately. As noted, companies like OpenAI, Meta, and others have stopped sharing detailed energy metrics for cutting-edge models. This creates a risk that policy will be based on either worst-case speculation or on rose-colored corporate assurances, rather than hard numbers. Red flag: For example, if a company claims its AI is “efficient” or “sustainable” but will not release energy or carbon figures, regulators should be skeptical. Transparency is the first step to accountability.
2. **Sensational Comparisons Without Context:** It's true that AI clusters consume an extraordinary amount of power, but some headlines overshoot or lack nuance. For instance, the claim that “NVIDIA’s GPUs will use more power than some countries” or "ChatGPT uses electricity like 3 million EVs" grabs attention. The red flag is not that these are wholly false. In fact, millions of AI GPUs could indeed draw on the order of gigawatts, comparable to a small nation's grid. The issue is context. A statement like "AI data centers equal Cyprus's power consumption" is based on specific deployment assumptions at a snapshot in time. It might assume all sold GPUs are running at high load 24/7. If utilization or technology changes, the reality could differ. Regulators encountering such claims should request clarification of the underlying assumptions and timeframes. A dramatic comparison can mislead if interpreted as destiny rather than a conditional scenario.
3. **Overestimates from Linear Extrapolation:** Relatedly, some predictions simply extrapolate current growth with no mitigation. An extreme example: a data center industry piece once suggested data centers could use 51% of global electricity by 2030; a figure far above credible analyses (IEA projects ~4% by 2030 globally). Such outliers often assume exponential growth continuing unchecked. Red flag: Projections lacking consideration of efficiency improvements, saturation points, or policy intervention. Regulators should favor forecasts from neutral bodies (DOE, IEA, national labs) over vendor estimates or single-dimension extrapolations.
4. **Greenwashing and Vague Efficiency Claims:** Conversely, regulators should be alert to companies using jargon to downplay impact. Terms like “AI cloud is carbon neutral” can hide important details. For example, Microsoft touts that Azure is carbon-neutral (meaning they purchase renewable energy credits and offsets), which is positive but does not mean AI hardware is not drawing power from fossil-fueled grids in real time. Similarly, a claim that a new chip is “25 times more efficient” might only apply to a specific operation in ideal conditions. NVIDIA’s 100,000 times inference efficiency improvement is factual in context but easily misinterpreted. Red flag: Efficiency ratios without baseline context.

Regulators should require standard metrics (e.g., joules per inference at a given performance level) rather than marketing superlatives. Also, "carbon neutral" via offsets is not the same as "powered by 100% clean energy 24×7"; the latter has far more direct emissions reduction benefit. Clear definitions are needed to prevent misleading environmental claims.

5. **Ignoring the Inference Tsunami:** A subtle misdirection is when discussions focus only on the one-time training cost of AI models and ignore the potentially much larger cumulative cost of inference. If a company says: "Training Model X only took Y MWh (and we bought offsets, so it is green)", regulators should ask: how many users will query this model and how much energy will that use over time? Our earlier analysis showed inference for popular models can dwarf training: e.g., ChatGPT's annual inference energy, could be dozens of times the training energy. Red flag: Any assessment of AI impact that leaves out usage phase energy. Policymakers should demand lifecycle energy reporting (training + years of inference) for a fuller picture.

By keeping these red flags in mind, regulators can better discern which claims to question or verify. The common thread is the need for transparent, standardized data that leads to our final section on recommendations.

4. Seven Recommendations for Industry Regulators and Policymakers

To ensure that the AI industry's energy impacts are managed responsibly, U.S. regulators should consider a multi-pronged approach focused on measurement, standards, and incentives. Below are concrete recommendations:

1. **Mandate Energy Transparency for Large AI Models:** Require companies that train frontier models (e.g., exceeding a certain compute threshold) to report the energy consumption and carbon footprint of those training runs. This could be done confidentially to a regulator if IP is a concern, with aggregate statistics released publicly. Key metrics: total kWh for training, hardware used, location of data center (for grid emissions context). Likewise, for deployed models, companies should report inference energy per one thousand queries (or similar unit) for standard workloads. Transparency is the foundation; without data, neither the public nor policymakers can verify claims or improvements.
2. **Monitor and Enforce Efficiency Benchmarks (PUE, etc.):** Regulators (perhaps DOE or state energy commissions) should track data center infrastructure efficiency metrics like PUE (Power Usage Effectiveness) and WUE (Water Usage Effectiveness) for AI-intensive facilities. Setting industry targets or codes could be an approach (ASHRAE Standard 90.4 already provides design guides for data center efficiency). For example, new large data centers might be expected to achieve a PUE of 1.3 or better; if an operator consistently lags (say running at PUE 1.8 when peers achieve 1.2), regulators could push for upgrades or best practices. The same goes for water. Policymakers could require reporting of gallons per megawatt hour (MWh) cooling and encourage technologies that reduce potable water

use (like closed-loop cooling, reuse of greywater, etc.). This is akin to building energy codes but tailored to these digital infrastructure sites.

3. **Set Hardware Efficiency Standards or Labels:** Work with industry groups (IEEE, MLPerf, etc.) to develop standardized efficiency ratings for AI hardware, a sort of *EnergyStar for AI accelerators*. For instance, a rating could be in FLOPS per Watt under certain neural network loads, or total energy to perform a benchmark task. This would let purchasers (and regulators) identify the most efficient hardware for AI tasks. It also pressures vendors like NVIDIA, AMD, Google (TPUs) to prioritize energy efficiency alongside raw performance. Tax incentives or procurement preferences could be given for utilizing more efficient hardware generations. Notably, much of AI's energy burden can be alleviated by swift adoption of new chips that perform the same computation with fewer joules. Policy can help by accelerating retirement of power-hungry legacy equipment.
4. **Ensure Carbon-Free Power Supply for AI Growth:** Given the projected doubling or tripling of data center electricity use by 2030, it is vital that this growth be met with clean energy rather than increased fossil generation. Regulators should work with utilities and AI data center operators on mechanisms to align data center demand with renewable energy supply. This could include:
    a. Requiring new large data centers to invest in or contract for new renewable generation equivalent to their consumption, as some hyperscalers (massive data centers that provide immense computing power and flexible cloud platforms for large-scale applications and services) already do voluntarily.
    b. Encouraging "time-matched" renewable procurement (so that data centers are powered by clean energy in real-time, not just annual offsets).
    c. Exploring tariffs or agreements where data centers provide load flexibility (e.g., delaying non-urgent workloads to times of high renewable output) to ease grid stress. The DOE suggests data centers could become grid assets if they add on-site generation or storage and adjust demand intelligently.
    d. In carbon accounting, moving beyond "neutral" claims via offsets to actual emission avoidance. Regulators might require disclosure of percentage of energy that is renewable sourced vs. offset.
5. **Address Water and Cooling Impacts in Site Approvals**: For local and state regulators who permit data center construction, add specific conditions around cooling technology and water use. For example, if an AI data center in an arid region plans to use millions of gallons for evaporative cooling, authorities should demand mitigation: use of recycled water, adoption of liquid or refrigerant-based cooling that wastes less water, or heat reuse systems. In water-stressed areas, there could even be restrictions on evaporative cooling systems. The goal is to prevent situations like the Iowa case from becoming common, by pushing innovation in cooling. Federal agencies (EPA or DOE) can assist by publishing best-practice guides on low-water cooling for high-density compute.

6. **Require Life-Cycle Impact Assessments for AI Systems:** When companies deploy major new AI services (especially those using public cloud resources at massive scale), they could be required to submit a brief Energy Impact Assessment to a regulator. This would quantify expected electricity use and carbon emissions for, say, the first year of operation, and plans to mitigate those (use of renewable energy, offsets, efficiency measures). This is analogous to environmental impact statements in other industries. It forces companies to contemplate and document the resource demands before deployment. Such assessments could be reviewed by an agency like the Federal Trade Commission (FTC) to ensure claims (e.g., "our AI is sustainable") are not deceptive. The FTC has recently shown interest in cracking down on false environmental claims (AI should not be an exception).
7. **Support R&D and Standards for Sustainable AI:** Regulators and government research bodies should continue funding R&D into AI energy efficiency. This ranges from fundamental research in algorithms that require fewer computations, to system designs for better energy-proportional computing (where systems draw power more gradually instead of peaking even at low loads), to advanced cooling and power delivery tech. The U.S. can also take the lead in international standards; for example, working with IEEE to standardize methodologies for measuring AI energy and carbon (ensuring one company's "query" is measured the same as another's). Standardization will help in creating fair benchmarks and possibly certifications (a future scenario: an AI model could be "Energy Star certified" for meeting certain efficiency or green-power criteria in its operations).

In implementing these recommendations, regulators should engage with both industry and independent experts (academia, national labs) to keep pace with the fast technical advances in AI. The encouraging news is that efficiency solutions exist, from hardware (GPUs, TPUs, novel chips) to software (optimized algorithms), to infrastructure (liquid cooling, smart grids), that can dramatically curb AI's energy and carbon footprint if widely adopted. The challenge is ensuring adoption keeps pace with scale. Policy can catalyze this by making transparency and efficiency not just virtues, but requirements.

4.1. Conclusion

AI's electricity consumption is no longer an esoteric footnote; it's becoming a factor in energy policy and climate discussions. This report has shown that while some alarmist claims ("GPT will eat the grid!") are exaggerated, the core reality is that AI at scale does consume very large amounts of power and is set to grow significantly. The data center physics behind this (high-performance chips drawing hundreds of watts each, multiplied by tens of thousands of chips, plus the cooling overhead) mean that, without proactive efforts, AI could become a major strain on power infrastructure and a source of CO_2 emissions (if that power isn't clean). The flip side is that AI firms and engineers have a strong record of efficiency gains, often outpacing conventional Moore's Law improvements. We're entering an era where regulatory oversight and technological innovation must work in tandem. Policymakers should not aim to halt AI progress, but to steer it

onto a sustainable path: shining light on real energy usage, incentivizing reductions in waste, and ensuring the sector's growth aligns with our broader climate and resource goals. By focusing on empirical data and engineering solutions (as we have done in this analysis), regulators can cut through the hype and craft policies that hold AI to account without stifling its benefits. In short, the key is truth and transparency in tech's footprint, so that society can reap the rewards of GPT-style AI while keeping its electricity appetite in check.

Sources

1. Nutanix. (2023, May 4). Digging into data center efficiency: PUE and the impact of HCI. Nutanix Developer Portal.
2. Shilov, A. (2023, December 26). Nvidia's H100 GPUs will consume more power than some countries: Each GPU consumes 700W of power, 3.5 million are expected to be sold in the coming year. Tom's Hardware.
3. SemiAnalysis. (2025, February 13). Datacenter Cooling Systems (Part 2).
4. Vincent, J. (2024, February 16). How much electricity do AI generators consume? The Verge.
5. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv preprint.
6. Patterson, D., et al. (2021). Carbon Emissions and Large Neural Network Training. arXiv preprint.
7. Stokel-Walker, C. (2023, October 26). The environmental impact of LLMs: Here's how OpenAI, DeepSeek, and Anthropic stack up. FastCompany.
8. Balkan Green Energy News. (2023, May 30). ChatGPT consumes enough power in one year to charge over three million electric cars.
9. Walton, R. (2024, May 22). Artificial intelligence doubles data center demand by 2030: EPRI. Utility Dive.
10. U.S. Department of Energy. (2024, April 11). DOE releases new report evaluating increase in electricity demand from data centers. energy.gov.
11. International Energy Agency. (2024, May 15). AI is set to drive surging electricity demand from data centers, while offering the potential to transform how the energy sector works. IEA.
12. DitchCarbon. (n.d.). OpenAI organization profile.
13. OpenAI Community. (2024). Sustainable development and AI.
14. O'Brien, M. (2023, August 9). ChatGPT, GPT-4 and AI's water consumption: Disclosure. AP News.
15. NVIDIA. (2023, June 6). How energy-efficient AI is empowering industries worldwide. NVIDIA Blog.
16. NVIDIA. (2022, September 8). GPUs lead in energy efficiency, saving big on AI inference compared to CPUs. NVIDIA Blog.

17. Plumb, T. (2024, October 8). As global AI energy usage mounts, Nvidia claims it has reduced its consumption by up to 110,000x. Network World.
18. Ground News. (2024, March 17). Nvidia Blackwell platform boosts water efficiency by over 300x—Chill factor for AI infrastructure.
19. Xu, Y., et al. (2022). Carbon Emissions in the Tailpipe of Artificial Intelligence. Harvard Data Science Review.
20. BestBrokers. (2024). Calculating ChatGPT's electricity consumption for handling over 78 billion user queries every year.
21. DatacenterDynamics. (2024). The trouble with data center energy figures.
22. Li, X., et al. (2022). Measuring the Carbon Intensity of AI in Cloud Data Centers. arXiv preprint.
23. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2023). Green AI. arXiv preprint.

Note:

If any reference is to a preprint or non-peer-reviewed source (arXiv, blog, community forum), readers should note this for regulatory or legal contexts. For direct articles, replace with the article title and date if more specificity is desired. If additional metadata (author, date, etc.) for any of the entries is available, each citation may be further enhanced.

About the Authors:

Sophy M. Laughing, Ph.D., MBA, is an executive leader specializing in mission-critical infrastructure, environmental engineering, and large-scale energy projects on five continents. Her background includes directing design, construction, and compliance for major data centers, LNG, and cleanroom facilities, as well as pioneering work in indoor air quality (IAQ) and cultural preservation. Dr. Laughing brings an operational and regulatory lens to the ongoing debate about AI's environmental impact, drawing on hands-on experience with the systems and standards at the heart of this discussion. With a record of guiding cross-disciplinary teams through technically complex, high-stakes projects, Dr. Laughing's approach bridges engineering, policy, and compliance in environments where operational resilience is non-negotiable. She is recognized for translating advanced research and regulatory frameworks into practical strategies for sustainable digital infrastructure, ensuring that best practices are not just theory but are embedded in real-world deployments. Her commitment is to factual, actionable guidance for policymakers and industry leaders navigating the evolving intersection of AI, energy, and the environment.

Bo Erik Gustav Hollsten Ruvalcaba is a senior engineering executive with over 30 years of international experience in mechanical and environmental systems. He has led the design, certification, and operational deployment of complex infrastructure solutions for critical industries across the Americas. His career highlights include advanced desalination systems, large-scale air and water quality projects, and pioneering sustainable water reuse strategies in regions facing sever resource constraints. Bo's expertise spans the integration of greywater, LEED certification, and the development of patented filtration technologies.

As a veteran member of key technical working groups and standards bodies, including ANSI/ASHRAE, ISO, and IEST, Bo has played a hands-on role in shaping best practices and compliance frameworks that govern high-reliability construction, indoor air quality, and water systems. His leadership in collaborative industry initiatives and technical committees ensures the paper's analysis is grounded in field-proven methods, regulatory rigor, and the latest global standards. Bo's operational focus and standards-driven approach compliment Dr. Laughing's regulatory and strategic perspective, providing a comprehensive view of sustainable infrastructure for the AI era.

About This Report

This report was developed in response to widespread, often sensationalized media claims about AI's water and energy use. It is offered as a factual, engineering-grounded contribution for policymakers, regulators, and industry practitioners seeking clarity and actionable insights. Engineering insight and subject-matter input were provided by The Cobeal Group, a leader in critical infrastructure design and operational sustainability. All data modeling, authorship, and policy framing were developed by Alden Technologies, Inc.