The AI Infrastructure Policy Report 2025

No, AI Doesn’t Drink a Bottle of Water per Prompt

A Policy Briefing on Cooling, Power, and the Design Realities of AI Infrastructure

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

Bo Erik Gustav Hollsten Ruvalcaba
Chief Engineer, The Cobeal Group

Water & power use in AI are design variables

Recent headlines claim AI systems consume half a liter of water per prompt or strain national power grids, but these narratives are based on outdated assumptions and worst-case scenarios. This policy briefing corrects the record with data from real-world engineering.

Today's best-in-class infrastructure uses closed-loop cooling, reclaimed water, and load balancing systems to achieve Water Usage Effectiveness (WUE) as low as **0.19 L/kWh** and Power Usage Effectiveness (PUE) below **1.2**. Contrary to media alarmism, AI's environment footprint is determined by **design choices**, not fixed technological costs.

This briefing lays out:

The origins of the "500 mL per prompt" myth and why it fails scientific scrutiny.

Regulatory recommendations for transparency, efficiency benchmarks, and location optimization.

Why this briefing matters

Policymakers, regulators, and industry leaders are currently making high-stakes decisions based on outdated, worst-case assumptions about AI's environmental footprint. These assumptions (particularly around water use) have made their way into legislative drafts and global media cycles, shaping public pressure and influencing digital infrastructure investment. This report presents an engineering-based correction. By separating media myth from operational fact, it equips stakeholders to regulate and innovate with accuracy, not alarm.

What this report delivers

In addition to correcting public misconceptions, this briefing concludes with a set of forward-looking policy recommendations. These include measurable efficiency benchmarks, transparent disclosure standards, and siting strategies to ensure AI infrastructure evolves sustainably, and credibly.

In a two-part series, *No, AI doesn't drink a bottle of water per prompt*, will cover:

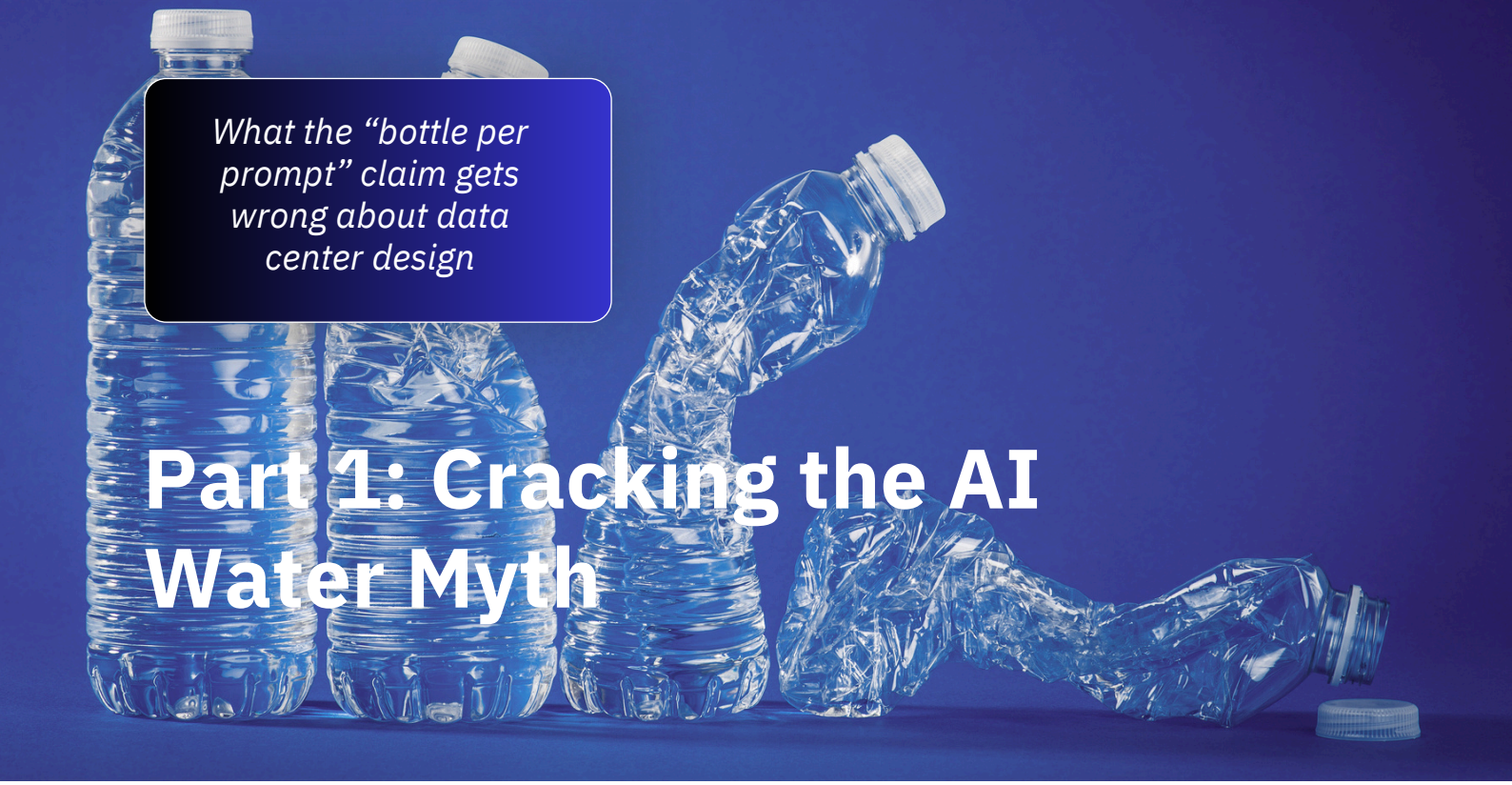
PART 1: CRACKING THE AI WATER MYTH

Tracks the origin of the half-liter bottle of water claim, dissects the thermodynamics behind it, and compares it against the cooling methods used in modern AI-ready data centers. The findings reveal that the "500 mL" figure only applies under legacy evaporative systems, and that state-of-the-art facilities consume an order of magnitude less, or none at all.

PART 2: INTRODUCING THE POWER REALITY

Examines AI's electricity demand and whether AI will overwhelm national grids. We examine how AI training and inference workloads differ in their draw, how those workloads are distributed, and how real-world deployments (especially those using advanced liquid cooling and intelligent scheduling) achieve low Power Usage Effectiveness (PUE).

The conclusion is clear: both water and power consumption are design variables, not fixed outcomes. Regulatory panic is premature, but engineering-based policy is overdue.



What the "bottle per prompt" claim gets wrong about data center design

Part 1: Cracking the AI Water Myth

5 Engineering Flaws in the "500 mL per Prompt" Claim

1-Assumes Outdated Cooling Design (Evaporative Towers Only)

The study bases its estimate on evaporative cooling, which inherently consumes water. But modern AI data centers increasingly use closed-loop, air-cooled, or hybrid systems that use little to no water. AWS, for instance, reports a WUE of just 0.19 L/kWh, not 1-2 L/kWh as implied by the Ren model.

2-Applies Static Efficiency to a Dynamic, Climate-Adapted System

Water use in real-world data centers is **not constant**. Modern operators dynamically adjust workloads and cooling modes (like air-side economization and adiabatic assist) to match environmental conditions, dramatically reducing water use during cooler seasons or at night.

3-Fails to Account for AI-Specific Design Innovations

High-density AI workloads are typically deployed in next-gen facilities designed for GPU heat flux. These often use direct-to-chip liquid cooling with dry coolers or heat reuse loops, not evaporative towers. As a result, water is a last resort, not a default cooling mechanism.

4-Treats Location and Timing as Invariant

The paper treats water intensity as fixed across regions and times of day. But in reality, training jobs can be scheduled in low-humidity, cool climates (like Sweden or Iowa), or run at night, cutting evaporation drastically. This flexibility invalidates fixed "per prompt" water costs.

5-Presents One-Time Peak Estimate as Universal Baseline

The "500 mL per 20-50 prompts" figure stems from worst-case stacking: high energy, evaporative cooling, and fossil-powered electricity. But AI queries today often run in stable, renewably powered, air-cooled infrastructure. There is no universal water cost per prompt - it depends entirely on design.

On-site Water Usage Efficiency (WUE)

Provider	WUE (L/kWh)
AWS (2022)	**0.19**
Microsoft (2022)	**0.49**
Google (2021 est.)	**~1.1**
Industry Avg (legacy)	**~1.8**
Fully air-cooled DC	**0.0**

"Evaporative cooling is optional. Modern AI data centers often use zero-water systems."

– ASHRAE Handbook, Data Center Cooling Standards

The Ren et al. Model Ignores Modern Cooling Practices

The “500 mL per prompt” figure originated from a 2023 preprint by Ren et al. (UC Riverside & UT Arlington), which used a broad-stack estimate combining ChatGPT’s compute load, legacy evaporative cooling, and fossil-fuel grid water use. But modern AI infrastructure doesn’t operate this way. For example, AWS reports a fleetwide on-site WUE of just 0.19 L/kWh, meaning a typical AI query would use a tenth of that figure, or none at all in air-cooled systems

Engineering First Principles Say Otherwise

Water use in cooling is not intrinsic to AI. It is a design parameter. Evaporative cooling uses water, yes - but dry coolers, air-side economization, and liquid-to-air heat exchangers do not. Many modern facilities are already using zero-water cooling, especially during cooler seasons or off-peak hours.

Climate-Aware Scheduling Makes a Measurable Difference

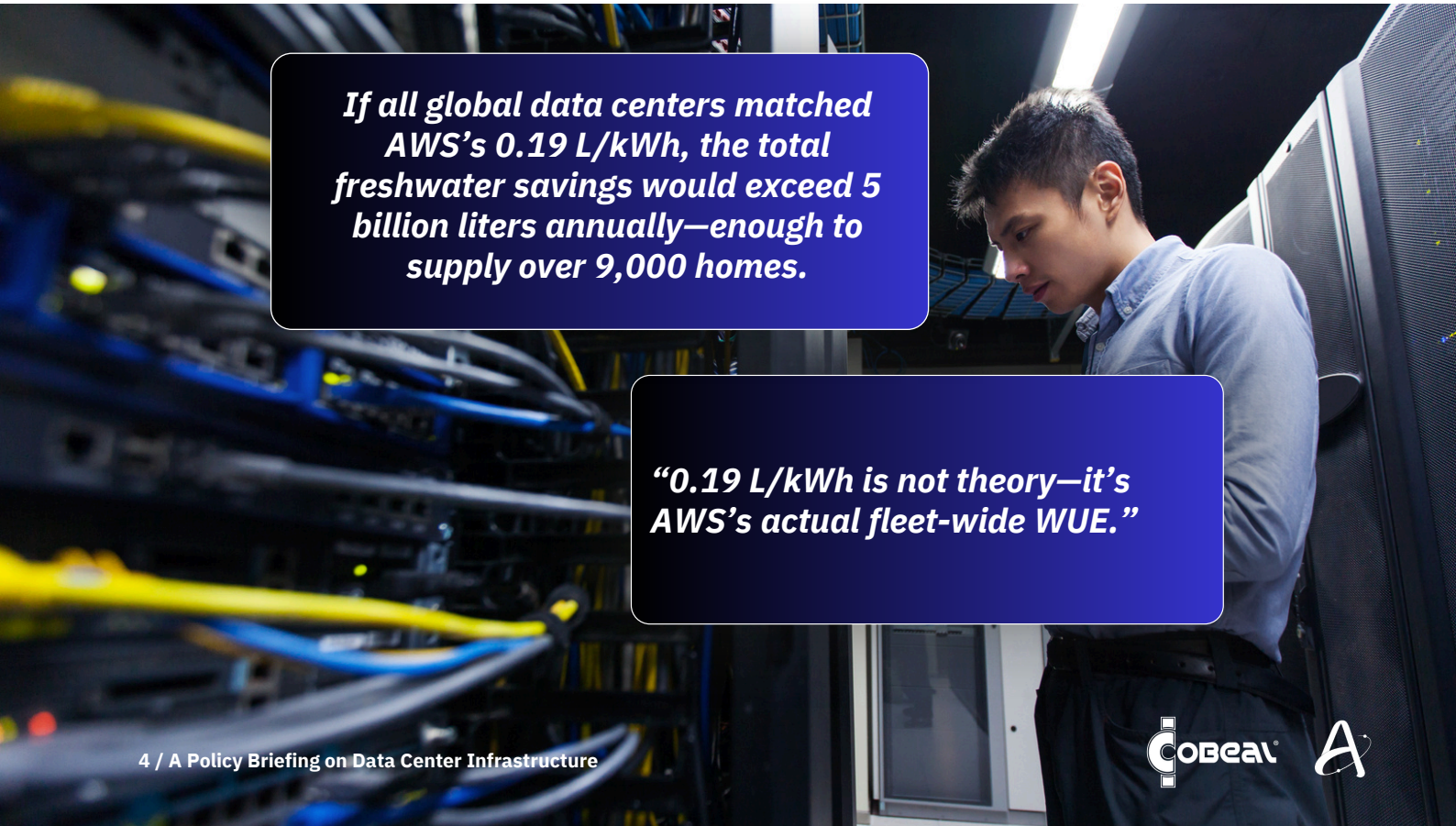
Even when evaporation is used, operators can reduce water use by timing AI workloads. Training runs are increasingly scheduled at night or in cool climates like Sweden, Oregon, or Iowa. This strategy, endorsed even by Ren et al., shows that “per prompt” water can vary 3x or more based on location and hour; meaning, averages are often misleading.

Modern AI Data Centers Are Actively Redesigning for Water Efficiency

Microsoft, Google, and Meta now deploy direct-to-chip liquid cooling and use reclaimed or non-potable water where possible.

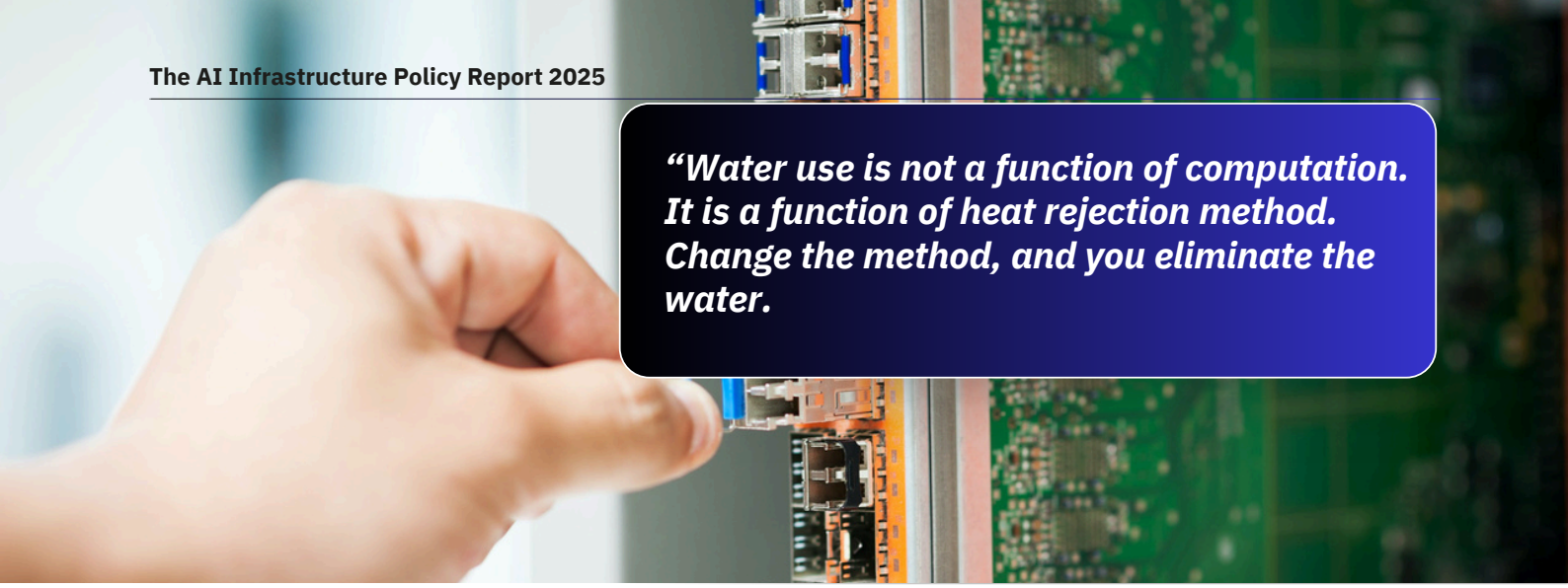
These systems use water only as a last resort, not as a baseline cooling method—cutting water use by 70–100% compared to legacy towers.

For full technical documentation, cooling cycle diagrams, and empirical WUE benchmarks, see:
No, AI Doesn’t Drink a Bottle of Water per Prompt



If all global data centers matched AWS’s 0.19 L/kWh, the total freshwater savings would exceed 5 billion liters annually—enough to supply over 9,000 homes.

“0.19 L/kWh is not theory—it’s AWS’s actual fleet-wide WUE.”



> ***"Water use is not a function of computation. It is a function of heat rejection method. Change the method, and you eliminate the water.***

Reframing Water Use: From Myth to Measurable Policy

The sensationalism around AI's "water footprint" has outpaced engineering reality. As shown in the Ren et al. (2023) study, worst-case assumptions stack high-legacy cooling towers, fossil power, and static scheduling. But those conditions do not reflect how modern AI systems are cooled, nor how water efficiency is achieved.

> ***"When regulators assume static infrastructure, they regulate against yesterday's data centers—not today's capabilities."***

Modern hyperscale data centers are designed for flexibility. They adapt cooling strategies in real-time, using economizers in cold weather, dry coolers where water is scarce, and reclaimed water instead of potable sources when needed. These design decisions aren't theoretical; they are the standard across Microsoft, Google, Meta, AWS, and NVIDIA's latest AI builds.

The issue is not AI itself. It's the assumptions regulators make when interpreting early-stage studies. Many headlines cite outdated averages and miss critical variables: climate, scheduling, heat reuse, or cooling architecture. Good policy must distinguish between facilities running state-of-the-art zero-water cooling and those using high-evaporation legacy systems.

Design flexibility is the defining feature of modern AI infrastructure. Cooling is no longer a one-size-fits-all constraint—it is a tunable parameter shaped by climate, load type, and sustainability goals. Leading hyperscalers dynamically balance efficiency and resilience by using reclaimed water, dry systems, or seasonally adaptive controls. These aren't theoretical solutions—they're deployed now. To regulate AI's impact accurately, policymakers must stop treating water consumption as an inevitable cost and start treating it as a solvable engineering input.

To ensure environmental oversight keeps pace with technical reality, policymakers must require visibility into how AI infrastructure is actually cooled. Today's best systems already minimize or eliminate water use, but these improvements are invisible in legacy permitting frameworks.

Without standardized WUE reporting, reclaimed water tracking, and cooling architecture disclosures, regulators are left regulating ghosts in the machine, based on estimates from outdated facilities with evaporative towers and fossil-fueled power. These five policy recommendations reflect current capabilities already deployed by major operators. The role of regulation is to recognize, reward, and require what is already technically possible.

- **"Require Transparent WUE Reporting"**
  Mandate standardized Water Usage Effectiveness (WUE) metrics for all AI-capable data centers to distinguish between legacy and efficient systems.
- **"Differentiate Potable vs. Non-Potable Use"**
  Count only potable water draw in regulatory assessments. Encourage use of reclaimed or non-potable sources through incentives.
- **"Align Siting Incentives with Climate-Aware Design**
  Offer policy support for facilities that locate in cooler climates or use economization to reduce annual water use.
- **"Incentivize Zero-Water Cooling Tech"**
  Provide R&D credits or expedited permitting for operators investing in air-cooled or closed-loop systems.
- **"Disclose Cooling Architecture in Environmental Reviews**
  Require data center projects to include detailed cooling systems specs (e.g., evaporative vs. dry) during EIR/EA processes to accurately project local impact.

Part 2: Introducing the Power Reality

Electricity, Not Water, Is the Backbone of AI Infrastructure

The public narrative has focused on water, but it is electricity—not liquid cooling—that defines AI's environmental profile. Training a large language model today consumes hundreds of megawatt-hours. But that figure is not a fixed cost; it varies widely depending on scheduling, siting, and architecture.

This section reframes AI's power demand using real-world engineering data. It distinguishes training from inference, and legacy data centers from modern liquid-cooled clusters. The findings show that AI workloads can be efficiently scheduled around grid availability, and that some data centers already operate at PUEs approaching 1.1 or lower—comparable to top supercomputing facilities.

Rather than triggering alarm, the growth in AI compute should motivate policy clarity. Without accurate metrics or meaningful transparency, well-designed facilities risk being mischaracterized by outdated national averages.

Understanding Power Use in AI Workloads

Electricity consumption in AI systems depends on the type of workload, model architecture, and data center design. Training large-scale models like GPT-3 can consume up to 1,300 MWh, but inference workloads are several orders of magnitude lower.

Recent benchmarking shows that:

- Training GPT-3 consumed ~1,287 MWh over two weeks.
- Inference using GPT-3 or GPT-4 typically draws 0.01-0.02 kWh per prompt when run on GPU clusters.
- This means training a model is equivalent to 60-80 million individual queries.

Best-in-class operators are addressing this delta with:

- Smart workload scheduling, shifting training jobs to nights or off-peak hours.
- AI siting strategy, placing clusters near renewable generation.
- Low-PUE facilities, now reporting fleet-wide PUEs as low as 1.1-1.2, with some approaching 1.05 in optimized sites.

Key Takeaways:

• AI power use is driven by training runs, not everyday prompts.

• Energy use varies up to 5x based on data center design.

• Modern deployments use direct-to-chip liquid cooling and align workloads with grid availability.

• PUEs below 1.2 are already being achieved by hyperscalers.

• Policy must separate outdated infrastructure from high-efficiency builds.

From Water Warnings to Power Reality

Unlike water, where outdated cooling models have dominated headlines, concerns over power draw are rooted in real, measurable demand growth. AI workloads require significant electricity for both training and inference, and the accelerating global buildout of GPU-intensive data centers has prompted fears that AI will outstrip national grids.

However, these concerns must be placed in context. Power consumption in AI systems is not an uncontrollable externality. It is a function of system design, workload management, and facility efficiency.

Just as cooled-loop cooling helped debunk the myth of the "500 mL per prompt," data shows that best-in-class AI data centers are already achieving record low PUEs (Power Usage Effectiveness), with intelligent scheduling and hardware advances slashing total energy consumption per compute unit.

Understanding PUE and Energy Performance in AI Workloads

Power Usage Effectiveness (PUE) is the industry-standard metric for evaluating the energy efficiency of a data center. It is defined as the ratio of total facility energy to energy used by IT equipment. The closer this value is to 1.0, the more efficiently a data center is operating.

Legacy data centers often reported PUEs above 1.6 or even 2.0, reflecting substantial overhead from cooling and facility operations. By contrast, state-of-the-art AI facilities from Google, Microsoft, and AWS report PUEs as low as 1.1 or even below. NVIDIA's DGX SuperPOD deployments—used for large model training—are designed to operate within this ultra-efficient range, using direct-to-chip liquid cooling and optimized airflow paths to reduce waste heat and fan power draw. Real-world GPT-scale models now consume hundreds of megawatt-hours per training cycle. But improvements in scheduling, rack density, and GPU utilization have driven down the energy cost per token by up to 70% compared to 2020 benchmarks. This shows that training large models is not fixed-cost intensive; it depends heavily on architectural design.

Electricity Use is the New Regulatory Frontier

With policy makers now considering moratoriums and quotas on new data centers, it is critical to distinguish between physical limits and design variables. Many of the most alarming power statistics—such as AI models "using as much energy as a small country"—are based on outdated infrastructure or worst-case stacking. In contrast, leading AI operators are deploying direct-to-chip liquid cooling, aligning compute loads with renewable generation, and locating facilities in regions with ample grid capacity.

AI infrastructure is not inherently unsustainable. But without standards for disclosure, performance benchmarking, and climate-aware siting, even well-engineered systems risk being lumped into inflated industry narratives. Page 8 will present real-world benchmarks for training vs. inference power loads and recommend pathways for transparent energy governance in AI development zones.

AI's energy footprint is not a law of nature. It is the result of compute intensity, model size, and system design. Better infrastructure = better outcomes.

Unaddressed Power Realities and Governance Risks

Regulatory concern around AI's electricity use has grown rapidly, but the loudest claims (comparing AI workloads to entire countries) often ignore context. While training GPT-3 consumed ~1,287 MWh over two weeks, Microsoft and Google's latest deployments are cutting that energy per training token by up to 70%, using direct-to-chip cooling and intelligent scheduling.

The real issue is not consumption itself, but the absence of disclosure standards. OpenAI, for example, has not published full training data for their GPT-4. Meanwhile, local utilities are beginning to push back, with Microsoft's Iowa deployment drawing 11.5 million gallons in one hot month, prompting the city to demand peak water cuts for future expansion. Without enforced transparency on training loads, cooling architecture, and energy sourcing, AI infrastructure remains at risk of mischaracterization—or overregulation.

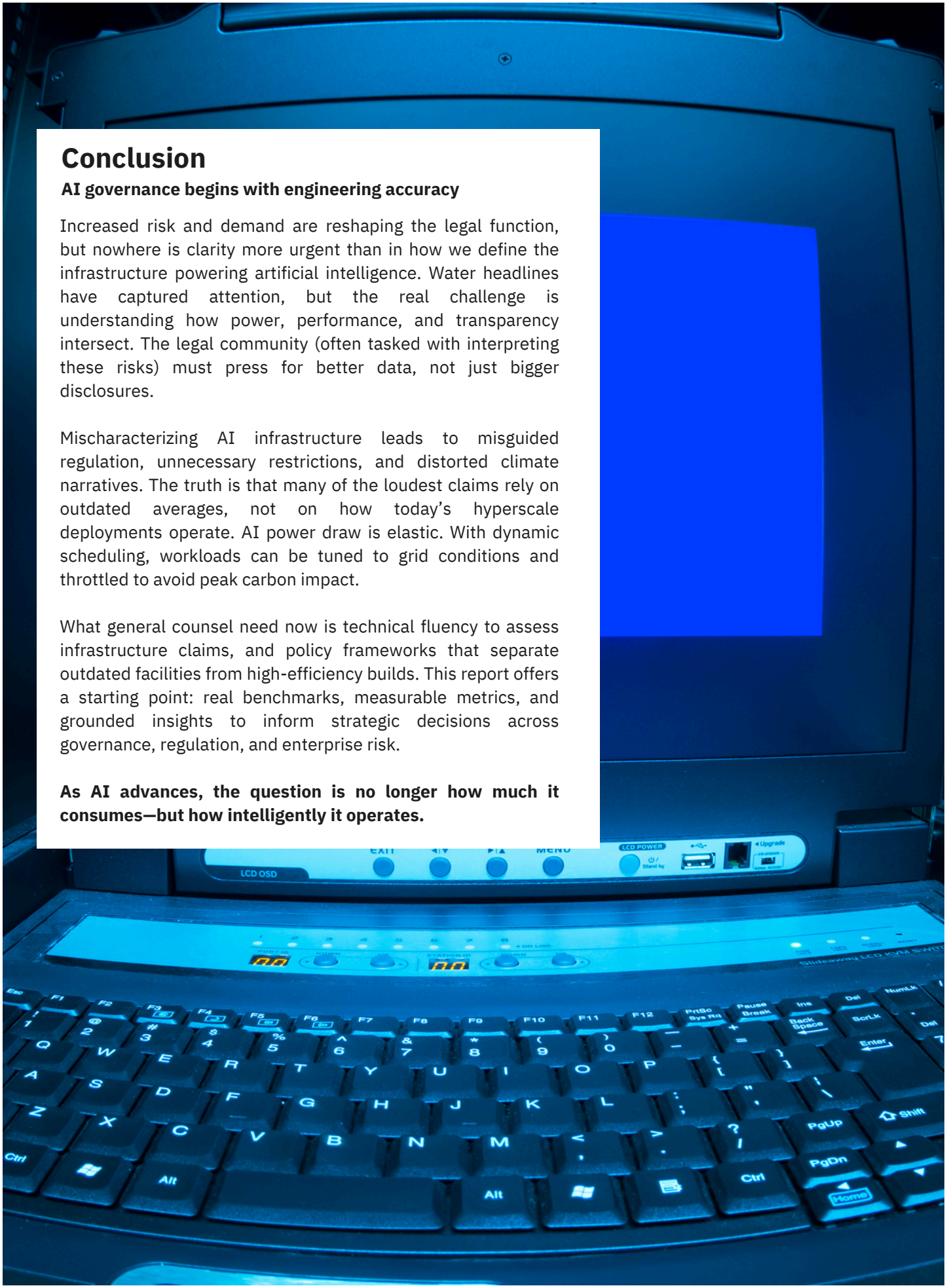
Why Transparency Matters: Modeling Policy After Performance

Policy debates around AI infrastructure often rely on outdated or generalized energy profiles that don't reflect what modern data centers achieve. Facilities running cutting-edge AI systems, such as NVIDIAs SuperPODs or hyperscaler AI clusters operate under a very different set of assumptions than the industry average. Without mandatory reporting of PUE, WUE, and energy sourcing, regulators may default to national baseliens that penalize efficient sites and discourage innovation.

Transparent benchmarking enables smarter zoning decisions, grid coordination, and investment into zero-carbon compute. It also helps separate normal energy scaling from true excess. Just as emissions targets have differentiated between Scope 1 and Scope 2 carbon, AI infrastructure needs a similar framework to parse training loads, inferencing loads, and cooling method impacts. Data-backed differentiation is the key to future-proof regulation.

Key Energy Benchmarks from Hyperscale AI Deployments

- U.S. Data Center Power Use surged from 58 TWh in 2014 to 176 TWh in 2023, with AI expected to push that to 325–580 TWh by 2028, representing as much as 12% of all U.S. electricity.
- Training GPT-3 used as much electricity as 130 U.S. homes consume in a year.
- Inference workloads now draw more power than training, with ChatGPT's live usage exceeding 400 GWh annually.
- A typical ChatGPT query uses 10× the power of a Google search (2.9 Wh vs. 0.3 Wh).
- AI model inference at global scale is now comparable to charging over 3 million electric vehicles annually.
- NVIDIA claims 300× improvement in water efficiency using closed-loop cooling for H100 systems.
- AWS fleet-wide PUE: ~1.1; WUE: 0.19 L/kWh. In contrast, legacy data centers exceed 1.8 L/kWh and PUE of 1.6–2.0.
- AI operators are strategically siting facilities near renewable generation zones and deploying climate-aware scheduling to match training with low-carbon energy availability.



Conclusion

AI governance begins with engineering accuracy

Increased risk and demand are reshaping the legal function, but nowhere is clarity more urgent than in how we define the infrastructure powering artificial intelligence. Water headlines have captured attention, but the real challenge is understanding how power, performance, and transparency intersect. The legal community (often tasked with interpreting these risks) must press for better data, not just bigger disclosures.

Mischaracterizing AI infrastructure leads to misguided regulation, unnecessary restrictions, and distorted climate narratives. The truth is that many of the loudest claims rely on outdated averages, not on how today's hyperscale deployments operate. AI power draw is elastic. With dynamic scheduling, workloads can be tuned to grid conditions and throttled to avoid peak carbon impact.

What general counsel need now is technical fluency to assess infrastructure claims, and policy frameworks that separate outdated facilities from high-efficiency builds. This report offers a starting point: real benchmarks, measurable metrics, and grounded insights to inform strategic decisions across governance, regulation, and enterprise risk.

As AI advances, the question is no longer how much it consumes—but how intelligently it operates.

About Sophy M. Laughing, Ph.D., MBA

Executive Leader Specializing in Mission-Critical Infrastructure

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.

For more information: Alden Technologies, Inc.

About Bo Erik Hollsten Ruvalcaba

Veteran Member of ANSI/ASHRAE, ISO, IEST

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.

For more information: The Cobeal Group