AI Literacy Lab

Bykov-Brett Enterprises

A BYKOV-BRETT BRIEFING

The Real Cost of the AI Cloud

AI, water, power, and the hum next door. A balanced briefing.

What AI actually costs the planet, measured honestly.

Turn Intent Into Impact.

A Bykov-Brett Enterprises briefing.

THE HONESTY CONTRACT

How to read this

- Fact vs estimate. When a number was actually measured, we say so. When it was worked out by a model or is a guess about the future, we flag it. A projection is a forecast, not a receipt.
- Green water vs blue water. Green water is rain a crop would have soaked up anyway. Blue water is water we pump from rivers, lakes, and aquifers (underground stores of water). They are not the same thing, and mixing them makes comparisons look scarier than they are.
- Ranges beat fake precision. Honest numbers usually come as a spread, like '2 to 3', not one exact figure. A number that looks super precise is often the one you should trust least.

MEASURED
A study counted it.

ESTIMATED
A model worked it out.

PROJECTED
A forecast of what might happen.

If a figure has no source and no range, treat it as a rumour.

THE SHORT VERSION

If you only read one page

AI's effect on the planet is real and growing. But one use of it is tiny, it is shrinking fast, and what matters most is where and how data centres are built, not whether you send a prompt.

- One question uses about five drops of water and a tiny sip of electricity. That is roughly ten times less than the 2023 scare stories said, and it is still falling.
- All the world's data centres together use about 1.5% of the world's electricity. AI is only a slice of that. Real and growing, but not boiling the planet.
- Where it plugs in matters more than the model. The exact same job is clean or dirty depending on the local power.
- The real water story is farming, not chatbots. Crops grown to feed cattle take about a third of the Colorado River on their own.
- The problems worth fixing are local: noise, strain on the power grid, and building in dry places. The lever is how and where we build, not your prompt.

▶ **Small per use. Big when you add it all up. Decided by where and how it runs.**

THE MEASUREMENT PROBLEM

Why is this so hard to see?

01 The stats are real, but old

The AI-and-environment numbers people share most come from real research papers. But they are often quoted wrongly, stripped of their warnings, or years out of date. A number measured back in the GPT-3 days gets repeated as if it were true today.

02 The cost hides upstream

Some of it is in the building's cooling. More of it is further back, in the power station that feeds the building. The electricity grid it plugs into (grid mix means what fuels make that area's power, like coal, gas, wind, or nuclear) can matter more than the AI model itself.

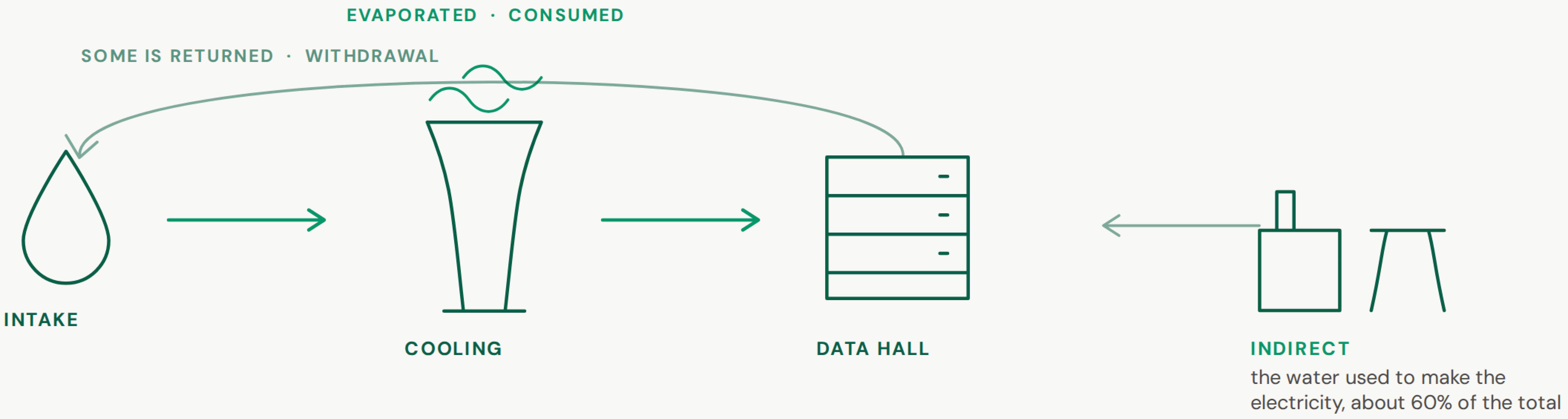
03 It always depends

One simple question, what does a single use of AI cost, has honest answers that vary by about ten times, depending on the hardware, the region, and the year.

▶ **Most scary AI stats are true numbers pointing at the wrong year or the wrong place.**

Water 101: two taps, not one

A data centre uses water in two ways, and people usually only count the first.



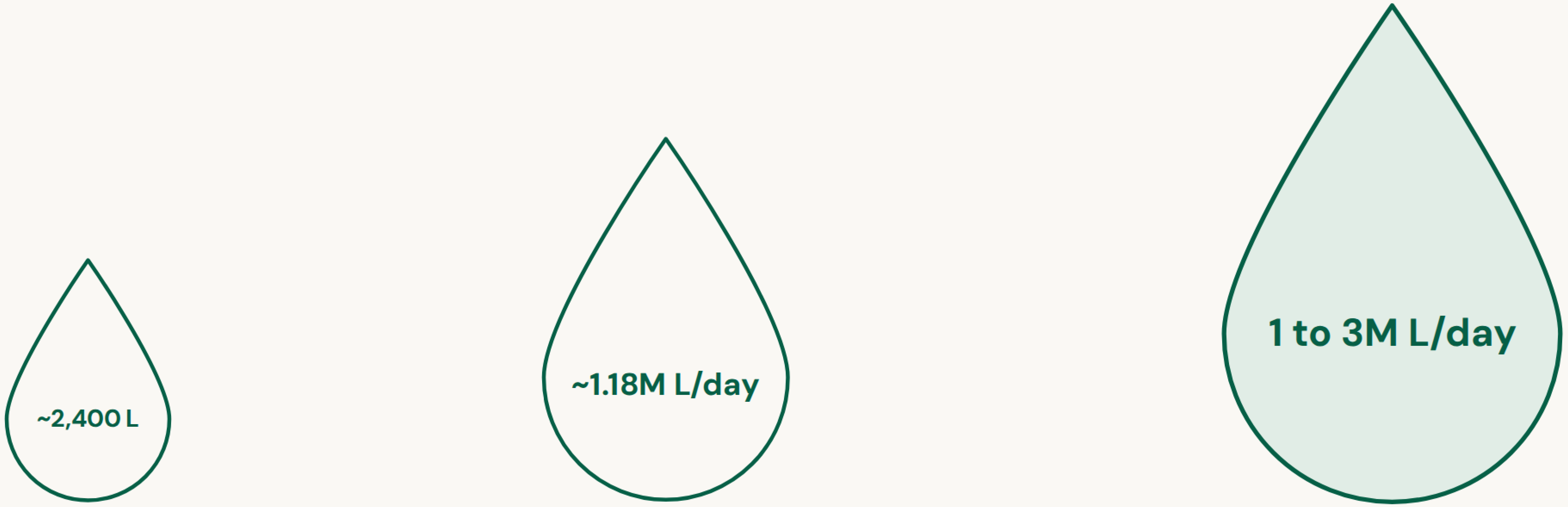
→ Withdrawn water is borrowed. It can go back to the river or lake it came from. Consumed water evaporates into the air and is gone for good. Headlines often quote the borrowed amount but make it sound like it was all used up.

▶ **Count both taps, and keep water borrowed separate from water burned.**

FOR SCALE

What a litre looks like elsewhere

Set side by side for scale. The drop sizes are rough, so read the figures. And the kind of water matters as much as the amount.



Beef burger (150g)
The whole chain; about 98% is water for the animal's feed, mostly green rain.
Mekonnen & Hoekstra 2012

Golf course (per day)
Mixed; much of it recycled.
USGA ~2015; GCSAA

Large data centre (per day)
Mostly blue; pumped supply
dgtlinfra 2024; EESI 2023

The burger's roughly 2,400 litres covers the entire process of making it, and about 98% of that is water to grow the animal's feed. Most is green water, meaning rain falling on grass. But the part grown with irrigation is blue water (pumped from rivers or underground). That blue share rises from about 6% of feed water worldwide to about 9% for grain-fed feedlot beef, and that blue part is what actually puts pressure on supply.

▶ **Look at the scale first, then ask which tap it came from. A bigger number is not automatically the worse one.**

How thirsty is AI, really?

~0.26 mL

PER PROMPT · GOOGLE 2025

A typical Gemini text prompt uses about five drops of cooling water at the data centre (Google, Aug 2025). 'Per prompt' means for one question you ask. The old '500 ml' figure was actually per 10 to 50 GPT-3 questions, and it is both old and quoted wrongly.

~700,000 L

TO TRAIN GPT-3 · 2020

A one-time figure from history, for a 2020 model. Training means the huge one-off job of building the model in the first place, and that is where water piles up. But no company today publishes the training water for its own models.

~32%

OF THE COLORADO RIVER

...goes to cattle-feed crops like alfalfa (Richter et al., 2024). Watered farming uses vastly more than data centres, though a data centre can still strain one particular local aquifer (underground water store).

THE HONEST READ

A single modern prompt is a few drops, and it is getting smaller. The water worth arguing about is where it is taken from and what it grows, not your chatbot. Google's figure counts only the cooling water at the site. Counting the whole journey, including the water to make the electricity, a peer-reviewed study of Mistral's model put it nearer 45 ml, about three tablespoons. So 0.26 ml is the floor and 45 ml is the honest ceiling.

▶ **A prompt is a few drops. The water that strains supply is farms and location, not your chatbot.**

POWER

How much power does AI really use?

415 TWh → 945 TWh

How much electricity the world's data centres used in 2024, and a forecast for 2030 (about 1.5% of the world's electricity today, more than doubling; about 1,200 TWh by 2035). IEA, 2025.

Where the growth sits

- AI chips (the special processors that run AI) account for about half of the extra data-centre electricity by 2030.
- Goldman Sachs estimates AI drives a roughly 165% rise in data-centre power demand by 2030.
- US data centres climb from about 3% of US power in 2022 toward about 8% by 2030.

Your prompt is tiny; the whole fleet is not

- A modern text prompt uses about 0.24 Wh (Google) to 0.3 Wh (Epoch), 2025. That is roughly 10 times less than the 2023 estimates that fed the scare stories.
- Thirty minutes of streaming video uses more energy than a whole day of your text prompts.
- Answering prompts is done in big batches (many prompts handled together to save power), and the data centres are built years ahead of time. So cutting your own prompts barely changes the total. The real lever is where data centres get built, not you holding back.

▶ **Per prompt, it fell about 10 times. Added all together, it is still doubling. Both are true.**

CARBON

Same model, different planet

552 t

GPT-3 training · 2020

On a US grid. Patterson et al. 2021

25 t

BLOOM training · 2022

On France's nuclear grid. Luccioni et al.

The biggest lever is the grid, meaning what fuel makes the local electricity. Coal puts out about 961 grams of CO_2 per kWh; wind and hydro put out under 10. The US grid averages about 429, the French nuclear grid about 57. So the exact same training run is clean or dirty depending on where it plugs in.

And today, per prompt: a typical Gemini query is about 0.03 g CO_2 (Google, Aug 2025), a figure Google says fell 44 times in a single year. There is still no trusted industry-wide average for grams of CO_2 per query, so the numbers that matter are the grid mix (the local fuel) and the total added up, not your single query.

Read the two panels fairly. GPT-3 is a 2020 model. BLOOM's 25t counts only the power used while running, and is nearer 50t once you add embodied carbon (the carbon from making the hardware itself). The two studies also used different methods. The grid mix is still the main reason for the gap.

▶ **Where you plug in can change the carbon by roughly ten times.**

A SENSE OF SCALE

So how big is AI, really?

Put next to the things around it, AI is small. That does not make it nothing. It means we should get the size right before we worry.

WHERE THE WATER GOES

69%

of the world's fresh water goes to farming

All industry together, which includes every data centre's cooling, is about 19%. Homes and cities are 12%.

In the US, cooling power stations and watering crops are so huge that data centres do not even get their own line in the national count.

FAO / AQUASTAT; USGS

WHERE THE POWER GOES

~0.2%

of the world's electricity is AI today

All data centres together are about 1.5%. AI is only a slice of that slice.

The whole growth in data centres to 2030 is less than half of one recent year of ordinary growth in world electricity.

IEA, 2025

▶ **Get the size right first. AI is a small user standing next to some very big ones.**

Cheapest, fastest, weakest rules

4,900+ MW Northern Virginia is the world's largest data-centre hub, more than twice the size of Beijing's.

WE DROPPED THIS

The 70% myth

You may have read that it carries 70% of the world's internet traffic. No one can trace where that number came from, so it is not in this briefing. Believable estimates put the real share nearer 13 to 23%.

THE PUBLIC COST

A billion in tax breaks

Virginia's data-centre tax break gave up about $1 billion in FY2023, and roughly 82% of it went to the five largest firms.

THE OTHER HALF

Even so, the industry puts more money into the state than it takes out, and data centres pay the full cost of their own power connections. The fair complaint is that they are crowded into one place and do not disclose enough, not that the sector is draining the state.

▶ **The building clusters where power is cheap and rules are light. Both of those things can change.**

WATER-STRESSED BUILDS

Building in the desert

2 billion gal

NORTHERN VIRGINIA, 2023

Data-centre water use, up about 63% since 2019. The growth is real.

4 million gal/day

ARIZONA (MESA)

A Google deal allows this from an aquifer (underground water store) that is already promised out to more users than it can supply.

THE OTHER HALF

For context, all industry put together took only about 6% of Mesa's drinking water in 2024, and operators are moving toward air cooling. The problem is being careful about where you build in dry areas, not water use everywhere.

Policy experts at ITIF call the data-centre water problem solvable, using recycled water, closed loops (water that circulates round and round instead of being thrown away), and smarter siting. The tools exist. The real question is whether anyone makes companies use them.

▸ **A thirsty build over a dry aquifer is a choice about where to build, and where to build is fixable.**

GRID STRAIN

Who pays for the wires

+$14 to $37 /mo

VIRGINIA HOUSEHOLDS

Virginia's own analysts modelled this by 2040, in today's money, for Dominion home customers, and only under a scenario where growth is left unchecked. Every one of those conditions matters.

226 GW

TEXAS ERCOT QUEUE

The queue of big new power users hit this by November 2025, roughly four times higher than a year earlier, about 77% of it data centres.

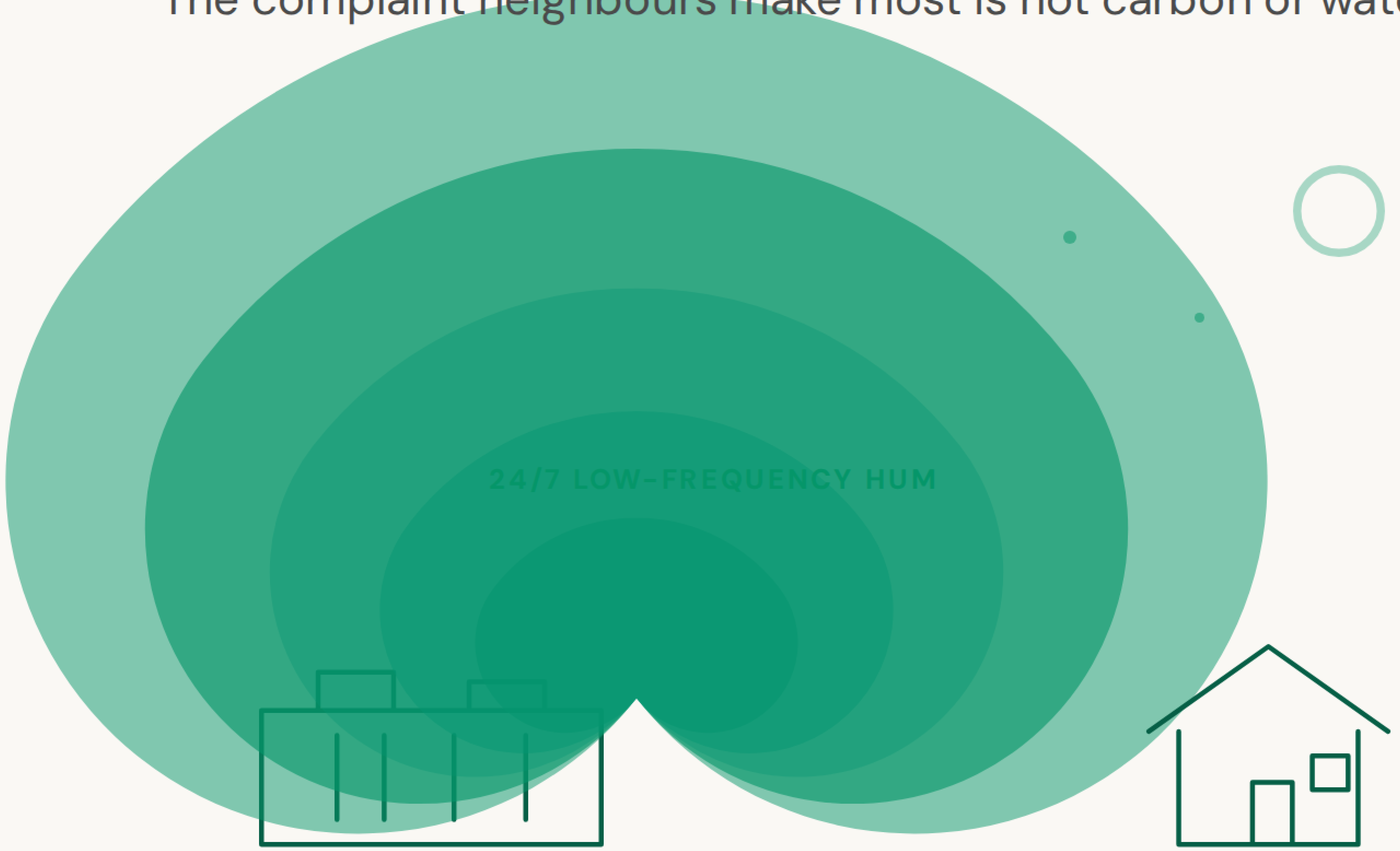
THE OTHER HALF

That Texas queue is power that has been asked for, not power that has been built. A lot of it is ghost requests that will never switch on. New rules are being brought in, and the state's line is that data centres pay their own connection costs, not households.

▶ **The bill risk is real under worst-case growth, and it is exactly what the new rules are written to prevent.**

The hum next door

The complaint neighbours make most is not carbon or water. It is noise.



Where it comes from

Cooling fans run non-stop, chillers and transformers add to it, and the backup diesel generators are the loudest of all. That transformer hum is magnetostriction, which just means the metal core of the transformer physically flexes a tiny bit with the electric current, over and over.

Why it carries

It is low-frequency sound, meaning a deep, low tone (also called infrasound when it is too low to hear clearly). Low sound travels a long way, passes through walls, and feels worse at night when everything else is quiet.

60 to 80 dB steady at the edge of the property

80 to 100 dB from generators, heard up to a mile away

Infrasound you feel but that standard sound meters miss

Best documented in Chandler, Arizona and Prince William County, Virginia, where some residents moved away.

▶ **It is the low, constant hum, not the loud peaks, that wears people down.**

WHAT CAN BE DONE

Quieting it down

The noise is one of the easier problems to fix in this briefing. There is a clear toolkit.

The toolkit

- Sound walls and low-frequency shields, sized for the deep hum, not just for high-pitched noise.
- Sound-proof boxes around the generators, and only testing them during the daytime.
- Bigger gaps between the site and homes. Distance is the most reliable fix, and some sites sit within 50 feet of houses.
- Liquid cooling, which is quieter by nature because it removes the walls of fans, though the pumps add a little noise back.

Rules that already work

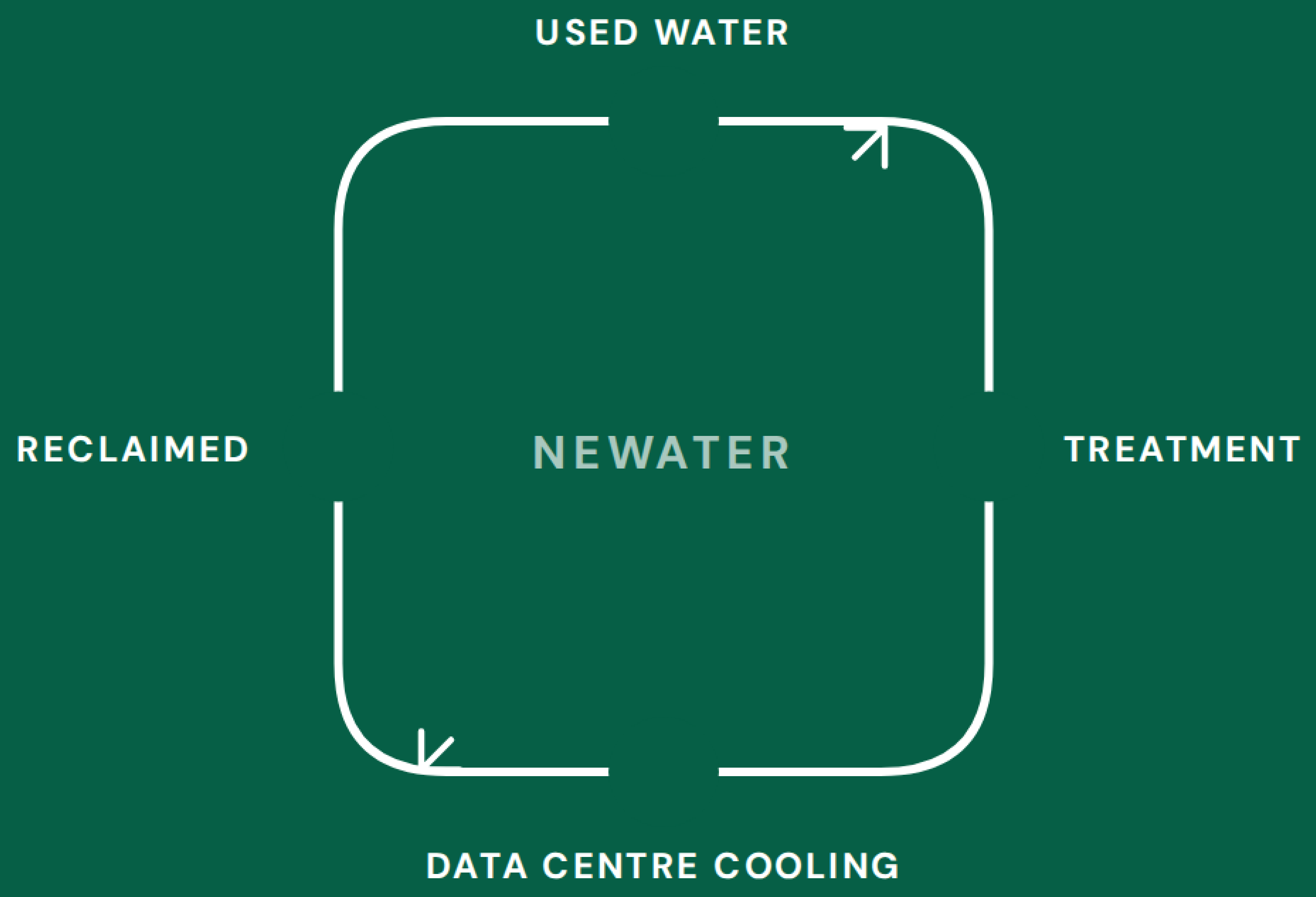
- Chandler's 2022 rule says a site must be no louder than the area was before it arrived, with checks for five years.
- Prince William caps noise at 52 dB by day and 47 dB at night.
- The UK and Ireland use BS 4142: background level plus 5 dB by day, plus 3 by night, with an extra 5 dB penalty for a tonal hum (a clear, steady note that annoys people more).

▶ **Distance plus enforceable night limits fixes most of it.**

DOING IT WELL

Singapore: doing it on hard mode

A small island that is short on water grows data centres without wrecking its power grid or its water supply.



Purple pipes

Recycled water runs through its own separate pipe network to industry and cooling, up to about 40% of the country's water demand. Cooling never has to touch drinking water.

A pause, then a plan

A moratorium (a temporary ban on new builds) from 2019 to 2022, then a system of conditional quotas and a Green Data Centre Roadmap in 2024 that allowed a further 300 MW, but only if efficiency conditions were met.

A standard with teeth

The Tropical Data Centre Standard, 2023, lets halls run at 26C and above. Every 1C you let the temperature rise saves roughly 2 to 5% of the cooling energy.

Its power share was about 7% in 2020, and is widely reported as possibly reaching 12% by 2030 if nothing is done. Singapore is small, rich, and run from the centre: proof that it can work under tight limits, not a plan that drops neatly onto every country.

▶ **Tight limits plus purple pipes plus a warmer set-point is a recipe that works, not a miracle.**

BRIGHT SPOTS

What good already looks like

Waste-heat reuse

Meta's Odense site heats about 6,900 Danish homes; the UK's Deep Green warmed an Exmouth swimming pool; Stockholm aims to get 10% of the city's heat from data-centre waste. The catch: server heat is only lukewarm, so it needs heat pumps and cold-climate heating networks to be useful.

Liquid cooling

Liquids carry roughly 3,000 times as much heat as air, and cooling the chip directly removes about 70 to 75% of a rack's heat. Cooling is about 40% of a data centre's power, so this is a big lever. Closed loops (water going round and round instead of being thrown away) cut water use, though how much depends on the design.

Clean siting

Google aims to run on carbon-free energy 24 hours a day by 2030, and is honest that it has not cracked it yet; Iceland runs on hydro and geothermal (heat from the ground). The catch is additionality: buying clean power can just push someone else onto fossil fuels, so no new clean power is really added.

Efficiency works

From 2010 to 2018, data-centre computing rose about 550% while the energy used grew only about 6%, because efficiency soaked up the growth. That gap between the two was real.

The best sites hit a PUE of about 1.1 (Google 2025). PUE is a score for how much extra power a data centre spends on cooling and overheads; 1.0 is perfect and higher is worse. But the industry average is stuck around 1.56 (Uptime Institute 2024), and a low PUE still does not mean a low total footprint.

▶ **The engineering answers already exist. The gap is rolling them out, not inventing them.**

WHAT AI CAN SAVE

Can AI cut more than it costs?

Almost every number about AI is the energy going in. But AI can also help other industries waste less. The IEA put real figures on both sides.

300 to 500 Mt

CO_2 that AI's data centres ADD each year by 2035

~1,400 Mt

CO_2 that AI could help the world AVOID by 2035

Mt means millions of tonnes of CO_2. Used widely to make buildings, transport, factories and the power grid run smarter, the savings the IEA models are several times bigger than the emissions from running the data centres themselves.

READ THIS CAREFULLY

The savings are a possibility, not a promise. They only happen if those tools are actually rolled out, and the IEA is clear it is still far smaller than what the world needs to fix the climate. A hopeful direction, not a free pass.

▶ **AI might cut more than it costs, if we actually use it that way.**

THE EFFICIENCY PARADOX

The catch: efficiency can backfire

There is a fair reply to every hopeful page in this briefing. It is called Jevons paradox: when something gets cheaper and more efficient, we often use so much more of it that the total goes up, not down. (This is also called the rebound effect.)

We have already seen it. Efficiency soaked up a decade of computing growth, yet the IEA still expects demand to roughly double by 2030. In some models the rebound is over 100%, which wipes out the saving completely.

So efficiency is needed, but it is not enough on its own. Without a cap, a limit, or clean power underneath it, doing more with less can still mean doing far more overall.

EFFICIENCY
per unit, falling

DEMAND
overall, rising fast

NET IMPACT

▶ **Efficiency buys us room. Only choices about where we build and what powers it decide whether we spend that room well.**

GOOD REGULATION

The rules that actually work

The pattern across the best examples is the same every time. It is not a ban. It is disclosure, protection, and discipline.

01 Disclosure first

Minnesota and Illinois make operators report their energy and water use, so the public can see the real numbers.

02 Ratepayer protection

Texas and Oklahoma shield households from the cost of connecting big new power users to the grid.

03 Siting discipline

Distance, night-noise limits, and rules against building over dry basins are the reliable levers, as Chandler and Prince William show.

04 Right-sized incentives

Illinois, Arizona, and Ohio paused their tax breaks, and others are scrapping them in 2026, once the giveaways cost more than they were worth.

IN PROPORTION These are course corrections on a real industry, not a crackdown. Analysts at ITIF judge the water problem solvable with exactly this toolkit. The key lever is being careful about where you build, applied early.

▶ **Make them disclose it, make them pay their own way, and choose where they build.**

WHAT A BUSINESS CAN DO

If you build with AI, measure it

You cannot manage what you do not measure. The good news is the measuring tool is free.

THE FREE TOOL

CodeCarbon

A free, open-source Python library. You wrap it around your code and it estimates the energy your CPU, GPU, and RAM use, multiplies that by how dirty your region's electricity is, and reports the result in kilograms of CO_2. Made by Mila, Yoshua Bengio's institute, with BCG GAMMA, Comet, and Haverford.

```
$ pip install codecarbon
$ tracker.start()  ...  tracker.stop()
✓ emissions: 0.42 kg CO2e
```

Four moves that help

- Choose providers who publish their PUE (their cooling-overhead score) and water use, and who run on clean grids.
- Run heavy jobs at low-carbon times or in low-carbon regions. The grid does most of the work.
- Right-size the model. Do not send a giant model to do a small job.
- Prefer providers who use recycled water or closed-loop cooling (water that circulates instead of being thrown away).

▸ **Measure first, then move the job to a cleaner grid and a smaller model.**

THE HONEST SCORECARD

What we know, and what we don't

WE KNOW

- ✓ Data centres use about 1.5% of the world's electricity today, and that is expected to more than double by 2030.
- ✓ The grid mix (the local fuel) can swing the carbon by roughly ten times.
- ✓ Running the models day to day, not training them, uses most of the lifetime energy. (Inference means answering your prompts; training means the one-off job of building the model.)
- ✓ Noise is real, local, and fixable.

WE DON'T KNOW CLEANLY

- ? One trusted figure for CO_2 per query. Estimates run from under a gram to several grams.
- ? The exact water per modern prompt. It is argued over, somewhere around a few millilitres.
- ? How much of the forecast demand actually shows up.

ONE UNDER-COUNTED COST

Generative AI could add up to about 2.5 million tonnes of electronic waste a year by 2030, as old chips and servers are thrown out. The same study finds that reusing and repairing that gear could cut it by 16 to 86%.

▶ **Real and growing, smaller per use than the panic implies, and decided by where and how it runs.**

CHECK THE RECEIPTS

Words to know, and where the numbers come from

WORDS TO KNOW

Data centre: a building full of computers that run things like AI, streaming and email.

Grid / grid mix: the electricity supply, and what fuels make it. A cleaner mix means lower carbon.

Blue vs green water: blue is pumped from rivers, lakes or underground; green is rain a plant would have used anyway.

Training vs inference: training is the one-off job of building a model; inference is answering your questions day to day.

PUE: a score for the extra power a data centre spends on cooling. 1.0 is perfect, higher is worse.

TWh, kWh and Wh: units of electricity. A watt-hour is tiny; a kilowatt-hour (kWh) is a thousand of them; a terawatt-hour is a billion times bigger.

WHERE THE NUMBERS COME FROM

Energy & carbon

IEA Energy and AI (2025); Google Gemini disclosure (2025); Epoch AI (2025); Patterson et al. (2021); Luccioni et al., BLOOM (2022); Masanet et al., Science (2020).

Water

Google (2025); Mistral peer-reviewed life-cycle study (2025); Li et al., Making AI Less Thirsty (2023); Mekonnen & Hoekstra (2012); Richter et al., Colorado River (2024); FAO/AQUASTAT; USGS.

Siting, grid & noise

JLARC Virginia (2024); Cardinal News (2025); ERCOT / Texas reporting; ITIF (2025); Uptime Institute (2024); UK/Ireland standard BS 4142.

Singapore & e-waste

PUB and IMDA (2023-24); Wang et al., Nature Computational Science (2024).

Every figure in this briefing traces to one of these named sources, with its year. Where sources disagree, we showed the range.

AI Literacy Lab | Bykov-Brett Enterprises

WHERE TO GO NEXT

Use AI well. Know what it costs.

Every figure here is sourced inline and numbered. We kept the caveats in on purpose. A statistic without its qualifier is how good research becomes a bad headline.

Bykov-Brett Enterprises helps organisations use AI in ways they can stand behind, technically, commercially, and environmentally.

Work with us | See our resources

bykovbrett.net/resources
jamie@bykovbrett.net

PRIMARY SOURCES

IEA, Energy and AI (2025)

Li / Ren, Making AI Less Thirsty (2023)

Patterson et al. (2021, 2022)

Luccioni et al., BLOOM (2022)

JLARC, Data Centers in Virginia (2024)

Cardinal News, the 70% debunk (2025)

Masanet et al., Science (2020)

Uptime Institute survey (2024)

PUB and IMDA, Singapore (2023 to 2024)

Wang et al., Nature Comp. Science (2024)

CodeCarbon (codecarbon.io)

Google, Gemini impact disclosure (2025)

Richter et al., Colorado River water (2024)

Epoch AI, energy per query (2025)

Turn Intent Into Impact.