The Hidden Cost of AI

How much energy, carbon and water does artificial intelligence really consume?

Artificial intelligence may feel weightless, but every prompt, generated image and automated decision depends on physical infrastructure: processors, data centres, electricity grids, cooling systems and water supplies. As AI adoption accelerates, businesses must measure not only what AI produces, but also what it consumes.

2024 data-centre demand	2030 projection	Median text prompt	AI water withdrawal, 2027
415 TWh	945 TWh	0.24 Wh	4.2-6.6 bn m³

1. AI’s electricity demand is accelerating

The International Energy Agency estimates that global data centres consumed about 415 terawatt-hours of electricity in 2024. Under its base-case scenario, demand could reach roughly 945 TWh by 2030—an increase of about 128% in six years. AI is expected to be the largest driver of this growth.

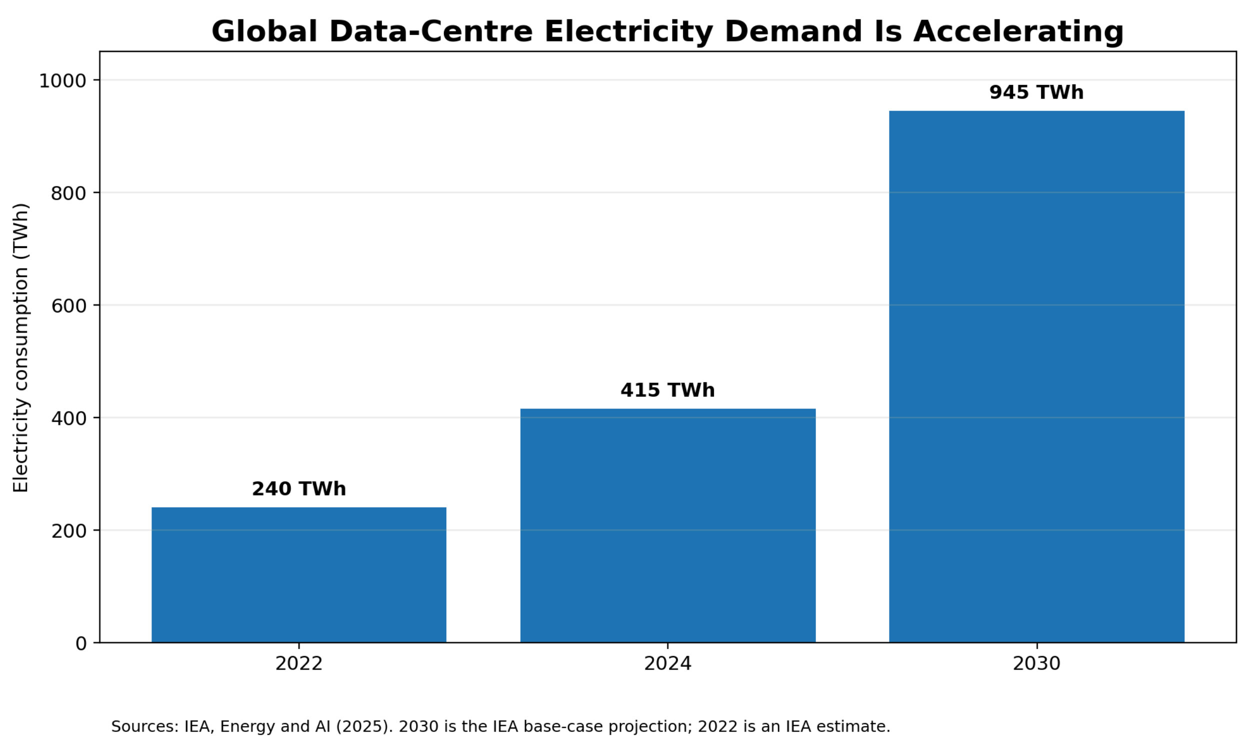

Figure 1. Global data-centre electricity demand. Source: International Energy Agency, Energy and AI (2025).

In the United States, data centres represented around 4.4% of electricity consumption in 2023. The US Department of Energy estimates that their share could rise to 6.7-12% by 2028. This growth places pressure not only on electricity generation, but also on transmission networks, local grids and energy prices.

Where does the electricity go?

- AI accelerators such as GPUs and TPUs

- Host processors, memory, storage and networking
- Cooling systems, pumps, chillers and fans
- Power conversion, backup systems and facility operations
- Idle capacity maintained for sudden peaks in demand

A complete calculation must therefore include more than the processor itself. Data-centre efficiency is commonly measured through Power Usage Effectiveness (PUE). A facility with a PUE of 1.2 consumes 1.2 units of total electricity for every unit used directly by computing equipment; a facility with a PUE of 1.8 consumes 1.8 units.

2. The upfront cost of training large language models

Training an LLM requires huge datasets to be processed repeatedly across thousands of specialised chips. The published model represents only part of the footprint: experimental runs, failed training jobs, evaluation, fine-tuning and safety testing also consume energy.

Meta's model cards provide one of the clearest public illustrations. Reported cumulative family-training figures rose from about 539 tonnes of location-based CO_2e for Llama 2 to 11,390 tonnes for Llama 3.3. The latter also involved approximately 39.3 million GPU-hours.

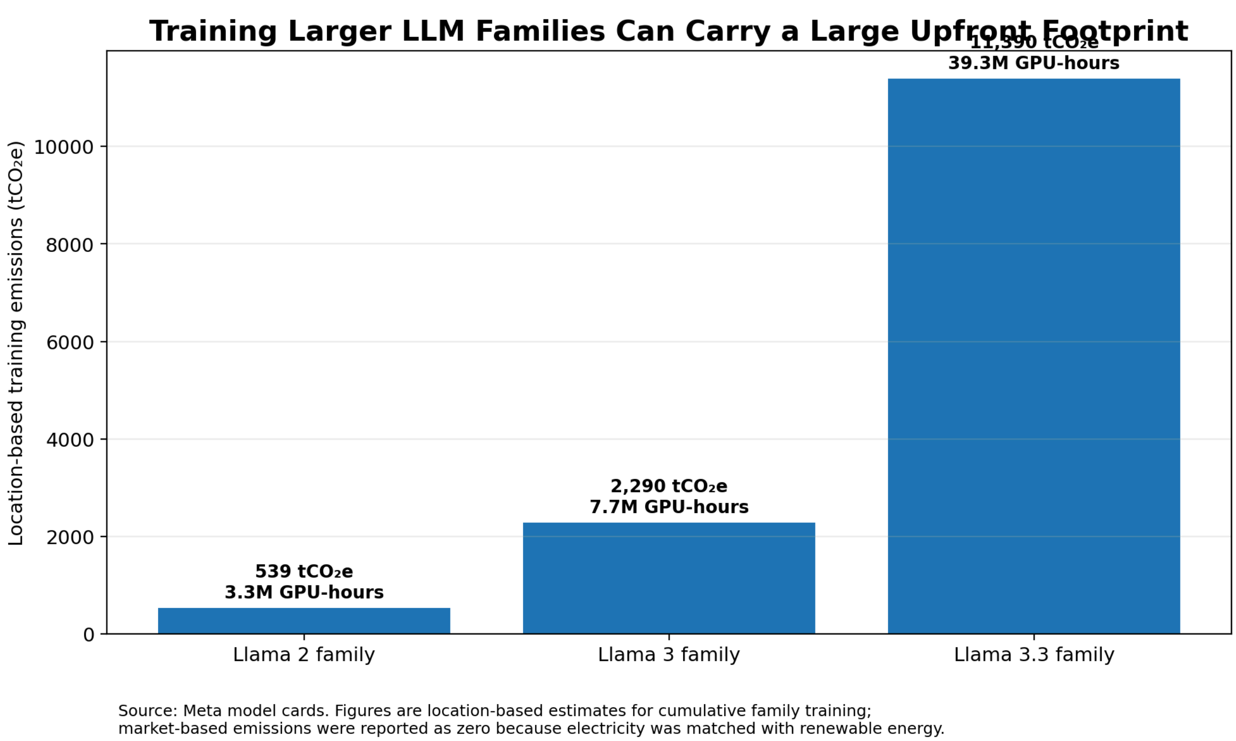

Figure 2. Disclosed location-based emissions for cumulative Llama family training. Source: Meta model cards.

The difference between 539 tonnes and 11,390 tonnes represents an increase of more than 21 times. These figures should not be treated as a simple model-quality ranking because the model families differ in size, training data and development scope. They do, however, show how quickly compute requirements can expand.

3. Inference: the cost that continues after training

Inference occurs whenever a model answers a user, summarises a document, writes code, generates an image or operates an AI agent. One request uses much less energy than training a model, but inference is repeated millions or billions of times.

A 2025 Google production study reported that the median Gemini Apps text prompt consumed approximately 0.24 watt-hours of electricity, produced around 0.03 grams of CO_2e and used about 0.26 millilitres of water. These are measured values for one production workload, not universal constants for every model.

How scale changes the calculation

Prompt volume	Illustrative electricity use at 0.24 Wh per prompt
1 prompt	0.24 Wh
1,000 prompts	0.24 kWh
1 million prompts	240 kWh
1 billion prompts	240 MWh
10 billion prompts	2.4 GWh

Actual consumption rises with longer prompts, larger models, reasoning steps, image or video generation, tool calls, long outputs and repeated regeneration. “One AI prompt” is therefore not a standard environmental unit.

4. Carbon emissions: operational, embodied and supply-chain

Operational emissions

These come from electricity used to run and cool AI infrastructure. The same workload can have very different emissions depending on whether the local grid is dominated by coal, gas, nuclear, hydropower, wind or solar energy.

Embodied emissions

Before a processor is switched on, emissions have already been generated through mining, semiconductor fabrication, server manufacturing, data-centre construction and equipment transport. These impacts are often excluded from per-prompt figures.

Supply-chain emissions

For companies purchasing cloud AI services, much of the footprint may fall within Scope 3 emissions. Moving computing to the cloud can reduce onsite energy use while shifting the environmental burden to a supplier. Procurement and ESG teams therefore need cloud-provider disclosures at model, region and workload level.

5. AI’s hidden water footprint

AI consumes water directly through data-centre cooling and indirectly through electricity generation. Evaporative cooling can reduce electricity demand but consume freshwater, creating a trade-off between carbon and water performance.

A widely cited academic study estimated that training GPT-3 in Microsoft’s US data centres could have directly evaporated approximately 700,000 litres of freshwater. The same study projected global AI water withdrawal of about 4.2-6.6 billion cubic metres in 2027. These are scenario-based estimates and vary with geography, cooling technology and power generation.

Water metric	Estimate	Important qualification
GPT-3 training	~700,000 litres	Scenario estimate for specific US facilities
Global AI, 2027	4.2-6.6 billion m³ withdrawal	Projection, not direct measurement
Median Gemini text prompt	~0.26 mL	Measured production workload; not universal

Water impact must be assessed in context. A litre consumed in a water-abundant location does not carry the same ecological and social consequence as a litre consumed in a drought-prone or water-stressed region.

6. Modern ways to reduce AI's footprint

Figure 3. Sustainable AI mitigation hierarchy: reduce demand first, then improve how remaining compute is supplied.

Use the smallest capable model

Simple classification, extraction, FAQs and document search often do not require a frontier model. Model routing can direct routine requests to smaller models and reserve high-compute systems for genuinely complex tasks.

Reduce tokens and repeated work

- Remove irrelevant prompt content
- Limit response length
- Retrieve only relevant document passages

- Cache common answers
- Summarise long conversation histories
- Prevent duplicated agent and API calls

Compress and optimise models

Quantisation, distillation, pruning, sparsity, speculative decoding and low-rank adaptation can reduce memory and compute requirements while maintaining acceptable performance for defined tasks.

Improve utilisation and scheduling

Batching requests, consolidating workloads and increasing accelerator utilisation reduce wasted energy. Flexible training and analytics jobs can also be shifted to cleaner-grid hours, cooler periods or regions with lower carbon intensity and water stress.

Modernise cooling and infrastructure

Direct-to-chip liquid cooling, immersion cooling, closed-loop systems, recycled water, free-air cooling and waste-heat recovery can improve performance. Each option should be assessed across both energy and water metrics to avoid shifting the burden from one resource to another.

Procure additional clean electricity

Annual renewable-energy certificates do not prove that a data centre operated on carbon-free electricity every hour. Stronger approaches include hourly matching, direct power-purchase agreements, storage, demand response and investment in new clean-energy capacity.

7. What corporate leaders should measure

Indicator	Why it matters
Energy per request or 1,000 tokens	Shows workload efficiency
Location-based carbon emissions	Reflects the physical grid supplying the facility
Market-based carbon emissions	Shows contractual energy procurement
Water Usage Effectiveness	Tracks water consumed relative to IT energy
Power Usage Effectiveness	Tracks total facility energy relative to IT energy
Hardware embodied emissions	Captures chip, server and construction impacts
Model-routing rate	Shows how often smaller models are used
Value per unit of compute	Links resource use to real business or social outcomes

Conclusion: from artificial intelligence to responsible intelligence

AI can support climate modelling, energy efficiency, precision agriculture, industrial optimisation and disaster response. But an application does not become sustainable simply because it is used for a sustainability-related purpose.

The future is not a choice between abandoning AI and accepting unlimited resource consumption. The opportunity is to build responsible intelligence: systems that create measurable economic and social value while consuming the least reasonable amount of energy, carbon, water and materials.

The next generation of AI leaders will not be defined by who deploys the largest model, but by who creates the greatest value from every unit of compute.

Sources

- International Energy Agency. Energy and AI (2025).
- US Department of Energy and Lawrence Berkeley National Laboratory. Data Centers and Electricity Demand.
- Meta AI model cards for Llama 2, Llama 3 and Llama 3.3.
- Google. Measuring the Environmental Impact of Delivering AI at Google Scale (2025).
- Li et al. Making AI Less Thirsty: Uncovering and Addressing the Secret Water Footprint of AI Models (2023).