# Computing energy and water requirements of commercial AI

*Training, inference and infrastructure at OpenAI, Anthropic, Google and the wider commercial AI ecosystem*

Revised edition — 7 October 2026

*Revision note. This edition preserves the September report's basic framework but updates the evidence through 7 October 2026. It adds new disclosures on hyperscale AI infrastructure, production-scale inference efficiency, direct versus indirect water use, supply-chain emissions, land and material footprints, and the rebound effect by which falling resource use per query can coexist with rising total demand.*

The enormous resource requirements of commercial AI apply both to training a model and to running it for users. They arise in different ways: training concentrates computation into creating or improving a model; inference uses computation every time the model processes an input and produces an answer; operating the service adds memory, networking, storage, spare capacity, power distribution, and cooling.

A single answer can consume relatively little electricity, while serving millions of people continuously requires enormous infrastructure. The International Energy Agency explicitly includes both training and deployment in its assessment of AI electricity demand. This report explains how the requirements change across the model lifecycle, how electricity and water are measured, and what the public disclosures for OpenAI, Anthropic, and Google can establish. [1 (https://www.iea.org/reports/energy-and-ai/executive-summary)]

If by **PoP** one means a **point of presence**, that is usually a network entry point near users. The demanding model computation normally happens deeper inside a data centre, on specialised servers. PoP describes a network location; training and inference describe what the computers are doing.

## Resource demands throughout the model lifecycle

| Stage | What happens | Main resources required | Pattern of use |
|---|---|---|---|
| 1 Data preparation | Collect, clean, filter, organise, and tokenise training material | Storage, networks, CPUs, and sometimes AI accelerators | Large processing campaigns |
| 2 Pretraining | Learn numerical parameters from enormous quantities of examples | Large accelerator clusters, fast memory, high-speed interconnections | Sustained computation, often over weeks or months |
| 3 Post-training | Improve instruction following, reasoning, behaviour, and safety | Accelerators, evaluation systems, human feedback, generated examples | Further training and experimentation |

| Stage | What happens | Main resources required | Pattern of use |
|---|---|---|---|
| 4 Evaluation and deployment preparation | Test the model and optimise it for serving | Inference and training hardware, storage, engineering systems | Repeated before and after release |
| 5 Inference | Process questions and generate responses | Accelerators, memory, CPUs, networking | Continuous, following demand |
| 6 Service operation | Maintain availability, store data, route requests, and remove heat | Supporting servers, storage, power equipment, cooling | Continuous |

These stages overlap at company level: one model may be answering customers while another is being trained. During pretraining, the model repeatedly makes predictions, measures its errors, and adjusts its parameters. During ordinary inference, it uses the parameters already learned.

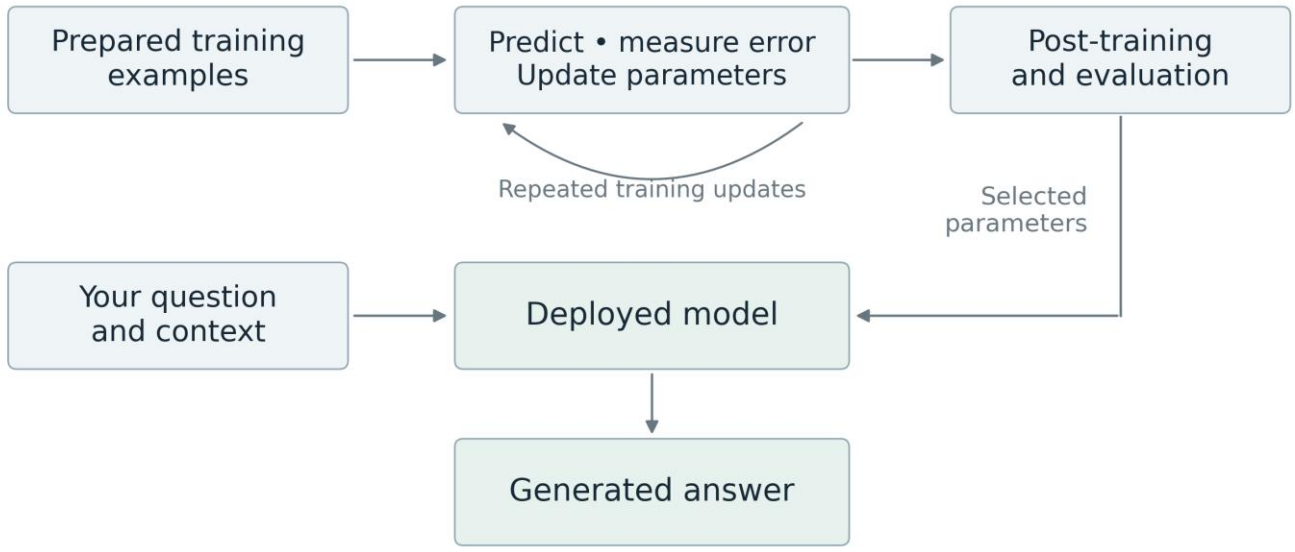



**Figure 1 From training examples to an answer. Schematic of the training loop and deployment process.**

The crucial distinction is that **answering a question does not require repeating the original training**. Nor does the model normally reread its entire training collection. It operates using its learned parameters and the information supplied in the current context. Searching the web or retrieving documents is an additional operation when the service uses those tools.

Post-training also deserves attention. Simple instruction tuning may be much smaller than pretraining, but advanced reinforcement learning involves generating and assessing many candidate solutions. It should not automatically be treated as a negligible finishing step. OpenAI reasoning research explicitly distinguishes additional computation spent during reinforcement learning from additional computation spent answering a question. [2 (https://openai.com/index/learning-to-reason-with-llms/)]

## Why training requires powerful computers

A large model contains billions of numerical parameters. Training requires calculating predictions, working backwards through the computation to determine parameter changes, and retaining substantial intermediate information.

The work must therefore be distributed across many chips. Those chips must exchange information quickly; otherwise, expensive processors spend their time waiting. The physical requirements include:

- **Arithmetic capacity:** performing huge numbers of matrix operations.
- **Memory capacity:** holding parameters, intermediate results, gradients, and training state.
- **Memory bandwidth:** moving those numbers quickly enough to keep processors occupied.
- **Interconnections:** exchanging information between chips and servers.
- **Reliable electricity and cooling:** keeping the entire installation operating.

This is why an AI training installation resembles a specialised supercomputer. Simply adding ordinary computers connected through a normal office network would create severe communication bottlenecks. NVIDIA describes how models are divided across devices to accommodate their computation and memory requirements. [3 (https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/)]

## What happens when the trained model answers

Running the trained model is easier than training it, but still involves substantial computation. An answer is calculated as it is generated. In a typical text model, inference has two main parts:

| Part | What the model does | Why resources are needed |
|---|---|---|
| Prefill | Processes the question, instructions, conversation context, and any supplied documents | Substantial computation over the input |
| Decoding | Generates successive output tokens | Repeated computation and movement of parameters and cached information through memory |

Input processing can exploit considerable parallelism. Generating an answer is more sequential because later tokens depend on earlier ones. Providers improve efficiency by batching different users' requests and caching intermediate results. [4 (https://jax-ml.github.io/scaling-book/inference/)]

Memory alone can be substantial. As an illustrative calculation, a model with **70 billion parameters stored at two bytes each requires about 140 GB just for its parameters**. That excludes working memory and conversation caches. Lower-precision representations can reduce the requirement.

A large model may consequently occupy several accelerators even when merely answering questions. The provider then runs many serving instances to handle concurrent users. There is no requirement for every question to involve the entire training cluster. A request is assigned to an appropriate serving system, which may itself span multiple chips.

## A point of presence and an AI server perform different jobs

A simplified request path separates the network functions from the principal model computation:

| Location | Typical role |
|---|---|
| Phone or PC | Display the interface and transmit the request |
| Network entry point or PoP | Receive traffic and provide network or security functions |
| Service infrastructure | Authenticate, assemble context, select a model, and route work |
| Accelerator servers | Perform the model's main computation |
| Supporting services | Retrieve documents, search, execute tools, or store results |

The network location nearest the user need not be where the model runs. Google's production study explicitly describes prompts arriving at its network edge and then being routed to AI computers inside its infrastructure. [5 (https://arxiv.org/html/2508.15734v1)]

Small models can also run on personal devices. Their local electricity use remains real, but they operate at a very different scale from a globally available frontier AI service.

## Reasoning and agents increase the work behind one question

Compare these requests:

- Translate this sentence.
- Analyse these documents and check your conclusions.
- Research this subject, compare sources, run calculations, and prepare a report.

The first may require a short model response. The others may involve lengthy input processing, internal reasoning, multiple model calls, searches, and tool execution.

A short final answer can therefore conceal a large amount of preceding computation. OpenAI demonstrated that reasoning performance could improve with additional computation at answer time, alongside improvements from further training. [2 (https://openai.com/index/learning-to-reason-with-llms/)]

An agent is particularly relevant here: one user instruction can initiate a sequence of operations. Counting only questions asked then becomes a poor way to estimate resource consumption.

## Computing performance electricity and water are different quantities

| Quantity | Common units | Meaning |
|---|---|---|
| Computing performance | FLOP/s | Arithmetic operations possible per second |
| Electrical power | W, kW, MW, GW | Rate of electricity consumption |
| Electrical energy | Wh, kWh, MWh, GWh | Electricity consumed over a period |

| Quantity | Common units | Meaning |
|---|---|---|
| Water withdrawal | Litres or cubic metres | Water taken from a supply |
| Water consumption | Litres or cubic metres | Water not returned to the immediate water environment, often because it evaporates |

The basic electricity calculation is:

$$\text{Energy} = \text{Average power} \times \text{Time}$$

For example, an installation averaging **10 MW for 1,000 hours consumes 10 GWh**.

Similarly, a facility actually drawing **1 GW continuously for a year would consume 8.76 TWh**. An announcement of 1 GW of capacity does not establish that it is already built, fully occupied, or operating continuously at that power.

## Cooling during inference and training

Almost all the electricity used by the computing equipment eventually becomes heat. Changing the workload from training to inference does not remove that heat.

There are two distinct cooling questions: **how is heat removed from the chips, and how is that heat ultimately released outside the building?**

At the chip, fans can move air over heatsinks, or liquid can circulate through sealed cold plates. AWS describes direct chip cooling in which coolant carries heat out of the server and then recirculates. [6 (https://www.aboutamazon.com/news/aws/aws-liquid-cooling-data-centers)]

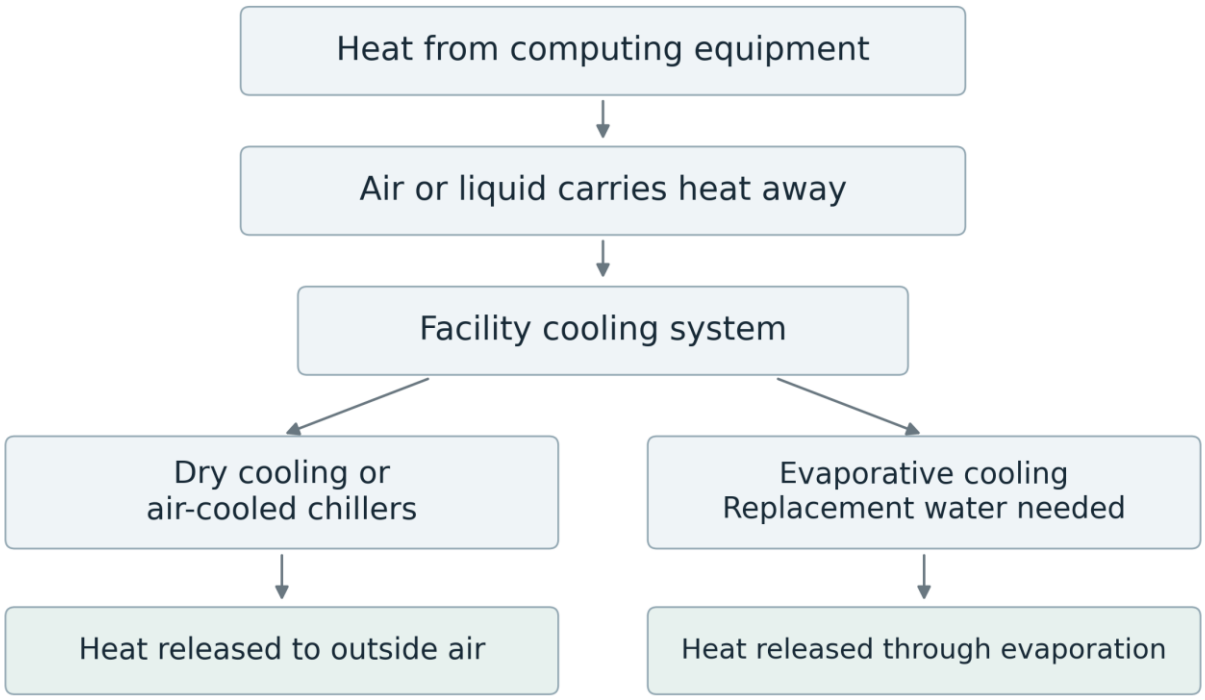



**Figure 2 Two ways to release heat outside a data centre. Simplified cooling paths showing why circulating liquid and consuming water are different.**

**Liquid circulating through pipes is not the same as water being consumed.** A sealed liquid loop can be coupled to dry heat rejection. Alternatively, a sealed loop inside the building can ultimately transfer its heat to an evaporative system outside.

Water also enters the environmental picture beyond the data centre:

- **Direct water use:** cooling and other operations at the facility.
- **Indirect water use:** generating its electricity.
- **Manufacturing water use:** producing chips, servers, and other equipment.

A facility with almost no direct cooling-water consumption can still have an indirect water footprint. Geography, weather, electricity supply, and cooling design all matter. [7 (https://arxiv.org/html/2304.03271v5)]

Cooling efficiency is often expressed through **PUE**, or **power usage effectiveness**:

$$\text{PUE} = \frac{\text{Total facility electricity}}{\text{IT equipment electricity}}$$

At PUE 1.2, every 100 kWh used by computing equipment corresponds to 120 kWh for the facility overall. The extra 20 kWh covers cooling and other infrastructure overhead. PUE describes facility efficiency; it does not tell us how efficiently the model answers a question. [8 (https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference)]

## Infrastructure at OpenAI Anthropic and Google

The three providers use the same broad physical principles, but different infrastructure arrangements. The examples below are documented installations and technologies, rather than an exhaustive inventory of each company's resources.

| AI provider | Documented infrastructure examples | Training and inference |
|---|---|---|
| OpenAI and ChatGPT | Microsoft Azure; Oracle infrastructure; NVIDIA-based Stargate capacity | Both |
| Anthropic and Claude | AWS Trainium infrastructure, including Project Rainier; Google TPUs | Both |
| Google and Gemini | Google data centres and custom Tensor Processing Units, or TPUs | Both |

OpenAI confirmed both workloads at Abilene; AWS describes both for Anthropic; Google describes TPUs supporting the training and serving of Gemini and Claude. [9 (https://openai.com/index/stargate-advances-with-partnership-with-oracle/), 10 (https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster), 11 (https://cloud.google.com/blog/products/compute/ironwood-tpus-and-new-axion-based-vms-for-your-ai-workloads)]

### OpenAI and the Abilene installation

Abilene in Texas provides a concrete example. In July 2025, OpenAI reported that the site had begun running early **training and inference workloads** using NVIDIA GB200 racks supplied by Oracle. Infrastructure of this scale serves both purposes. [9 (https://openai.com/index/stargate-advances-with-partnership-with-oracle/)]



**Figure 3 Server infrastructure pictured in Crusoe's account of Abilene. Photograph published by Crusoe with its Abilene facility description [12 (https://www.crusoe.ai/resources/blog/an-inside-look-at-the-abilene-ai-data-center)].**

Crusoe describes Abilene's design as a closed, non-evaporative system using air-cooled chillers. Its published plan specifies an initial water fill and smaller maintenance replenishment, while avoiding continuous water consumption for heat rejection. This illustrates why a large AI facility's electricity demand does not automatically imply equally large evaporative water consumption. [12 (https://www.crusoe.ai/resources/blog/an-inside-look-at-the-abilene-ai-data-center)]

## Anthropic and AWS Project Rainier

AWS's Project Rainier illustrates the scale of specialised computing. AWS describes a cluster spread across several US data centres containing nearly half a million Trainium2 chips. Its announcement explicitly covers building and deploying Claude, including training and inference. The chip count is a cluster resource; it is not the number used for an individual answer. [10 (https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster)]



**Figure 4  Data centre exterior shown in the Project Rainier announcement. Photograph by Amazon illustrating infrastructure supporting Anthropic [10 (https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster)].**

AWS also describes seasonal cooling at Rainier's Indiana site: extensive use of outside air, with cooling water needed during warmer periods. Thus, the same computing workload can have a different direct water requirement in winter and summer. [10 (https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster)]

## Google Gemini and TPU hardware

Google designs both the models and much of the underlying computing infrastructure. Its TPUs are specialised AI accelerators. The Ironwood generation supports pretraining, reinforcement learning, and inference, notwithstanding its strong emphasis on serving AI at scale. [11 (https://cloud.google.com/blog/products/compute/ironwood-tpus-and-new-axion-based-vms-for-your-ai-workloads)]

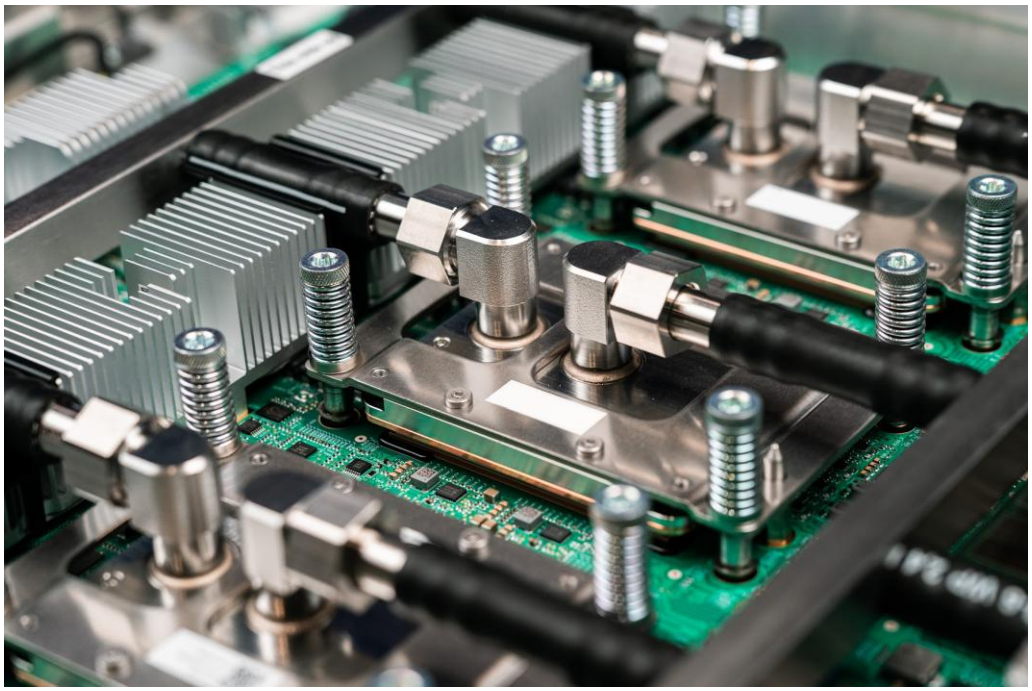


**Figure 5  Google Ironwood board with liquid cooling attached to its TPU modules. Photograph by Google showing cold plates and coolant hoses [11 (https://cloud.google.com/blog/products/compute/ironwood-tpus-and-new-axion-based-vms-for-your-ai-workloads)].**

The cold plates and hoses show how heat is removed close to the processors. Notice the different scales in these photographs: **a processor assembly, a server hall, and a campus**. Headlines frequently move between these scales without making the distinction clear.

## Published electricity and water figures per answer

The public figures below are dated disclosures from 2025. They are not measurements of every model or of the exact services running today.

| Service | Published electricity figure | Published water figure | What the disclosure represents |
|---|---|---|---|
| ChatGPT | About 0.34 Wh per query | About 0.32 mL per query | Sam Altman's statement about an average query; his post does not provide a detailed measurement methodology |
| Gemini Apps | 0.24 Wh per text prompt | 0.26 mL per text prompt | Google's reported median measure using May 2025 data |
| Claude | No comparable first-party figure located | No comparable first-party figure located | An equivalent public measurement was not established in this review |

Sources: Sam Altman's The Gentle Singularity and Google's August 2025 disclosure. The ChatGPT water value converts Altman's stated US gallons into millilitres. [13 (https://blog.samaltman.com/the-gentle-singularity), 8 (https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference)]

**These numbers do not establish that Gemini is more efficient than ChatGPT**, and missing Claude data do not imply low or high consumption. The workloads, model mixtures, statistical measures, and accounting boundaries differ.

Google's study is particularly useful because it separates the components of serving energy:

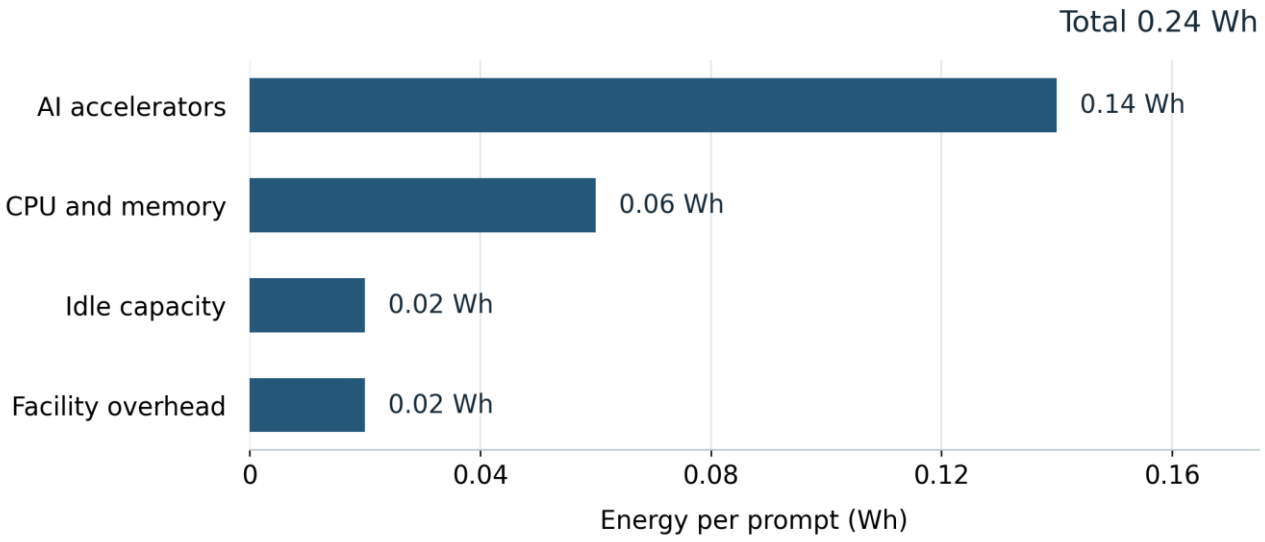



**Figure 6 Components of Google's reported 0.24 Wh per prompt. Median Gemini Apps text prompt in May 2025; data from Elsworth and colleagues, Table 1 [5 (https://arxiv.org/html/2508.15734v1)].**

The figure includes supporting computers and allocated idle capacity. Google's energy boundary excludes training and end-user devices; its water measure concerns data-centre consumption rather than electricity-generation or manufacturing water. [5 (https://arxiv.org/html/2508.15734v1)]

A particular request could differ substantially because of:

- The chosen model and how much of it is activated.
- The length of the input and output.
- Additional reasoning and repeated attempts.
- Tool calls or generated images and video.
- Hardware utilisation, batching, and caching.
- The facility and conditions under which the work runs.

A published median should not be multiplied by total traffic to estimate total electricity: that calculation requires an appropriate **arithmetic mean** or, preferably, measured total consumption.

## When continuous use overtakes model development

The following is a deliberately hypothetical example, not an estimate for any of the three companies:

- Developing a model consumes **10 GWh**.
- It subsequently handles **100 million requests per day**.
- Each request averages **0.3 Wh**, including the serving overhead being counted.

Daily serving electricity would be:

100,000,000 × 0.3 Wh = 30 MWh per day

At that rate, cumulative serving energy reaches the original 10 GWh after about **333 days**.

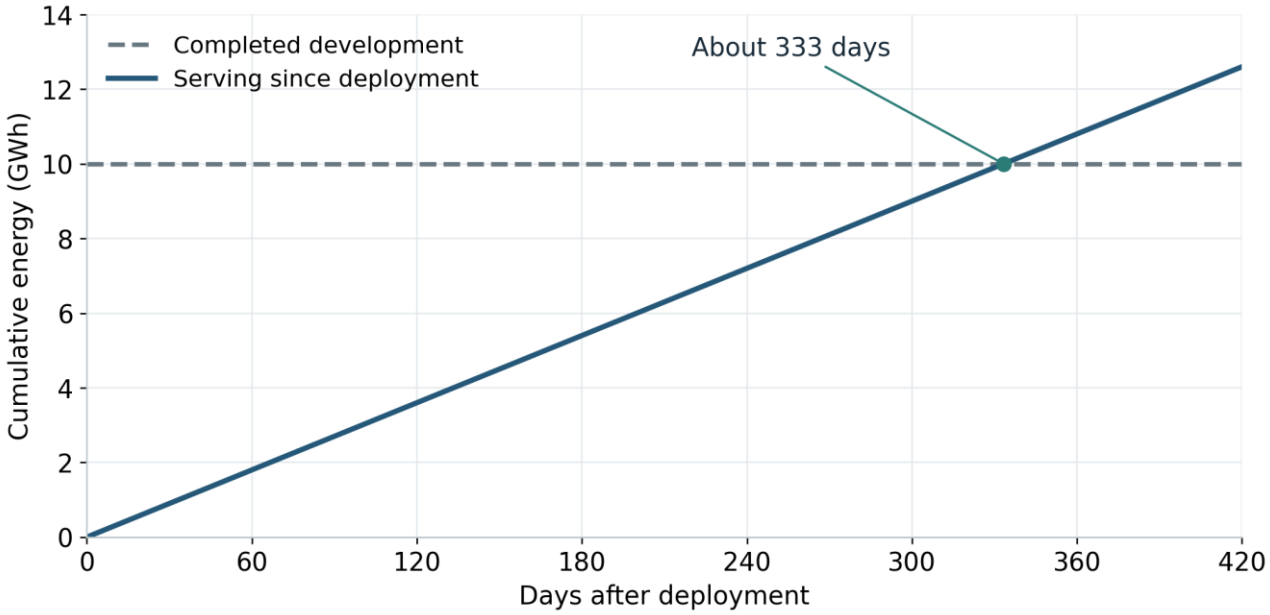



**Figure 7 How serving can overtake model development. Calculated example only; day zero is deployment and subsequent retraining is excluded.**

With ten times as much daily inference work, the crossover would occur about ten times sooner. With little use, or a much larger development expenditure, it could occur much later.

There is therefore **no universal percentage split between training and inference**. It depends on the model's development history, popularity, workload, efficiency, and useful lifetime. Companies also train successive models, so their development expenditure continues alongside serving.

## Global scale and interpretation

For perspective, the IEA's 2025 report estimated that **all data centres worldwide used 415 TWh in 2024** and projected **around 945 TWh in 2030**. Those totals include conventional cloud services and other computing, with AI identified as the largest driver of the projected increase. They are not totals for chatbots alone. [1 (https://www.iea.org/reports/energy-and-ai/executive-summary)]

The environmental consequences depend on where that demand occurs. A moderate global share can still create a large local requirement for electrical connections, generation, and water infrastructure.

When assessing a claim about AI resources, the most revealing questions are: **Which stage? Which workload? Over what period? At which location? And what has been included in the measurement?** A credible comparison of OpenAI, Anthropic, and Gemini would need matched tasks and quality levels, measured serving energy, separate training figures, and clearly defined direct and indirect water accounting.

# October 2026 update: what has changed since the first edition

## 1. The central conclusion remains valid

The September report's main distinction remains sound: training, post-training, inference, and service operation impose different resource demands, and the environmental footprint of a commercial AI service cannot be inferred from accelerator power alone. New production-scale measurements strengthen this conclusion. The main revision is therefore not a reversal but an expansion: AI is becoming simultaneously more efficient per unit of useful work and much larger as a physical infrastructure system.

## 2. Infrastructure scale has moved into the multi-gigawatt era

OpenAI's 2026 disclosures make the scale of frontier-AI infrastructure unusually concrete. In April, OpenAI said that Stargate had already surpassed its original objective of securing 10 GW of U.S. AI infrastructure by 2029. In July, Project Camellia in Georgia was described as a 3.2 GW power project, delivered in phases from 2028 to 2032. In August, OpenAI announced an agreement to secure approximately 8 GW-IT at the PORTS-Pike Technology Campus in Ohio. These are project or contracted-capacity figures, not measurements of continuous electricity consumption, but they show why gigawatts rather than megawatts are becoming the relevant unit for frontier-AI infrastructure planning.

## 3. Per-query inference energy is lower than many early estimates

Recent production-scale measurements indicate that ordinary text inference can be substantially more efficient than widely repeated early estimates. Google reported that a median Gemini Apps text prompt used about 0.24 Wh of energy, emitted about 0.03 g CO2e, and consumed about 0.26 mL of water under its comprehensive accounting methodology. Microsoft researchers reported a median of about 0.31 Wh per query (interquartile range 0.16–0.60 Wh) for optimized frontier-scale inference under realistic large-scale deployment assumptions. These numbers should not be treated as universal constants: energy varies with model, input and output length, reasoning effort, hardware, batching, utilization, and data-centre design.

## 4. Reasoning and agentic workloads change the inference picture

A short conversational query is increasingly a poor proxy for all AI inference. Microsoft's 2026 analysis estimates that long reasoning and agentic queries can increase energy consumption by more than an order of magnitude because they generate many more tokens and can reduce serving concurrency. In one test-time scaling scenario with roughly 15 times more tokens, median energy rose to about 4.32 Wh, around 13 times the typical-query figure. Consequently, future environmental accounting should report workload classes rather than a single number labelled 'energy per AI query'.

## 5. Water accounting needs a three-part distinction

Water figures are particularly easy to misinterpret. A revised analysis should distinguish (a) direct on-site water used by the data centre, principally for cooling; (b) indirect water associated with electricity generation; and (c) broader lifecycle water associated with semiconductor fabrication, construction and supply chains. Some new facilities use closed-loop or very low-water cooling, which can greatly reduce direct operational consumption. That does not eliminate the indirect and lifecycle water footprint. A 2025

Nature Sustainability analysis of U.S. AI-server expansion estimated that indirect water could represent about 71% of the modelled total water footprint, illustrating why on-site water figures alone are insufficient.

## 6. Efficiency and total consumption can move in opposite directions

The revised report should make explicit an apparent paradox. Model architectures, serving software, batching, routing, and new accelerators can reduce energy per query dramatically. Microsoft estimates that combined near-term improvements across model design, serving and hardware could reduce energy per query by roughly 8–20 times in the configurations studied. Yet lower cost and higher capability make many more AI uses economically attractive. Total demand can therefore rise even while each unit of work becomes cheaper and more efficient. This is a rebound or Jevons-type effect, and it is central to interpreting the environmental consequences of AI.

## 7. Supply-chain emissions deserve more weight

Operational electricity is only part of the footprint. Google's 2026 Environmental Report states that it signed agreements for more than 12 GW of net-new clean energy in 2025 and reduced operational emissions by 2% despite substantial electricity-demand growth. At the same time, its supply-chain emissions increased 25% year over year, reflecting in part the scale of AI infrastructure expansion and carbon-intensive manufacturing supply chains. This strengthens the case for lifecycle accounting that includes semiconductor fabrication, servers, networking equipment, buildings, construction and replacement cycles.

## 8. Carbon is not the only environmental metric

The United Nations University's 2026 report, Environmental Cost of Artificial Intelligence: Carbon, Water, and Land Footprints, argues for a broader physical accounting of AI. Electricity sources differ not only in carbon intensity but also in water and land requirements; a low-carbon option is not automatically low-water or low-land. The report also highlights critical-mineral supply chains, e-waste, local resource competition and environmental-justice questions. These dimensions should be treated as complements to energy accounting rather than as interchangeable impacts.

## 9. Scenario estimates remain uncertain but policy-relevant

A Nature Sustainability study modelling U.S. AI-server deployment from 2024 to 2030 estimated annual water footprints of roughly 731–1,125 million cubic metres and additional annual carbon emissions of about 24–44 Mt CO2e, depending on expansion scenarios. The authors emphasize deep uncertainty arising from deployment scale, server location, grid decarbonization, efficiency, cooling and utilization. These values are projections rather than observed industry totals; their value lies in showing the plausible magnitude and the sensitivity of outcomes to engineering and policy choices.

## 10. Electricity-system effects are becoming part of AI policy

Anthropic stated in February 2026 that training a single frontier model would soon require gigawatts of power and projected that the U.S. AI sector would need at least 50 GW of capacity over the following several years. The 50 GW figure is Anthropic's projection, not an independently established consensus forecast, but the policy issue is real: large AI loads can require new generation, transmission, substations

and grid upgrades. Companies are increasingly making commitments about paying project-specific infrastructure costs and avoiding cost transfers to residential ratepayers.

## A revised framework for measuring commercial AI

| Layer | What should be measured | Typical unit | Why it matters |
|---|---|---|---|
| Compute | Accelerators, CPUs, memory, networking; utilization and idle capacity | GPU/TPU-hours, FLOPs, utilization | Hardware power alone does not describe service energy. |
| Electricity | IT load plus power conversion and facility overhead | kWh; PUE | Connects computation to grid demand. |
| Direct water | On-site cooling and humidification | L; WUE | Highly dependent on cooling design and climate. |
| Indirect water | Water associated with electricity generation | L/kWh; total L | Can exceed direct cooling water. |
| Carbon | Operational and purchased-energy emissions | kg or Mt CO2e | Depends strongly on grid and procurement. |
| Embodied footprint | Chips, servers, buildings and construction | kg CO2e; materials | Becomes more important during rapid build-out. |
| Land and materials | Site footprint, generation footprint, minerals and supply chains | area; tonnes/material flows | Captures impacts missed by carbon-only accounting. |
| Service demand | Queries, tokens, modalities, reasoning/agentic work | queries/day; tokens; workload mix | Determines whether efficiency gains reduce or merely slow total growth. |

## The efficiency–growth paradox

The most important conceptual addition to the September report is the distinction between intensity and scale. Resource intensity asks how much electricity, water or carbon is associated with a unit of AI work. Scale asks how much AI work society performs. The total footprint is approximately the product of the two. If energy per useful task falls by a factor of ten while the amount of AI work rises by a factor of

twenty, total energy use still doubles. Conversely, efficiency can prevent an even larger increase. This is why statements such as ‘a prompt uses only a fraction of a watt-hour’ and ‘AI data centres require gigawatts’ are not contradictory: they refer to different levels of analysis.

The effect is especially relevant as AI shifts from short question-answer interactions toward long-context reasoning, coding agents, multimodal generation, autonomous research and persistent enterprise workflows. These applications may perform many model calls for a single user-visible task. A future report should therefore increasingly use ‘energy per completed task’ or ‘energy per unit of useful work’ alongside energy per token or per prompt.

## Updated conclusions

- The original report remains broadly correct, but commercial AI infrastructure is scaling faster than a September snapshot alone conveys.
- Production-scale text inference can be sub-watt-hour per ordinary query, while long reasoning and agentic workloads can be an order of magnitude more demanding.
- Direct cooling water should never be equated with total water footprint; indirect electricity-related and lifecycle water must be separated.
- Per-query efficiency is improving rapidly, but aggregate demand can still grow through rebound effects, expanding usage and more compute-intensive workloads.
- Lifecycle impacts—especially semiconductor/server manufacture, construction, supply chains, land and materials—deserve a larger role in environmental accounting.
- Provider disclosures are valuable but should be distinguished from independent measurements and scenario studies; uncertainty should be reported explicitly.

## References

## Sources and photograph credits

[1] International Energy Agency. 2025. Energy and AI (https://www.iea.org/reports/energy-and-ai/executive-summary).

[2] OpenAI. 12 September 2024. Learning to reason with LLMs (https://openai.com/index/learning-to-reason-with-llms/).

[3] NVIDIA. 17 November 2023. Mastering LLM Techniques Inference Optimization (https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/).

[4] JAX scaling book. Accessed 11 September 2026. All About Transformer Inference (https://jax-ml.github.io/scaling-book/inference/).

[5] Elsworth and colleagues. 21 August 2025. Measuring the environmental impact of delivering AI at Google Scale (https://arxiv.org/html/2508.15734v1).

[6] Amazon Web Services. Accessed 11 September 2026. AWS rolls out liquid cooling in data centers (https://www.aboutamazon.com/news/aws/aws-liquid-cooling-data-centers).

[7] Li and colleagues. Manuscript originally published in 2023 version 5. Making AI Less Thirsty Uncovering and Addressing the Secret Water Footprint of AI Models (https://arxiv.org/html/2304.03271v5).

[8] Google Cloud. 21 August 2025. How much energy does Google AI use We did the math (https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference).

[9] OpenAI. 22 July 2025. Stargate advances with 4.5 GW partnership with Oracle (https://openai.com/index/stargate-advances-with-partnership-with-oracle/).

[10] Amazon Web Services. Announcement published October 2025. AWS activates Project Rainier One of the largest AI compute clusters comes online (https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster).

[11] Google Cloud. 6 November 2025. Announcing Ironwood TPUs General Availability and new Axion VMs to power the age of inference (https://cloud.google.com/blog/products/compute/ironwood-tpus-and-new-axion-based-vms-for-your-ai-workloads).

[12] Crusoe. 5 August 2025. An inside look at the Abilene AI data center (https://www.crusoe.ai/resources/blog/an-inside-look-at-the-abilene-ai-data-center).

[13] Sam Altman. 10 June 2025. The Gentle Singularity (https://blog.samaltman.com/the-gentle-singularity).

Newer references

1. OpenAI. (29 April 2026). Building the compute infrastructure for the Intelligence Age. https://openai.com/index/building-the-compute-infrastructure-for-the-intelligence-age/
2. OpenAI. (22 July 2026). Building AI infrastructure with the Effingham County community (Project Camellia). https://openai.com/index/building-ai-infrastructure-with-the-effingham-county-community/
3. OpenAI. (17 August 2026). OpenAI joins PORTS-Pike project. https://openai.com/index/openai-joins-ports-pike-project/
4. Google. (30 June 2026). 2026 Environmental Report / 11th annual Environmental Report. https://blog.google/company-news/outreach-and-initiatives/sustainability/2026-environmental-report/
5. Vahdat, A., & Dean, J. (21 August 2025). Measuring the environmental impact of AI inference. Google Cloud. https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference
6. Oviedo, F., et al. (2026). Energy use of AI inference, efficiency pathways, and test-time scaling. Joule. Microsoft Research publication page: https://www.microsoft.com/en-us/research/publication/energy-use-of-ai-inference-efficiency-pathways-and-test-time-scaling/
7. Anthropic. (11 February 2026). Covering electricity price increases from our data centers. https://www.anthropic.com/news/covering-electricity-price-increases
8. Aczel, M., Chamanara, S., Matin, M., Farsi, A., Marwala, T., & Madani, K. (2026). Environmental Cost of Artificial Intelligence: Carbon, Water, and Land Footprints. United Nations University Institute for Water, Environment and Health. https://doi.org/10.53328/INR26RMA002
9. Li, P., et al. (2025). Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. Nature Sustainability. https://doi.org/10.1038/s41893-025-01681-y