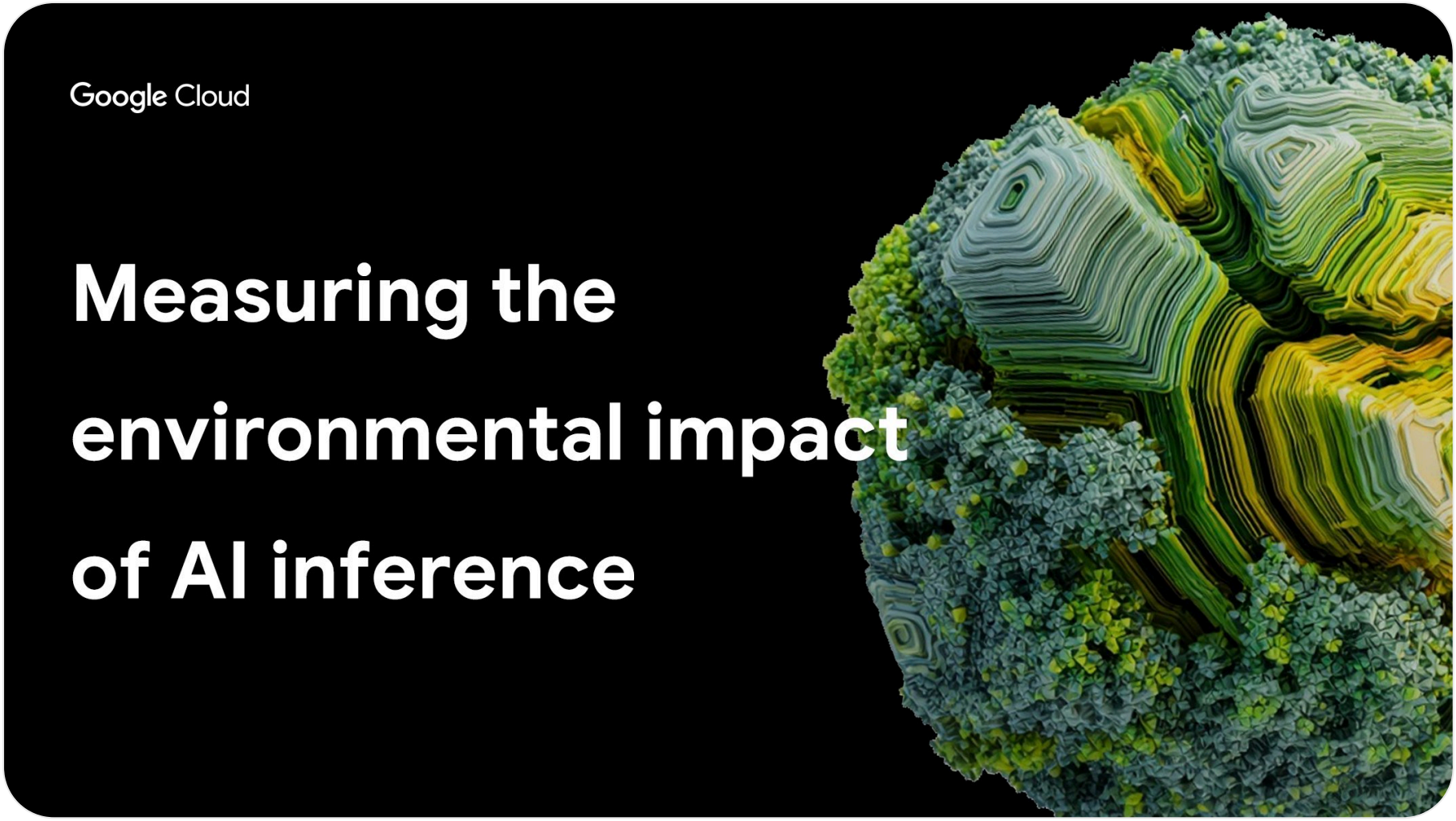
Infrastructure

# How much energy does Google's AI use? We did the math

August 22, 2025




**Amin Vahdat**
VP/GM, AI & Infrastructure, Google Cloud

**Jeff Dean**
Chief Scientist, Google DeepMind and Google Research

AI is unlocking scientific breakthroughs, improving healthcare and education, and could add trillions to the global economy. Understanding AI's footprint is crucial, yet thorough data on the energy and environmental impact of AI inference — the use of a trained AI model to make predictions or generate text or images — has been limited. As more users use AI systems, the importance of inference efficiency rises.

That's why we're releasing a technical paper detailing our comprehensive methodology for measuring the energy, emissions, and water impact of Gemini prompts. Using this methodology, we estimate the median Gemini Apps text prompt uses 0.24 watt-hours (Wh) of energy, emits 0.03 grams of carbon dioxide equivalent ($gCO_2e$), and consumes 0.26 milliliters (or about five drops) of water[1] — figures that are substantially lower than many public estimates. The per-prompt energy impact is equivalent to watching TV for less than nine seconds.

At the same time, our AI systems are becoming more efficient through research innovations and software and hardware efficiency improvements. For example, over a recent 12 month period, the energy and total carbon footprint of the median Gemini Apps text prompt dropped by 33x and 44x, respectively, all while delivering higher quality responses. These results are built on our latest data center energy emissions reductions and our work to advance carbon-free energy and water replenishment. While we're proud of the innovation behind our

efficiency gains so far, we're committed to continuing substantial improvements. Here's a closer look at these ongoing efforts.


## Calculating the environmental footprint of AI at Google

Detailed measurement lets us compare across different AI models, and the hardware and energy they run on, while enabling system-wide efficiency optimizations — from hardware and data centers to the models themselves. By sharing our methodology, we hope to increase

industry-wide consistency in calculating AI's resource consumption and efficiency.

Measuring the footprint of AI serving workloads isn't simple. We developed a comprehensive approach that considers the realities of serving AI at Google's scale, which include:

- **Full system dynamic power:** This includes not just the energy and water used by the primary AI model during active computation, but also the actual achieved chip utilization at production scale, which can be much lower than theoretical maximums.
- **Idle machines:** To ensure high availability and reliability, production systems require a degree of provisioned capacity that is idle but ready to handle traffic spikes or failover at any given moment. The energy consumed by these idle chips must be factored into the total energy footprint.
- **CPU and RAM**: AI model execution doesn't happen solely in ML accelerators like TPUs and GPUs. The host CPU and RAM also play a crucial role in serving AI, and use energy.
- **Data center overhead:** The energy consumed by the IT equipment running AI workloads is only part of the story. The infrastructure supporting these computations — cooling systems, power distribution, and other data center overhead — also consumes energy. Overhead energy efficiency is measured by a metric called Power Usage Effectiveness (PUE).

- **Data center water consumption**: To reduce energy consumption and associated emissions, data centers often consume water for cooling. As we optimize our AI systems to be more energy-efficient, this naturally decreases their overall water consumption as well.

Many current AI energy consumption calculations only include active machine consumption, overlooking several of the critical factors discussed above. As a result, they represent theoretical efficiency instead of true operating efficiency at scale. When we apply this non-comprehensive methodology that only considers active TPU and GPU consumption, we estimate the median Gemini text prompt uses 0.10 Wh of energy, emits 0.02 $gCO_2e$, and consumes 0.12 mL of water. This is an optimistic scenario at best and substantially underestimates the real operational footprint of AI.

Our comprehensive methodology's estimates (0.24 Wh of energy, 0.03 $gCO_2e$, 0.26 mL of water) account for all critical elements of serving AI globally. We believe this is the most complete view of AI's overall footprint.

## Our full-stack approach to AI — and AI efficiency

Gemini's dramatic efficiency gains stem from Google's full-stack approach to AI development — from custom hardware and highly efficient models, to the robust serving systems that make these models possible. We've built efficiency into every layer of AI, including:

- **More efficient model architectures:** Gemini models are built on the Transformer model architecture developed by Google researchers, which provide a 10-100x efficiency boost over the previous state-of-the-art architectures for language modeling. We design models with inherently efficient structures like Mixture-of-Experts (MoE) and hybrid reasoning. MoE models, for example, allow us to activate a small subset of a large model specifically required to respond to a query, reducing computations and data transfer by a factor of 10-100x.
- **Efficient algorithms and quantization:** We continuously refine the algorithms that power our models with methods like Accurate Quantized Training (AQT) to maximize efficiency and reduce energy consumption for serving, without compromising response quality.
- **Optimized inference and serving:** We constantly improve AI model delivery for responsiveness and efficiency. Technologies like speculative decoding serve more responses with fewer chips by allowing a smaller model to make predictions that are then quickly verified by a larger model, which is more efficient than having the larger model make many sequential predictions on its own. Techniques like distillation create smaller, more efficient models

(Gemini Flash and Flash-Lite) for serving that use our larger, more capable models as teachers. Faster machine learning hardware and models enable us to use more efficient larger batch sizes when handling requests, while still meeting our latency targets.

- **Custom-built hardware:** We've been designing our TPUs from the ground up for over a decade to maximize performance per watt. We also co-design our AI models and TPUs, ensuring our software takes full advantage of our hardware — and that our hardware is able to efficiently run our future AI software when both are ready. Our latest-generation TPU, Ironwood, is 30x more energy-efficient than our first publicly-available TPU and far more power-efficient than general-purpose CPUs for inference.
- **Optimized idling:** Our serving stack makes highly efficient use of CPUs and minimizes TPU idling by dynamically moving models based on demand in near-real-time, rather than using a "set it and forget" approach.
- **ML software stack:** Our XLA ML compiler, Pallas kernels, and Pathways systems enable model computations expressed in higher-level systems like JAX to run efficiently on our TPU serving hardware.
- **Ultra-efficient data centers:** Google's data centers are among the industry's most efficient, operating at a fleet-wide average PUE of 1.09.

- **Responsible data center operations:** We continue to add clean energy generation in pursuit of our 24/7 carbon-free ambition, while advancing our aim to replenish 120% of the freshwater we consume on average across our offices and data centers. We also optimize our cooling systems, balancing the local trade-off between energy, water, and emissions, by conducting science-backed watershed health assessments, to guide cooling type selection and limit water use in high-stress locations.

## Our commitment to efficient AI

Gemini's efficiency gains are the result of years of work, but this is just the beginning. Recognizing that AI demand is growing, we're heavily investing in reducing the power provisioning costs and water required per prompt. By sharing our findings and methodology, we aim to drive industry-wide progress toward more efficient AI. This is essential for responsible AI development.

---

*1. A point-in-time analysis quantified the energy consumed per median Gemini App text-generation prompt, considering data from May 2025. Emissions per prompt was estimated based on energy per prompt, and applying Google's 2024 average fleetwide grid carbon intensity. Water consumption per prompt was estimated based on energy per prompt, and applying Google's 2024 average fleetwide water usage effectiveness. These findings do not*

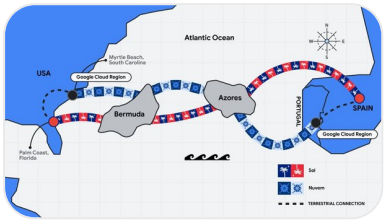
*represent the specific environmental impact for all Gemini App text-generation prompts nor are they indicative of future performance.*

*2. The results of the above analysis from May 2025 were compared to baseline data from the median Gemini App text-generation prompt in May 2024. Energy per median prompt is subject to change as new models are added, AI model architecture evolves, and AI chatbot user behavior develops. The data and claims have not been verified by an independent third-party.*

---

Posted in Infrastructure—AI & Machine Learning—Sustainability

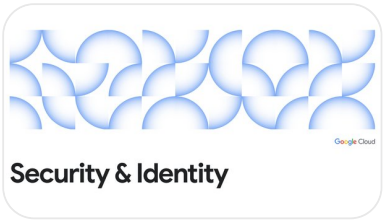
## Related articles




Infrastructure

### Strengthening network resilience with the Sol transatlantic cable

By Brian Quigley • 4-minute read




Security & Identity

### Just say no: Build defense in depth with IAM Deny and Org Policies

By Kevin Schmidt • 8-minute read


Infrastructure

### Hej Sverige! Google Cloud launches new region in Sweden

By Tara Brady • 6-minute read


Systems

### How we got to 100 million cells in our global Li-ion rack battery fleet

By Christina Peabody • 3-minute read