

AI Energy Per Query: 5 Myths, From Google's 0.24 Wh to the 27x Gap Nobody Reports

Google reports a median Gemini text prompt at 0.24 Wh. A study published November 3, 2026 measured 86 open-weight checkpoints and found configurations hitting the same benchmark score while burning 27x more energy than each other.

THE SHORT VERSION

Per-prompt energy numbers like Google's 0.24 Wh (August 21, 2025) and Sam Altman's 0.34 Wh for ChatGPT (June 2025) describe one vendor's deployment on one day. They are not a property of a model. Across 86 open-weight checkpoints and 136 serving configurations, the November 3, 2026 study found a 27x spread between setups reaching the same score, with a median 2.6 to 4.8x excess for configurations off the efficiency frontier. Energy per query is something you configure.

A per-prompt number that is 27x wrong will not survive your electricity bill

If you are a CTO putting an AI inference workload into a sustainability report, a procurement comparison, or a capacity plan, the gap between 0.24 Wh and whatever your stack actually draws is not a rounding error. The published measurement studies put the plausible spread at one to two orders of magnitude depending on serving configuration, hardware utilisation, task type and whether reasoning mode is switched on. A number that is 27x wrong does not survive contact with a real electricity bill or a real GPU reservation.

There is a second cost, less visible. If you believe per-query energy is fixed by the model, you stop looking for the lever. The measurement work from the last two months says the lever is large, it sits in your deployment configuration, and most people are not pulling it. That is the useful finding, more useful than any single Wh figure.

Myth one: 0.24 Wh is what a prompt costs

THE CLAIM	THE REALITY
Google measured 0.24 Wh, 0.03 gCO□e and 0.26 mL of water per median Gemini Apps text prompt, so that is roughly what an AI query costs.	Google's figure, published August 21, 2025 in its technical paper on AI at Google scale, is a median over Google's own traffic, on Google's own hardware, at Google's own utilisation. Google also reported that the median fell 33x in twelve months (May 2024 to May 2025) and total carbon footprint 44x. A quantity that moved 33x in a year through engineering work is a deployment outcome, not a constant.

I want to be fair to Google here: publishing a full-stack figure including idle machines and water is more disclosure than almost anyone else offered, and the paper is specific about scope. The problem is downstream. The number escaped the paper and became a universal per-prompt coefficient in slide decks about systems that share almost nothing with Gemini Apps serving.

Note also what "median text prompt" excludes. The "Where Do the Joules Go?" study (46 models, 7 tasks, 1,858 configurations on H100 and B200 GPUs, updated November 3, 2026) found task choice alone drove up to 25x energy differences, and video generation sometimes consumed more than 100x the energy of image generation. A median across a text-prompt distribution tells you nothing about the request your product actually sends.

Myth two: two vendors agreeing means the number is validated

THE CLAIM	THE REALITY
Altman said ~0.34 Wh per average ChatGPT query in June 2025, Epoch AI independently estimated ~0.3 Wh for a typical GPT-4o text query, and Google landed at 0.24 Wh. Three figures clustering around 0.3 Wh means we know the answer.	Convergence on text queries to then-current chat models says nothing about the general case. As reported August 9, 2025, OpenAI did not disclose GPT-5's energy use and it could be higher than past models. The Epoch estimate was for GPT-4o specifically, and the agreement does not validate the figure for later or reasoning models.

Three estimates of the same narrow thing agreeing is weak evidence that the thing is stable, and no evidence at all about adjacent things. The open-weight study makes the point concretely: enabling "thinking" mode raised energy use by roughly 3.7x on average, varying substantially by model family and output budget. Apply a 0.3 Wh assumption to a reasoning-heavy workload and you are out by a factor before you have considered anything else.

Myth three: the fastest GPU is the cheapest per token

THE CLAIM	THE REALITY
Newer accelerators are more efficient per token, so moving to the latest hardware reduces both cost and energy.	InferenceProfiler (University of Washington, WOSAI 2026, October 1, 2026) found the newest or fastest GPU is not automatically cheapest: at 1,000 concurrent requests, an A10G instance hit the lowest cost at $0.1829 per million tokens. Energy told a different story again: 35.7 kWh per million tokens on the A10G configuration against 28.4 kWh per million tokens on an L40S configuration.

That last pair of numbers is the one I would put in front of a finance team. The cheapest configuration by dollar was not the cheapest by joule. If your optimisation target is cost, you land on one instance type; if it is energy or carbon, you land on another. Many organisations have never decided which target they are optimising, so they optimise neither and inherit whatever default their platform team picked.

Does the same model on the same GPU always use the same energy?

No. The "Where Do the Joules Go?" measurements found GPU utilisation differences alone produced 3 to 5x energy swings, making the same model on the same hardware a variable rather than a constant. Batch size, concurrency, and how long cards sit warm and idle between requests move the per-token figure by multiples before model choice enters the picture.

This is where the single-number framing breaks hardest. A per-query Wh value implies the denominator is the query. In practice the denominator is a serving system with a utilisation profile, and the same query costs several times more on a lightly loaded cluster than on a saturated one. Anyone who has sized a GPU fleet for peak traffic knows where that leaves the average.

So the metric worth reporting is not a per-prompt Wh figure at all. Report energy per million output tokens for your own deployment, measured at your own observed concurrency, with reasoning mode state recorded alongside it. The InferenceProfiler work shows that number is obtainable from instance-level measurement, and the open-weight study shows it is the number that actually varies.

Myth four: regulation will force the numbers out

THE CLAIM	THE REALITY
The EU AI Act's GPAI obligations include energy documentation, so comparable per-inference figures are coming and there is no point building your own measurement now.	The GPAI energy-consumption documentation duties applied from August 2, 2025, with added AI Office enforcement powers from August 2, 2026, but there is no universal per-inference Wh disclosure rule. Documentation duties on providers are not a standardised, comparable, per-query metric you can drop into a procurement comparison.

Myth five: most deployments are already near optimal

Inference serving has been heavily optimised by vendors and frameworks, the argument goes, so the remaining headroom is marginal. The November 3, 2026 open-weight study found the opposite. Most of its 136 serving configurations sat off the score to energy Pareto frontier, using a median 2.6 to 4.8x more energy than another configuration achieving at least the same benchmark score. The largest measured gap was 27x. **The headroom is the default case, not the exception.**

Read that carefully, because it is the single most actionable result in the whole set. These are not configurations trading quality for efficiency. They score the same and burn multiples more. Someone chose a serving setup, it worked, and nobody measured the alternative that would have produced identical output for a fraction of the power.

What holds up as of November 2026

Here is what I think survives scrutiny if you have to defend it in a board paper.

Per-query energy for a mainstream text chat prompt on a well-utilised hyperscaler stack is plausibly in the low hundreds of milliwatt-hours: Google's 0.24 Wh and Altman's 0.34 Wh both sit there, and Epoch's ~0.3 Wh estimate for GPT-4o is consistent with them. That band is real but narrow in scope. It applies to short text prompts, to those vendors, at those dates.

Outside that band, the variation is enormous and systematic. Reasoning mode multiplies energy by roughly 3.7x on average. Task type varies by up to 25x, with video generation exceeding image generation by 100x or more. Utilisation moves it 3 to 5x. Serving configuration at equal quality moves it up to 27x. These stack. They do not neatly multiply, since the studies measured them separately on different hardware and benchmarks, but a workload that is reasoning-heavy, running on a poorly utilised cluster, with an unexamined serving configuration, is nowhere near a factor of two of 0.24 Wh. I would not put a specific multiplier on the stack without measuring it, and neither should anyone selling you one.

The other thing that holds: measurement scope is now good enough that "we don't know" is no longer an honest answer for your own system. The Watt Counts benchmark covers 50 LLMs across 10 NVIDIA GPUs and 5 architectures, with over 5,000 experiments, 370 model to GPU pairs and more than 14 million power-draw samples. The methodology exists. It is published. The reason your energy figure is a guess is that nobody has been asked to produce a real one.

Why a 27x finding lost to a 0.24 Wh finding

My take: single numbers win because they are quotable and because the people who need them are not the people who can measure them. A sustainability lead needs one coefficient for a spreadsheet; a procurement officer needs one row for a comparison. "It depends on your serving configuration by up to 27x" is true and unusable in both contexts, so the usable falsehood circulates and the accurate finding stays in the paper. I also suspect vendor incentives play a part: a low, stable, flattering per-prompt figure is a better communications asset than an honest range, and Google's own 33x year-over-year improvement claim is more persuasive precisely because the endpoint is a single clean number.

So treat energy per query the way you already treat latency. Nobody accepts a vendor's single global latency figure for their own p99. You measure it on your traffic, you track it per configuration change, and you know which knob moved it. The InferenceProfiler and Watt Counts work give you the measurement shape to copy.

Three practical consequences I would act on. First, record reasoning mode state in your telemetry, because a 3.7x average multiplier that varies by model family and output budget

is the largest single thing most teams are not tracking. Second, before you upgrade hardware for efficiency reasons, measure: the A10G hit the lowest cost per million tokens at 1,000 concurrent requests while using more energy per token than the L40S configuration, so your choice depends entirely on which of the two you are actually optimising. Third, run the Pareto check on your own serving configuration, because the base rate from the 86-checkpoint study says you are probably spending 2.6 to 4.8x more energy than a configuration that scores the same.

This connects to something I wrote in September about agentic reasoning burning up to 136.5x more GPU energy per query than a single-turn call. The pattern is consistent across independent research groups: the more control you hand to the model over how much computation to spend, the less a per-prompt average means. A single-turn chat completion has a bounded cost. An agent loop does not, and neither does a configuration you have never benchmarked against its alternatives.

One honest uncertainty. The three November 2026 studies measured open-weight models on self-managed hardware. Google and OpenAI's figures come from proprietary serving stacks with optimisations nobody outside has measured. It is possible the hyperscalers genuinely sit near the Pareto frontier and the 2.6 to 4.8x median excess applies mainly to self-hosted deployments. I cannot rule that out from the published data, and I would not assume the opposite either. What I can say is that if you self-host, the research says the headroom is probably in your stack, and the way to find out costs a few days of instrumented benchmarking rather than a quarter of debate.

WHAT TO DO WITH THIS

Stop sourcing your AI energy figure from a vendor press release and start measuring energy per million output tokens on your own traffic, with reasoning mode and concurrency recorded. If you want a second pair of eyes on where your inference stack sits relative to that frontier, I am happy to talk it through. Book a call →



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →