Digging Deeper – Energy and AI Studies

We generated these estimates using several large language models (LLMs). Also note: To account for the energy used in training the models, which is often excluded from standard calculations, we doubled the resulting estimates.

Even so, the academic literature suggests a wider range of possible electricity use than the LLM-generated estimates indicate. This may point to a potential bias toward lower estimates. For example, one study from 3 years ago by Alex de Vries published in *Joule* suggests electricity consumption could be roughly ten times higher than our estimates:

https://www.cell.com/joule/fulltext/S2542-4351(23)00365-3

Below is a response from one of the queries we pursued. The results generally support our rough estimates, particularly of our assumption using **2 kWh per month** for our kind of studies as a baseline. But there remains considerable uncertainty. If you have relevant expertise, data, or feedback that could help improve these estimates, we would welcome your input. Please contact us at **info@footprintnetwork.org**.

QUESTION:

When querying Gemini or ChatGPT, it tells me that **Energy Use Per Query is as follows:**

- **Google Gemini:** Uses about **0.24 Wh** per median text prompt.
- **ChatGPT / Claude:** Use about **0.3 to 0.34 Wh** per standard text prompt.

Yet when looking at this academic publication https://www.cell.com/joule/fulltext/S2542-4351(23)00365-3 it concludes that it would take 10-20 times more energy use per query than the estimates above. Can you reconcile the two? What might be the range? What is the monthly use of a basic subscriber? Might it even possible that the subscription costs do not even cover the energy costs?

Yes. They can be reconciled, and there is now a particularly useful development: **a 2026 paper in *Joule* revisits essentially this exact question and concludes that the older estimates were too high for ordinary, production-scale queries.**

The short answer

For a **normal text question and answer today**, I would use roughly:

0.3–0.7 Wh per query

rather than either treating 0.24 Wh as universally correct or continuing to use 2.9 Wh.

For more computationally intensive uses, the range becomes much wider:

Type of interaction	Reasonable current range
Short/ordinary text query	**~0.2–0.7 Wh**
Longer ordinary text query	**~0.5–2 Wh**
Reasoning/"thinking" query	**~2–7 Wh**
Very long reasoning/agentic task	**10–40+ Wh**
Old 2023 ChatGPT estimate	**2.9 Wh**

The new *Joule* paper estimates **0.31 Wh median** for standard frontier-model queries, with an interquartile range of **0.16–0.60 Wh**. For reasoning queries producing around 5,000 output tokens, the median rises to **3.91 Wh**, with an IQR of **2.15–7.05 Wh**.

So the answer has changed substantially since 2023.

Why the 2023 paper was so much higher

The paper you linked is Alex de Vries' 2023 *Joule* commentary, *The growing energy footprint of artificial intelligence*. Importantly, **it did not measure the electricity consumed by an individual ChatGPT query**.

Instead, it relied on a SemiAnalysis estimate that early ChatGPT required 3,617 Nvidia A100 servers, consuming **564 MWh/day**, and divided that by an estimate of **195 million requests/day**. That produced the now-famous figure of **"at most 2.9 Wh per request."**

There were several assumptions embedded in that calculation. Among them, the underlying workload assumed roughly **4,000 input tokens and 2,000 output tokens per request**—much larger than the ordinary ChatGPT interaction. It also assumed older **A100 hardware**, a large dense model, and did not reflect the highly optimized batching and inference systems now used by commercial AI services. Epoch AI went back through the calculation and concluded these differences explain most of the roughly tenfold discrepancy.

In other words:

2023: “Given the hardware and assumptions we have, how much power might ChatGPT's infrastructure be consuming?”

versus

2025–26: “How much energy does an actual or representative production query consume on modern infrastructure?”

Those are quite different questions.

There has also been enormous efficiency improvement. Google reports that between May 2024 and May 2025, electricity use for its median Gemini prompt fell **33-fold**. Its May 2025 measurement was **0.24 Wh**, and unlike many earlier estimates it included the accelerator, CPU/RAM, idle capacity and data-center overhead.

And Google's number is no longer an outlier. The current evidence lines up surprisingly well:

- Google production measurement: **0.24 Wh**
- Epoch AI GPT-4o estimate: **~0.30 Wh**
- OpenAI/Sam Altman disclosure: **~0.34 Wh average**
- 2026 *Joule* modeling: **0.31 Wh median, 0.16–0.60 IQR**

That makes me reasonably comfortable saying **~0.3 Wh is the right order of magnitude for an ordinary contemporary text query**.

One caveat: Google's 0.24 Wh is a **median**, rather than a mean. Long queries create a large right-hand tail, so 0.24 Wh shouldn't simply be multiplied by every query in the world. The 2026 *Joule* study explicitly warns that averages can conceal this workload distribution.

Reasoning models change the picture

This is probably the most important qualification for the way people increasingly use ChatGPT.

A simple question such as:

“What is the capital of Senegal?”

might genuinely be in the few-tenths-of-a-Wh range.

But something like:

“Read these three 40-page documents, research the issue, compare the organizations and write me an analysis.”

is not remotely equivalent to one ordinary “query.”

Epoch estimates that a 10,000-token input could push a GPT-4o interaction to about **2.5 Wh**, while processing roughly 100,000 input tokens could approach **40 Wh** under its assumptions.

And the new *Joule* paper finds that a reasoning response around 15 times longer than a normal response raises energy consumption about **13-fold**, from ~0.31 to ~3.91 Wh.

This is why I would **not** put “one AI query = 0.3 Wh” into a presentation without qualification.

I'd phrase it more like:

A typical short text interaction with a modern AI model uses roughly 0.3–0.7 Wh. Complex reasoning, long-document and agentic tasks can use several Wh to tens of Wh.

That is much harder to misinterpret.

What does that mean for one subscriber per month?

There doesn't appear to be public data giving a meaningful “average Plus subscriber's queries per month,” and usage patterns are hugely different. But we can create some useful scenarios.

For **ordinary text use**, using **0.3–0.7 Wh/query**:

Usage	Queries/month	Electricity/month
Light: 10/day	300	**0.09–0.21 kWh**
Regular: 30/day	900	**0.27–0.63 kWh**
Heavy: 100/day	3,000	**0.9–2.1 kWh**

That is remarkably small at the **individual** level.

Even using an illustrative electricity cost of **$0.10/kWh**, 3,000 ordinary queries would cost only about:

$0.09–$0.21 of electricity per month.

Even suppose somebody somehow did **100 reasoning queries every day**, at ~4 Wh each:

3,000 × 4 Wh = **12 kWh/month**

or roughly **$1.20 of electricity at $0.10/kWh**.

So your last question has an interesting answer.

Could a $20 subscription fail to cover the electricity?

For electricity alone: almost certainly not for normal usage.

ChatGPT Plus remains $20/month. At $0.10/kWh, $20 buys **200 kWh** of electricity.

At 0.34 Wh/query, consuming 200 kWh would require roughly **588,000 ordinary queries in one month**—around **20,000 per day**.

Even using the old 2.9 Wh figure, you'd need around **69,000 queries/month**, or more than **2,000 per day**, just to consume $20 of electricity.

But there is an important distinction:

Subscription revenue absolutely can fail to cover the total cost of providing the AI service.

Electricity is only one part of the cost. There are extremely expensive GPUs, depreciation of servers and data centers, networking, memory, storage, redundancy, engineering, model training, and so forth. Reasoning models and products such as Deep Research also consume much more compute than an ordinary chat response.

Indeed, in January 2025 Sam Altman said OpenAI was actually losing money on its **$200/month Pro subscriptions** because subscribers were using the service more heavily than expected. That was about **total compute/service cost**, not their electricity bills.

So I would separate these two ideas:

Individual energy footprint: surprisingly small for normal text interactions.

System-level AI electricity demand: potentially very large, because billions of interactions plus training, reasoning, agents and rapidly expanding usage multiply those small numbers enormously.

That second point is important. The 2026 *Joule* paper estimates that even if only **10% of queries become long reasoning queries**, aggregate inference electricity demand can more than double.

The conclusion I would use

I think the 2023 paper should **no longer be used as the best estimate of the energy required for an ordinary ChatGPT query**. Interestingly, *Joule* itself has now published research saying essentially that: current production-scale estimates are **4–20 times below widely cited older estimates**.

But I also wouldn't use **0.24 Wh as "the energy use of AI."** A better representation is:

Ordinary text AI: ~0.3–0.7 Wh/query
Reasoning/complex AI: ~2–7 Wh/query
Very long-context/agentic work: potentially 10–40+ Wh/task

QUESTION:

Does the estimate of how many seconds the system worked on a problem, like in this case 1 min 45 seconds, give an indication of energy use? In other words, is there an equivalent between seconds of reasoning and energy use?

Yes—but **only as a weak proxy**. There is no reliable equivalence such as "one second of ChatGPT thinking = X watt-hours."

The reason is that the number you see—such as **1 minute 45 seconds**—is **elapsed wall-clock time**, whereas electricity consumption depends on how much accelerator compute was actually devoted to your request.

A useful way to think about it is:

Energy ≈ computational work × energy per unit of computation

rather than:

Energy ≠ elapsed seconds × fixed wattage

Why 105 seconds doesn't translate directly into energy

Several things can be happening during those 105 seconds:

- The model may be generating **hidden reasoning tokens**, which do require computation.
- It may be processing a long conversation or document before answering.
- Multiple users' requests may be **batched together on the same GPUs**.
- The system may be waiting for a web search or another tool. In the previous answer, for example, I did web searches; some of that elapsed time was retrieval rather than the language model continuously "thinking."
- Different models can generate reasoning tokens at different speeds and use very different amounts of hardware per token.
- There can also be networking, scheduling and other latency that adds seconds without proportionally adding GPU electricity.

OpenAI itself cautions that its displayed "thinking time" isn't directly comparable across different models; models are independently tuned for different amounts of thinking.

There is nevertheless a relationship

Within **the same model, on similar tasks**, longer reasoning generally means more hidden tokens and therefore more energy.

The 2026 *Joule* study gives us perhaps the best illustration:

- ordinary frontier-model query: about **0.31 Wh**
- reasoning workload producing about **5,000 output tokens:** about **3.91 Wh**
- thus, roughly **13× as much electricity**

The important variable in that study is **tokens/computation**, not seconds.

OpenAI likewise describes reasoning tokens as additional internal tokens generated before the visible answer. So a response can look quite short while having required thousands of hidden reasoning tokens.

Here's a striking example of why seconds can mislead

Epoch AI estimated that a typical 500-token GPT-4o response requires roughly **one second of H100-equivalent GPU time**, corresponding to around **0.3 Wh**.

But the person using ChatGPT obviously waits considerably longer than one second for the complete response. That's because GPUs process many queries simultaneously in batches—the 15 or 30 seconds you experience are not 15 or 30 seconds of an entire GPU being dedicated to you.

So it would be badly wrong to say:

H100 ≈ 700 watts × 105 seconds = 20 Wh

because you did **not** have an H100 to yourself for 105 seconds.

What about your 1:45 example?

I would interpret that as evidence that the interaction was **probably more computationally intensive than an ordinary 0.3 Wh text prompt**, but I would not be comfortable deriving an exact figure from the timer.

For illustration, if a 105-second interaction represented something roughly comparable to the *Joule* study's ~5,000-token reasoning task, then **~4 Wh** would be a reasonable order-of-magnitude estimate. But it could readily be lower or higher because we don't know how many reasoning tokens were generated, what model/hardware served them, or how much of those 105 seconds was spent waiting on searches and other operations.

I would therefore use something like this framework:

What you observe	What it tells you about energy
5-second simple answer	Probably low energy
30-second reasoning answer	Probably more
1–2 minutes of active reasoning	Likely substantially more
5 minutes involving searches/tools	**Cannot infer energy from time alone**
Very long uploaded document	Could be high energy even if response arrives quickly

What I wish ChatGPT showed instead

If you're trying to track your own AI electricity footprint, **"thinking seconds" isn't actually the metric we'd want**. A much better disclosure would be something like:

Input: 2,300 tokens
Reasoning: 4,700 tokens
Output: 650 tokens
Estimated electricity: 3.6 Wh

The API already tracks reasoning tokens separately for some models, even though ChatGPT does not expose those counts to you in the normal interface.

And that suggests something potentially useful for your larger question about AI's environmental footprint: **the really important distinction may not be "number of queries" at all.** A person doing 100 simple queries could conceivably use less electricity than someone doing five intensive reasoning/research queries.

So I would increasingly think in terms of **compute-intensive tasks rather than queries**. The 0.3 Wh/query number is useful for ordinary chat, but it becomes misleading as ChatGPT shifts toward reasoning, research, documents, and agents.