AI vs. an extra minute in the shower

The sustainability of large language models is no new topic. I think most of us have heard of how immensely energy demanding the training process of these models is; Training GPT-3 consumed ~1.3 TWh [1], producing greenhouse gas emissions of around half a thousand metric tonnes of CO_2 [2]. However, we know that they are also quite energy demanding during their inference phase. I'm not too fond of the current "AI" hype, but I recognize its value as a tool, so I'd like to explore what (if any) environmental concerns I should have with my personal use.

For me, and I think most people, the most common way to interact with LLMs is via a chat interface, like e.g. ChatGPT. At the time of writing, the default free model in ChatGPT is `gpt-4o`, so I will assume that that is the most typical model in use today. There are few public details available on it's architecture, and little open research done on it, but it achieves somewhat similar results as Llama 3. (`llama3-70b` gets a score of 82.0 [3] on the MMLU benchmark, while `gpt-4o` gets 88.7 [4].) I don't think there's reason to believe their energy usage is that different. A quite recent publication by Husom[1] et al. [5] shows that a server-grade machine running `llama3-70b` consumes an average of 2.26 Wh per response. This aligns well with a an estimate from 2023 [6], putting the energy use of the ChatGPT service to *at most* 2.9 Wh per response (though these numbers are notably from when ChatGPT used GPT-3). The responses in the study by Husom et al. had an average length of 251.46 tokens [5], which translates to roughly 9 mWh of energy consumed per token[2]. Another study by Samsi et al. [7] shows around 7 J (~2 mWh) per token, but that looks at an older version of Llama. As you understand, the literature is limited and there are a lot of unknowns here, for example the fact that hardware varies a lot and that a larger data center can operate a lot more efficiently than a single machine. However, let's use the rough range of 2-9 mWh per token as a starting point.

Let's proceed, then, with two types of users:

- **A conservative user**: Uses a model that has an energy use of 2 mWh per token and that leans towards 200 tokens on average per response. The user performs 10 queries per day.
- **A heavy user**: Uses a model that has an energy use of 9 mWh per token and that has longer responses of on average 1000 tokens. The user performs 500 queries per day.

With the numbers found above, the conservative user would have an energy footprint of 4 Wh per day, from their use of LLMs. The heavy user, on the other hand, will have a footprint of 4.5 kWh per day. 4 Wh is less than an efficient LED bulb will use in an hour, while 4.5 kWh is about the amount of energy my panel heater uses to keep my bedroom at 22 °C on a typical winter day. (I live in Norway.) The average data center uses 1.7 liters of water per kWh of energy consumed [2], which means the conservative user spends an extra 7 mL of water a day on their LLM use, while the heavy user spends 7.6 L — about the minutely water consumption of an efficient shower. Neither of these estimations are huge, but I am surprised at how the heavy user's consumption is on the level of something I would go out of my way to save by other means (for example showering less, being smart about heating, etc.). Remember, these are *very rough* estimations, and there are probably a lot of factors I'm missing here, but I don't consider my numbers as too unreasonable.

Things change and technology evolves; the R1 model by Chinese DeepSeek made a lot of splash lately, in part namely because of its energy efficiency (though their claims have been put into question). Additionally, the use of language models in day-to-day life for an individual admittedly has a small environmental impact compared to other activities. However, I think it is important to compare apples to apples, and highlight the fact that many of the use-cases presented by LLMs replace other, vastly more energy-efficient tools that largely achieve the same job. Asking ChatGPT how a transformer works probably doesn't consume a lot of energy, but searching that up on Wikipedia consumes a lot less. (I would also claim that you will learn more from the latter.) Not only that, but some AI products attempt to replace the work of us humans, and while we are not as energy efficient as a data center, one can question whether that is the right place to conserve energy. In my case, I'll continue to use ChatGPT when it helps me solve an issue I would struggle to solve otherwise, but for the most part I'll keep my use of LLMs to the minimum. I'd rather just use that extra minute in the shower to think the problem through.

References

[1] D. Patterson *et al.*, "Carbon Emissions and Large Neural Network Training," 2021, *arXiv*. doi: 10.48550/ARXIV.2104.10350.

[2] C. C. Walther, "The Hidden Cost of AI Energy Consumption," Knowledge at Wharton. Accessed: Jan. 27, 2025. [Online]. Available: https://knowledge.wharton.upenn.edu/article/the-hidden-cost-of-ai-energy-consumption/

[3] AI@Meta, "Llama 3 model card," 2024, Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[4] Wikipedia contributors, “GPT-4o — Wikipedia, the free encyclopedia.” 2025. Available: https://en.wikipedia.org/w/index.php?title=GPT-4o&oldid=1275995932

[5] E. J. Husom, A. Goknil, L. K. Shar, and S. Sen, “The Price of Prompting: Profiling Energy Use in Large Language Models Inference,” Jul. 24, 2024, *arXiv*. doi: 10.48550/ARXIV.2407.16893.

[6] A. De Vries, “The growing energy footprint of artificial intelligence,” *Joule*, vol. 7, no. 10, pp. 2191–2194, 2023, doi: 10.1016/j.joule.2023.09.004.

[7] S. Samsi *et al.*, “From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference,” Oct. 04, 2023, *arXiv*: arXiv:2310.03003. doi: 10.48550/arXiv.2310.03003.

[1] Thanks to Erik Johannes Husom for helping me with the research for this blog post. He researches green AI at SINTEF and there’s a lot of interesting posts on his blog about the topic. ↩

[2] The energy is not split equally over the input and output tokens. To simplify my calculations, I’ve chosen to abstract that away and base them only on the number of output tokens. ↩

home · src · feed · blogroll · fediverse

Here are some posts from sites I follow

Speech-to-Phrase brings voice home - Voice chapter 9

Welcome to Voice chapter 9 🎉 part of our long-running series following the development of open voice. We’re still pumped from the launch of the Home Assistant Voice Preview Edition at the end of December. It sold out 23 minutes into our announcement - wow…

via Home Assistant February 13, 2025

2024 State of Rust Survey Results

Hello, Rustaceans! The Rust Survey Team is excited to share the results of our 2024 survey on the Rust Programming language, conducted between December 5, 2024 and December 23, 2024. As in previous years, the 2024 State of Rust Survey was focused on gatheri…

via Rust Blog February 13, 2025

[NOR] KI og klimaavtrykk på NRKs Abels Tårn

Forrige uke var jeg så heldig å få fortelle om vår forskning innen mer bærekraftig kunstig intelligens på en av Norges største populærvitenskapelige radioprogram, nemlig Abels Tårn på NRK! Først og fremst var det veldig gøy å få være med, og det var en flo…

via Erik Johannes Husom February 10, 2025