LLM Water And Energy Use

Milja Moss - July 2024

Sources used:

- [1] https://arxiv.org/pdf/2304.03271
- [2] https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/
- [3] https://www.nrel.gov/docs/fy04osti/33905.pdf
- [4] https://arxiv.org/pdf/2005.14165

Here, I'm putting together various sources to properly contextualize how much energy and water you actually spend when using services like ChatGPT. This is not a rigorous analysis; while I've based my numbers on the above sources as much as I could, I've also introduced wiggle room by having to make some educated guesses.

This analysis addresses the claims of "AI uses a bottle of water every time you prompt it."

Server cooling evaporates around 1L / kWh when using cooling towers

> On average, depending on the weather conditions and operational settings, datacenters can evaporate about 1 – 9 liters per kWh of server energy (about 1 L/kWh for Google's annualized global number and 9 L/kWh for a large commercial data center during the summer in Arizona). [1]

On the other hand, Microsoft claims that their US datacenters use only around 0.5 liters of water per kWh for cooling on average [2]. Different corporations report different numbers for their datacenters, so this is a bit tricky to put a definite number on. I've chosen to double the reported 0.5 liter figure given by Microsoft for my calculations here because OpenAI's datacenters were funded by Microsoft, and I'm leaving some generous margin for error.

It should be understood that when datacenters take in water, they don't spit it out as toxic radioactive sludge. Cooling towers are, in the most simple terms, giant buckets of water. Heat is pumped into these buckets from inside the datacenter, transfering the heat to the water. As the hot water evaporates, it releases the heat into the environment. New water is then added in to continually replace the lost water. The evaporated water enters the water cycle, and rains back down somewhere else.

What the paper is concerned about is disruption to local ecosystems and water availability. If datacenters in one location grow too big, they risk evaporating too much water, drying up the local freshwater sources. The same goes for all manner of datacenters and industrial water use regardless of application, of course.

Sidenote

Cooling towers have other environmental impacts outside the scope of this analysis. Namely, the water in cooling towers is treated to prevent microbial

and fungal growth, as well as prevent corrosion and other unwanted damage to the cooling infrastructure. If this treated water is dumped back into the environment without proper recycling, these chemicals will cause environmental harm.

7.6L of water is used per 1 kWh of energy generated.

This is another important point to realize. The "bottle of water for each response" claims originate from this line:

> GPT-3 needs to “drink” (i.e., consume) a 500ml bottle of water for roughly 10-50 responses, depending on when and where it is deployed [1]

And this figure considers both "scope 1" as well as "scope 2" water usage. Scope 1 means the water used for cooling directly, as described above. Scope 2 means water used for generating the energy itself. That means water used by the power plants that supply the datacenter is also taken into account.

> The national weighted average for thermoelectric and hydroelectric water use is 2.0 gal (7.6 L) of evaporated water per kWh of electricity consumed at the point of end use [3]

This means that ALL electricity usage, across the United States, spends 7.6 liters of water per one kilowatt-hour.

To put this into perspective, these activities spend more or less 1 kilowatt-hour of electricity:

- Playing a PC game for 2 hours
- Using an oven for 30 minutes
- Microwaving food for 1.5 hours
- Running 1 washing machine cycle
- Boiling 12 kettles of tea
- Watching Netflix on your TV for one hour
- Watching TV for 9 hours (not streaming, playing a bluray or similar)
- Having your fridge on for one day
- Vacuuming for an hour
- Running a space heater for 30 minutes

Each time you do one of these things, you use 7.6 liters of water (on average across the entire US).

Inference energy cost

> The official estimate shows that GPT-3 consumes on the order of 0.4kWh electricity to generate 100 pages of content (e.g., 0.004kWh per page) [1]

This refers to the "Language Models Are Few-Shot Learners" paper:

> with the full GPT-3 175B, generating 100 pages of content from a trained model can cost on the order of 0.4 kW-hr [4]

Note two things here:

1. The model being talked about is the 175 billion parameter base model GPT-3, NOT GPT-3.5-Turbo or GPT-4-Turbo, which are of unknown parameter counts.

2. The estimate comes from "pages of content" generated with GPT-3, not from individual responses from ChatGPT.

It is not clear how many tokens a "page" is worth, the paper does not define what a page is. I'm going to assume they mean a standard page in a book, which is around 500 tokens. This corresponds to 1-3 responses from ChatGPT, depending on context. The following results are somewhat fuzzy given there are no concrete numbers available.

Based on this I'll estimate that, given a 175B parameter chat model, one 300-token response will consume around 2.5 watt-hours. The current distilled "turbo" models for both ChatGPT 3.5 and 4 are almost certainly smaller than this, at least for 3.5. But we'll stick with the 2.5 watt-hour figure per response for now.

With our 1 liter per kWh estimated for cooling, 2.5 watt-hours neatly corresponds to 2.5 milliliters of cooling water evaporated per request. If we take into account the water spent on electricity as well, the number jumps to 21.5 milliliters of water per request. This figure is in line with the paper's "GPT-3 needs to 'drink' a 500ml bottle of water for roughly 10-50 responses" [1], allowing us 24 responses for 500 ml of water use.

As I said, the current production models (at least for GPT-3.5-Turbo) are smaller than 175 billion parameters. A wild guess for 3.5-turbo, given its performance compared to open source models, is in the ballpark of 30-50 billion parameters. So odds are that ChatGPT-3.5 gets you around 100 responses per 500ml of water.

In summary:

- GPT-3-175B: You get ~400 responses per kilowatt-hour (and 8.6 liters of water used, if counting electricity generation).
- GPT-3.5-Turbo: You get ~1600 [citation needed] responses per kilowatt-hour.

When determining whether you should be using services like ChatGPT, the question then becomes: "Are 1600 responses from ChatGPT as useful to me as 2 hours of video games?"

(For posing the same question about GPT-4o, you should use the numbers for GPT-3-175B as a guideline. There is no confirmed information for the size of GPT-4o; rumors say it is a mix of 200 billion parameter models, where the most fitting model is chosen on a per-task basis. This would make it more or less equivalent to the original GPT-3 model in terms of power use.)