Addressing the knowledge gap of environmental costs by using GPT with an informative browser extension

Linthe van Rooij
s3989429
Leiden University
s3989429@vuw.leidenuniv.nl

December 18, 2024

The rise of ChatGPT and its environmental costs

When ChatGPT was released in November 2022, Large Language Models (LLMs) gained much popularity. This interface was one of the first to be easily accessible to a wide public [1]. Individuals experienced its direct potential by interacting with this model on a website. In the first week after its release already 1 million users were using ChatGPT. And this appeared only to be the beginning. One year later, ChatGPT gained up to 100 million users. But this fast-rising and growing use comes with its consequences. While users benefit daily from the fast developments of AI, huge data centers are continuously running behind the scenes to meet the demands of the models [2].

Running these data centers is energy and water-intensive [2, 3, 4]. Both training and inference (using the model) use varying amounts of environmental resources and it depends on multiple factors which, and how much of these resources are used.

Firstly, energy consumption depends greatly on the way the model was designed. More complex models (more parameters, more layers in a deep neural network) require more computational resources and thus more energy [3]. With a growing demand on the user's side, tech companies are incentivized to produce better outputs with more accurate models, which often comes down to more complex models, requiring more energy. Before inference, months of training are needed to produce these accurate outputs [5]. Training a LLM, in particular, is energy-intensive, due to the vast amounts of training data needed to fuel the models. Furthermore, each query with inference builds up quickly to a significant energy use if the amount of users is large. Both the training and inference contribute to the total energy consumption, depending on the complexity of the model in question.

Moreover, a substantial amount of energy and water is also used for cooling down data centers [2, 5]. Data centers produce a lot of heat and it is crucial to keep an optimal temperature for these centers to work properly. Different cooling techniques, such as air conditioning, free air, liquid, or water require varying amounts of energy and water resources [2]. Using a lot of freshwater resources to cool down data centers is problematic, especially in areas where water is scarce. The amount of cooling needed is also highly dependent on the location of the data centers. In general, hotter regions require more cooling thus more energy or water.

Although tech companies are incorporating ways to handle water and energy resources more sustainably [6, 2], the demand and pace at which LLMs are developing is simply too large to supply with renewable energy sources [7]. And yet,

insight into true (unsustainable) environmental costs is often not publicly available [3]. Despite the lack of insight, there remains a growing awareness of the costs that come with AI. Thus, it is important to not only acknowledge the environmental costs but also provide more transparency about the resources used for the whole process.

Knowledge gap of exact environmental costs by using generative AI

With the increasing public availability of Generative AI systems, especially ChatGPT, as highlighted in the press, awareness is growing about their planetary costs [7, 8] and their broader contribution to global climate change [9]. In the following section, I will further elaborate on this critique and describe the sociotechnical system that is involved in this matter.

Electricity and water

The excessive electricity usage of AI [10] contributes to the overall electricity demand which negative impact these emissions have on climate change [11]. With the burning of fossil fuels to fulfill the energy demand [12], there is an increase in carbon emissions, causing rising global temperatures and imbalances in nature. This in place, affects species, such as fish in the ocean, and planetary health. Another concern of the energy demand is that 80% of this demand depends on fossil fuels. Besides their negative effect on the environment is that these resources are finite. With the current usage levels and expanding growth, this is becoming another critical point.

On the other hand, there are growing concerns about GPT's water footprint [13, 14]. Both direct water consumption for cooling data centers and indirect water consumption by producing hardware contribute to the total water footprint of GPT. The consequences of this water usage are water scarcity, its impact on local ecosystems, water quality, and energy use. Large amounts of water are extracted from rivers in regions with data centers, directly affecting the local population's availability of water resources for their own purposes such as drinking water and agriculture. The exact water consumption varies per location and model. For example, the consumption of 500 mL of water for GPT-3 allows a data center placed in Sweden to produce around 20 inferences compared to 50 in Georgia and 70 in Ireland [13].

Responsibility and transparency

Accompanying this awareness and concerns about the impact on the climate is the growing responsibility of the manufacturers to inform their users about the specifics of their AI services. Despite this responsibility, the exact environmental impact of GPT remains difficult to track down [15, 7, 16]. The relevant stakeholders, including OpenAI and its competitors, have an interest in maintaining the status quo. They resist the call for transparency and stricter environmental regulations [17]. This hold-back of information can be due to the growing AI hype and the competition that derives between several companies [3]. Another reason can be a way to only keep the public in the "blue magic divine", instead of showing the real impact it has on our planet. Researchers in ethical and sustainable AI call for a good provision of information, to leave the decision with the users [3, 18]. However, without this provision of information about the environmental costs of AI, users cannot make a fair trade-off and often use these services without knowing the environmental consequences.

It is especially problematic for the environment when the demand for ChatGPT continues to grow at this pace. Developers tend to prioritize performance and innovation, often viewing environmental considerations as secondary to achieving breakthroughs in AI capabilities. Moreover, governments also play a role in resisting environmental costs insight, since this would most likely result in defining more strict regulations and sustainability standards, which could greatly reduce AI innovation and endanger the economic growth of a country and its success compared to other countries in this matter.

With this growing trend, GPT tends to replace other services that serve the same goal but might have a more sustainable background. For example, when comparing ChatGPT with a Google search, ChatGPT uses 10 times more energy for a single query than a plain Google search [10]. Without knowing the ChatGPT query uses much more energy than a Google search would do, why change your habits?

Some restrictions already exist on the use of more complex models to reduce the enormous demand on the energy net. While the less complicated models are often free of charge and come without rate limits [1], the recent, more complex models come with limited access for free accounts, providing some regulations on their unlimited usage. OpenAI also offers paid accounts, providing some monetary incentive to pay for the more complex models which are more likely to use more resources overall. While the free accounts may sound beneficial for users who don't want to spend money on these services, there is also a downside. Free accounts with a less complicated model can be used by users with no rate and time limit, or monetary expenses. With the free availability 24 hours a day. This unlimited access is also restricting AI services from implementing their services more sustainably with renewable energy sources [19]. Renewable energy sources are often fixed to specific conditions (sun, wind) making it impossible for these sources to deliver enough consistent and high-quality power to fulfill ChatGPT's total demand.

Action required

Taken together, the growing demand for powerful AI systems such as GPT—from the public, manufacturers, developers, and even nations—has significant environmental consequences. This demand increases reliance on energy sources, often fossil fuels, and worsens water scarcity in vulnerable regions. Alongside the rapid developments, there is a growing awareness of these environmental impacts, combined with a pressing need for more transparency from big tech companies such as OpenAI. To address the knowledge gap about the environmental costs of using ChatGPT, I propose a browser extension that provides users insight into the water and electricity costs of a GPT model on the current website they are visiting. To goal of this intervention is to make users aware and ultimately to reduce the environmental impact on the user's side.

GreenPlanetTracker

Functionalities

A browser extension is easy to access and downloadable for every interested user. The download option provides basic information about the extension and the permissions granted when downloading this (such as accessing your data). After downloading, the extension will be available in your browser toolbar as shown in Fig. 1. The user can adjust the extension's available active and inactive states with one click, making it easy to turn the extension offline if not wanted at a specific moment. When activated, the extension will search for traces of using a GPT model on the current website. Suppose any model is detected and the version of the model

[1] `https://chatgpt.com/#pricing`

is also known. In that case, it will be displayed in the extension as seen in Fig. 2 - 4. The extension comes with 3 tabs, each providing meaningful information about the relation of using GPT and the consequences for the environment.

Figure 1: Availability of extension after downloading in the toolbar.

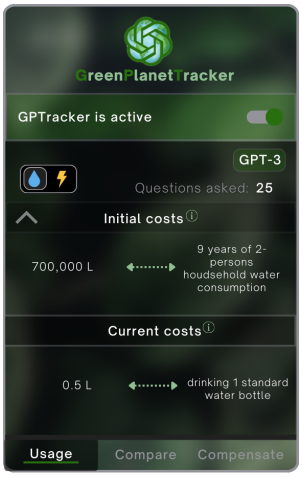

Figure 2: Your current water and electricity usage tab

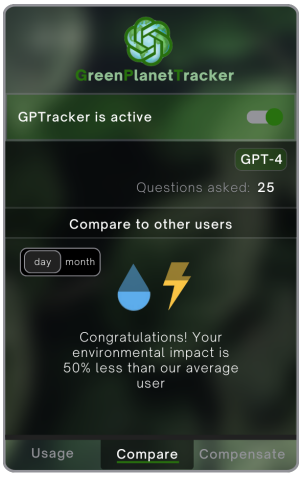

Figure 3: Compare usage patterns with other users tab.

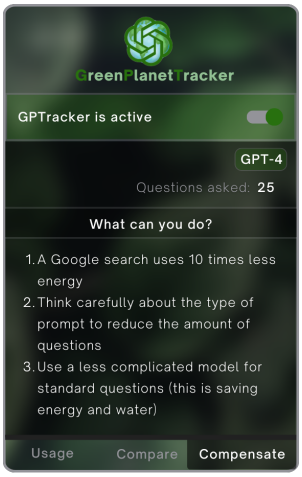

Figure 4: Compensate your usage with concrete actions tab.

The first tab provides insight into the user's cost patterns when using the model on the website (see Fig. 2). This will start tracking the number of questions/ prompts processed to show your real-time current costs (inference costs). Next to the inference costs, the initial costs (training costs) are also shown, as they vary in the type of model used and are insightful to be able to use the model in the first place. The user can choose between water and electricity costs to be displayed. The water and electricity costs are also translated into comprehensible data, such as water usage in a household, or electricity usage in charging your phone x times. This is done to make this data easily interpretable for any user. The combination of training and inference costs will give real-time information about the user's current total usage patterns.

In the second tab, users can compare their usage patterns to other users (see Fig. 3), using the same tracked user data to translate into water and electricity costs. Additionally, the data of other users is also shown, to provide more insight into how your usage patterns relate to other users who used this extension. Users can view this data over a shorter period (a day), or a longer period (a month), enabling them to compare their usage with others and balance out their peak consumption. Next to the symbolic representation of these comparable costs, it will provide textually how much more or less impact you have on the environment compared to other users, making it relevant and incentivizing to use less than others.

Lastly, the compensate tab provides varying small tips to compensate for your current costs. This ranges from using these models more efficiently to other services that can be used instead of GPT models. This enables users to not only get more insight into usage patterns but also to know concrete actions that can be done to minimize their impact, without completely abandoning GPT models.

Impact

Suppose you now want to judge a system on its overall usage based on inference and training costs in water usage, and electricity [7] it is highly complicated to consider all the parameters and make an informed decision as to whether or not use an AI model, let alone to know what your contribution might be to the impact on our planet. Visibility and insight into the usage of AI models create more awareness around the environmental costs and are a step in the right direction to reduce its unnecessary usage and thus impact significantly. The goal of this visibility of costs is not to limit the potential of AI but to find ways in which this can be done in a hopeful and sustainable manner. This browser extension will leave the responsibility of using these models to the users, letting users decide how and when they want to use it.

With this extension, the existing knowledge gap of exact environmental costs by using GPT models is addressed, providing meaningful information about usage patterns and guidance on how using GPT models can be achieved more sustainably. This extension is creating more awareness by providing a good overview of information about the environmental costs of AI. Moreover, it enables users to make a trade-off about whether the costs that accompany using generative AI are worth the use. Altogether, this intervention enables users to reduce their environmental impact.

Reflection

Speculative interventions, like this project, have been very eye-opening in terms of rethinking what AI is and what it could be. It allowed me to step back and imagine alternative paths for technology, focusing on values like sustainability and ethics instead of just efficiency or profit. I have put a lot of work into making the project as applicable as possible. Despite it being speculative, this helped me create something that does make an impact.

Before starting this project, I felt quite helpless because of the negative side effects of AI that we discussed in class. I asked myself whether I wanted to even use AI in the future, despite how much I use it every day. This project allowed me to rethink this mindset. To not only think and be frustrated about the negative consequences but to also imagine what could be done to guide this process in a different, more sustainable way.

Working on this project made me much more aware of how important it is to align our use of AI with our values. I realized that, just like buying clothes or food, using AI carries hidden costs that we often ignore. This made me think more carefully about what kind of information or tools I would need to adjust my behavior and reduce my impact.

I wanted to align my project with permacomputing, which is an idea inspired by permaculture. This is a holistic approach aimed at integrating computing into human civilization in a sustainable and meaningful way, ensuring its alignment with the health of the planetary biosphere [20]. I found this concept truly inspiring because it showed me that not all advancements in technology have to come at a high cost to the planet. There are ways to design systems that respect the environment and still meet human needs (although this can also be viewed as a Pfaffenberger myth that all technology needs to fulfill human needs [21]). For me, projects aligning with permacomputing create a hopeful view, that next to the technological advancements at a high pace, there are movements that create alternative visions that are less impactful on the planet, to live more with nature than going against it.

Overall, this project helped me to realize that besides the negative consequences of AI, it is important to also focus on alternative ways in which these consequences could be less harmful. Speculative work like this is important because it pushes us

to think about what we truly value and how technology can fit into that vision. It remains important to not be passive users of these models but to always align our use with the overarching values that we find important.

References

[1] Fabio Duarte. *Number of ChatGPT Users (Dec 2024)*. Dec. 2024. URL: https://explodingtopics.com/blog/chatgpt-users#top.

[2] Vinod G R Kunmar. *ChatGPT Data Centers Are Consuming Staggering Amount Of Water, Study Warns — by Vinod Kumar G R — Medium*. Apr. 2023. URL: https://medium.com/@vinodkumargr/chatgpt-data-centers-are-consuming-staggering-amount-of-water-study-warns-6cc1861c19ee.

[3] James Vincent. *How much electricity do AI generators consume? - The Verge*. Feb. 2024. URL: https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption.

[4] Brian Calvert. *AI already uses as much energy as a small country. It's only the beginning. — Vox*. Mar. 2024. URL: https://www.vox.com/climate/2024/3/28/24111721/climate-ai-tech-energy-demand-rising.

[5] Arooj Ahmed. *New Research Shows ChatGPT Uses Too Much Water to Keep its Data Centers Cool Which Can Have Harmful Effects to Environment*. Sept. 2024. URL: https://www.digitalinformationworld.com/2024/09/new-research-shows-chatgpt-uses-too.html.

[6] Tate Cantrell. *The true cost of AI innovation — Scientific Computing World*. URL: https://www.scientific-computing.com/analysis-opinion/true-cost-ai-innovation.

[7] Alva Markelius et al. "The mechanisms of AI hype and its planetary and social costs". In: *AI and Ethics* (2024), pp. 1–16.

[8] Sasha Luccioni, Yacine Jernite, and Emma Strubell. "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" In: *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024* (June 2024), pp. 85–99. DOI: 10.1145/3630106.3658542. URL: http://arxiv.org/abs/2311.16863%20http://dx.doi.org/10.1145/3630106.3658542.

[9] David Patterson et al. "Carbon Emissions and Large Neural Network Training". In: (Apr. 2021). URL: https://arxiv.org/abs/2104.10350v3.

[10] Alex de Vries. "The growing energy footprint of artificial intelligence". In: *Joule* 7 (10 Oct. 2023), pp. 2191–2194. ISSN: 2542-4351. DOI: 10.1016/J.JOULE.2023.09.004.

[11] Kari C Nadeau et al. *Climate change: a call to action for the United Nations*. 2022.

[12] Jerry L. Holechek et al. "A Global Assessment: Can Renewable Energy Replace Fossil Fuels by 2050?" In: *Sustainability* 14 (8 2022), p. 4792. ISSN: 2071-1050. DOI: 10.3390/SU14084792. URL: https://www.mdpi.com/2071-1050/14/8/4792.

[13] Pengfei Li et al. "Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models". In: *arXiv preprint arXiv:2304.03271* (2023).

[14] A. Shaji George, A. S. Hovan George, and A. S. Gabrio Martin. "The environmental impact of AI: A case study of water consumption by ChatGPT". In: *Partners Universal International Innovation Journal* 1.2 (2023), pp. 97–104.

[15] UNEP. *AI has an environmental problem. Here's what the world can do about that.* 2024. URL: https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about.

[16] Earth.org. *The Green Dilemma: Can AI Fulfil Its Potential Without Harming the Environment?* 2024. URL: https://earth.org/the-green-dilemma-can-ai-fulfil-its-potential-without-harming-the-environment/.

[17] Matthew Elmore. "The Hidden Costs of ChatGPT: A Call for Greater Transparency". In: *The American Journal of Bioethics* 23 (10 2023), pp. 47–49. ISSN: 15360075. DOI: 10.1080/15265161.2023.2250335. URL: https://www.tandfonline.com/doi/abs/10.1080/15265161.2023.2250335.

[18] Chris Stokel-Walker. *TechScape: Turns out there's another problem with AI – its environmental toll.* 2023. URL: https://www.theguardian.com/technology/2023/aug/01/techscape-environment-cost-ai-artificial-intelligence.

[19] Andrew R. Chow. *How AI Is Fueling a Boom in Data Centers and Energy Demand — TIME.* June 2024. URL: https://time.com/6987773/ai-data-centers-energy-usage-climate-change/.

[20] Permacomputing Community. *Permacomputing.* https://permacomputing.net. Accessed 2024-10-19. 2024.

[21] Bryan Pfaffenberger. "Social anthropology of technology". In: *Annual review of Anthropology* (1992), pp. 491–516.