AI chatbots may be fun, but they hava a drinking problem

Aniruddha Das
Computer Science and Engineering
University of Engineering &
Management,
Kolkata, India

***Abstract*—Artificial Intelligence (AI) has revolutionized the world by providing unprecedented benefits to industries ranging from healthcare to finance. Just like organic life, artificial intelligence (AI) cannot exist without water. It uses water directly to cool its massive server rooms and indirectly at the power stations that produce electricity for those servers. The total water consumption of AI is called its 'water footprint'. Popular new AI tools like ChatGPT and BARD fall in the category of 'large language models' and have a huge water footprint . These models are trained on massive language datasets that are hosted on stacks of energy-hungry servers. Their operation produces a lot of heat. Server work best at 10-27 degrees Celsius, and to maintain this temperature range server farms employ large cooling towers. The study distinguishes between "Withdrawal" and "Consumption". Withdrawal refers to the physical extraction of water from rivers, lakes and other sources, while consumption relates to the water loss due to evaporation when used in data centers. It will also measure the 'water footprint' of large AI models like Open Ai's ChatGPT. It found that the water consumed to run ChatGPT, which is used by billions of users worldwide is "extremely large". The study highlights the importance of addressing water footprint of AI models in order to address global water crisis. Finally present some ideas to how to reduce the water footprint of AI models.**

Keywords—Artificial Intelligence(AI), water footprint

I. INTRODUCTION

The increasing popularity of AI tools such as OpenAI's ChatGPT has raised concerns about their environmental impact. Training Large AI models like GPT-3 can consume up to 700,000 liters of clean fresh water [12]. According to the down to earth report, this amount water is equivalent to producing 370 BMW cars or 320 Tesla electric vehicles.[7] The upcoming GPT-4 which is expected to be even larger, is predicted to further increase water consumption, although specific estimates are challenging due to limited data availability. While AI activities occur digitally, the physical storage and processing of data in data centers generate heat , requiring water-intensive cooling system. These systems use pure freshwater and also require significant water for power generation.

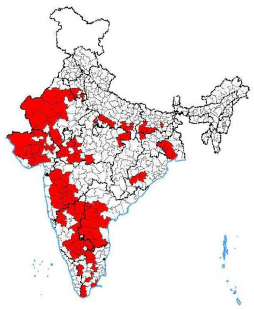

Fig.1: Map of drought prone districts of India for August , 2022.[20]

Fig. shows that 42% of the India area can be under severe drought or worse [13], where hundreds of data centers. Clearly results in a huge environmental impact on regional water system. Additionally, due to the aging public water infrastructure, the need for water conservation remains equally important, even in non drought areas. More-over, it is extremely costly to expand the aging public water infrastructure that is already operating near limits in many parts of the world .The addition of water-thirsty data center to accommodate new AI model development can certainly worsen the situation .

II. HOW COOLING TOWERS WORKS

Cooling towers works on the same principle as traditional room coolers. When water evaporates it absorbs heat from its surrounding and reduces the ambient temperature. The water vapor rises inside the cooling tower and is released into the atmosphere. As a result, the water used by data centers is lost and cannot be recycled. This is doubly problematic because cooling towers at data centers can use only clean fresh water. Say, from rivers and lakes, sea water is not an option because its high salt content would cause corrosion, damaging sensitive equipment at data centre.

. III. THE WATER CONSUMPTION IN DATA CENTERS

The water consumption in data centers has two parts – on site direct water and off-site indirect water.

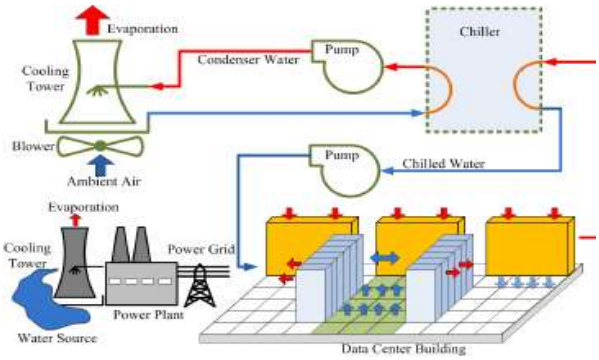

Fig 2.: Data center water footprint: on-site water consumption for data center cooling,and off-site water consumption for electricity generation.[23]

A. *On-Site direct water consumption*

There are two water loops: one closed loop between the chiller and data center server rooms, and one open loop between the cooling tower and the chiller. Within the closed loop, water is not lost-it is pumped from the chiller into the data center to cool down the air handling unit supply air in order to maintain a proper server inlet temperature, and warm water that absorbs the server heat returns to the chiller direction . Through a heat exchange at the chiller, the heat is transferred from the closed loop to the open loop.

Along the open loop, some of the water gets evaporated (consumed) in the environment. Additionally there is a process called "blown down" that drains the cooling water to reduce salt concentration.

B. *Off-Site direct water consumption*

Data centers are held accountable for carbon footprint because of their (non-renewable) electricity usage. Likewise, electricity generation requires a huge amount of water, thus resulting in off-site indirect water consumption for data center.

IV. Estimating Water Footprint of AI Models

To obtain an AI Models total water footprint, We consider both on-site WUE and off-site WUE.

A. *On-Site WUE*

Cooling towers are most commonly used as the heat rejection mechanism for data centers. In general, the on-site WUE of cooling towers depends on multiple factors, such as temperature approach setting (i,e, difference between the cold water temperature and entering wet bulb temperature), Cycles of concentrations (i,e water recirculation times before "blow down") , water flow rate, air pressure, humidity, wet-bulb temperature, and wind speed, among many others. Due to the lack of operational data from major data centers, we focus on the impact of outside wet-bulb temperature. On the on-site WUE and present an empirical model based on a commercial cooling tower[18]. Specifically, following recommended operational setting , the on-site WUE can be approximated as

$$\text{WUE } on = s/s\text{-}1(6x10^{-5} \cdot T_w^3 - 0.01 \cdot T_w^2 + 0.61 \cdot T_w - 10.40) \quad (1)$$

Where S is the cycle of concentrations and T_w is the outside wet bulb temperature (in Fahrenheit) .

B. *Off-Site WUE*

We now present the off-site indirect WUE measured in term of EWIF (Electricity Water Intensity Factor). The same way as AI models are accountable for carbon footprint associated with off-site electricity generation. Specifically, the off-site WUE depends on the energy fuel mixes (e.g., coal, nuclear, hydro) as well as cooling techniques used by power plants [10]. Since electricity produced by different energy fuel becomes non-differentiated once entering the grid, we consider the average EWIF, which can be estimated as

$$\text{WUE } off = \frac{\sum_k bk * EWIFk}{\sum_k bk} \quad (2)$$

Where b_k denotes the amount of electricity generated from fuel type k for the grid serving the data center under consideration , and $EWIF_k$ is the EWIF for fuel type k [14].

A. *Water footprint*

The on-site direct water consumption can be obtained by multiplying AI's energy consumption with the on-site WUE, while the indirect water consumption depends on the electricity usage 10 minutes to an hour depending on how frequently we want to assess the water footprint, and T is the total length of interest (e.g., training stage, total inference stage, or a combination of both). At time t, suppose that an AI model uses energy e_t . (which can be measured using power meters and/or servers' built-n tools), the on-site WUE is $WUE_{on,t}$, the off-site WUE is $WUE_{off,t,}$ and the data center hosting the AI model has a power usage effectiveness (PUE) of PUE_t that accounts for the non-IT energy such as cooling system and power distribution losses. Then, the total water footprint W of the AI model can be written as.

$$W = W_{on} + W_{off} = \sum_{t=1}^{T} et \,.\, WUEon,t + \sum_{t=1}^{T} et \,.\, PUEt \,.\, WUEoff,t \quad (3)$$

TABLE I. Estimated EWIF for Common Energy Fuel Types in the US [17]

Fuel Type	Coal	Nuclear	Natural Gas	Solar(PV)	Wind	Other	Hydro
EWIF	1.7	2.3	1.1	0	0	1.8	68

Our methodology for estimating AI models' water footprint is general and applies to data centers with any type of cooling systems. For example, if the data center uses a cooling tower other than the one we model, we only need a different $WUE_{on,t}$.

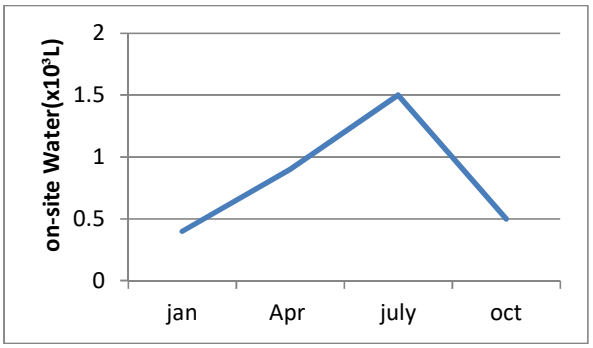

Fig 3.: On-site Water Footprint of Pune location in 2022

V. Carbon/Water Trade-off

Data centers located in countries such as Sweden and Finland use less water because of the naturally cooler condition . But in the Asis-Pacific region, where a lot of the AI action is now concentrated,higher ambient temperatures push up the need for water. There often a trade-off between carbon efficiency and water efficiency. You can generate more solar energy to run servers in the afternoon[21]so,smaller carbon footprint, but then you need more water for cooling as it is the hottest time of the day.

AI model developers may want to train their models during the noon time when solar energy is more abundant, but this is also the hottest time of the day that leads to the worst water efficiency, In other words, adopting renewable energy can sometimes come in the way of water conservation. The challenge, then, is to find a way to balance carbon and water efficiencies through new approaches to sustainable AI.

VI. What Companies are Doing

Most AI companies have pledged to make their system sustainable by 2030.One solution, could be to run AI model training in different locations at different points in time.
Microsoft says its data centers in Phoenix, Arizona, which hosted the training of GPT-3 and its advanced version ChatGPT4, saved water by using outdoor air to chill servers for most of the year. They otherwise cool through direct evaporation, which uses a fraction of the water required by other, more traditional, water-based cooling system like cooling towers.
Microsoft further plans to save a million liters of water daily by switching from conventional energy to Solar energy from the "Sun Stream 2 Solar Project."Operated by local partner Long road Energy. Google, meanwhile, uses a mix of air cooling, water cooling, refrigerants, or some combination of them, to reduce its water consumption. The decision is based on hyper local condition, and a data-driven approach to local hydrology, topography, energy ,and emissions issues,

Conclusion

In this paper, we recognize the enormous water footprint as a critical concern for socially responsible and environmentally sustainable AI, and make the first-of-its-kind efforts to uncover the secret water footprint of AI models. Specifically, we present a principled methodology to estimate the fine-grained water footprint, and show that AI models such as GPT-3 can consume a stunning amount of water in the order of millions of liters. In addition, we point out the need of increasing transparency of AI models' water footprint, and highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI.

Acknoledgement

I had a great deal of help from others among the way for an earlier [6] that eventually led to this version of article.Pengfei Li, Jianyi Yang , Mohammad A. Islam , Shaolei Ren shared information about data center water footprint.They did a great a deal of work to measure the on-site and off-site water footprnt . Arindam Giri provided helpful suggestion on how to write better technical content.

References

[1] Facebook .Sustainability report, 2020, https:// sustainability.fb.com/wp content/uploads/2021/06/2020_FB_Sustainability-Report.pdf.

[2] Amazon.Sustainability – water stewardship,2023, https:// sustainability.aboutamazon.com/environment/the-cloud/water-stewardship.

[3] Google. Water commitments,2023,https://sustainability.google / commitments/water/.

[4] UN Water Conference. How 'aquapreneurs' are innovating to solve the water crisis,2023,https://www.weforum.org/agenda/2023/03/un-water-conference-aquapreneurs-innovation.

[5] U.S.White House.White House action plan on global water security,2022,https://www.whitehouse.gov/wpcontent/uploads/2022/06/water-action-plan_final_formatted.pdf.

[6] David Rolnick.Priya L. Donti, Lynn H. Kaack,Kelly Kochanski,Alexandre Lacoste, Kris Sankaran,Andrew Slavin Ross,Nikola Milojevic-Dupont,Natasha Jaques,Anna Waldman-Brown,Alexandra Sasha Luccioni,Tegan Maharaj,Evan D.Sherwin,S.Karthik Mukkavlli,Konarad P.Kording,Carla P.Gomes,Andrew Y.Ng,Demis Hassabis Hassabis, John C. Platt, Felix Creutzig,Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. ACM Comput. Surv, 55(2), feb 2022

[7] A. Shehabi, S.J.Smith, N Horner, I. Azevedo, R. Brown,J. Koomey,E. Masanet, D.Sartor,M.Harrlin,and W.Lintner. United States data center energy usage report. Lawrence Berkeley National Laboratory,Berkeley,California. LBNL- 1005775,2016.

[8] David Patterson,Joseph Gonzalez,Urs Holzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia,Daniel Rothchild,David R.So,Maud Texier,and Jeff Dean. The carbon footprint of machine learning training will plateu, then shirnk.

[9] Roy Schwartz,Jesse Dodge, Noah A.Smith, and Oren Etzioni.Green ai. Common . ACM,63(12):54-63, nov 2020

[10] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP.In Proceedinga of the 57th Annual Meetting of the Association for Computational Linguistics, pages 3645-3650,Florence, Italy,July 2019. Association for Computational Linguistics.

[11] Peter Henderson,Jieru Hu, Joshua Romoff,Emma Brunskill,Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of maching learning. J.Mach.Learn. Res, 21(1), jan 2020

[12] The water resources group: Background, impact and the way forward, http://ww3.weforum.org/docs/WEF/WRG_Background_Impact_and_Way_Forward.pdf.

[13] United Nations . Water scarcity, https:// www. Unwater.org/water-facts/water-scarcity.

[14] United Nations Children's Fund(UNICEF). Water scarcity – addressing the growing lack of available water to meet children's needs., https:// www. unicef.org/wash/water-scarcity.

[15] Mohammad.A.Islam,Shaolei Ren,Gang Quan , Muhammad Z.Shakir,and Athanasios V. Vasilakos. Water-constrained geographic load balancing in data centers. IEEE Trans. Cloud Computing, 2015.

[16] Md Abu Bakar Siddik,Arman Shehabi, and Landon Marston. The environmental footprint of data centers in the United States. Environmental Research Letters, 16(6): 064017,2021.

[17] Urs Helzle. Our commitment to climate-conscious data center cooling 2022, https://blog.google/outreach-initiatives/sustainability/our-commitment-to-climate-conscious-data-center-cooling/.

[18] BMW Group. Next level mobility,2022, https://www.bmwgroup.com/En/report/2022/downloads/BMW-Group-Report-2022-en.pdf.

[19] Tesla.Impact report,2021,https://www.tesla.com/ns_videos/2021-tesla-impact.pdf

[20] researchgate.net/figure/Water-consumption-in-datacenter-fig-272158244

[21]DatacenterMap.Colocation USA, http:// www.datacentermap.com/usa/.

[22] Palm Beach Daily News. Drought dilemma: Palm Beach leaders struggle to cap water use, https://www.palmbeachpost.com/story/weather/2011/06/09/water-managers-shortage-near-crisis/7179447007/.

[23] dreamstime.com/cooling-towers-data-center-building-air-conditioning-cooling-towers-front-building-fins-to-front-industrial-image 193250314

,