A Water Efficiency Dataset for African Data Centers

Noah Shumba
Carnegie Mellon University, Africa

Opelo Tshekiso
Carnegie Mellon University, Africa

Pengfei Li
University of California, Riverside

Giulia Fanti
Carnegie Mellon University

Shaolei Ren
University of California, Riverside

Abstract

AI computing and data centers consume a large amount of freshwater, both directly for cooling and indirectly for electricity generation. While most attention has been paid to developed countries such as the U.S., this paper presents the first-of-its-kind dataset that combines nation-level weather and electricity generation data to estimate water usage efficiency for data centers in 41 African countries across five different climate regions. We also use our dataset to evaluate and estimate the water consumption of inference on two large language models (i.e., Llama-3-70B and GPT-4) in 11 selected African countries. Our findings show that writing a 10-page report using Llama-3-70B could consume about **0.7 liters** of water, while the water consumption by GPT-4 for the same task may go up to about 60 liters. For writing a medium-length email of 120-200 words, Llama-3-70B and GPT-4 could consume about **0.13 liters** and 3 liters of water, respectively. Interestingly, given the same AI model, 8 out of the 11 selected African countries consume less water than the global average, mainly because of lower water intensities for electricity generation. However, water consumption can be substantially higher in some African countries with a steppe climate than the U.S. and global averages, prompting more attention when deploying AI computing in these countries. Our dataset is publicly available on Hugging Face.

1 Introduction

With the rapid growth of artificial intelligence (AI) and digital services, the demand for data centers has increased substantially [1]. While data center infrastructure was historically lacking in Africa, the continent's burgeoning digital economy has recently led to a surge in data center constructions, with a projected market growth of 50% by 2026 compared to 2021 [2].

Nonetheless, data centers are notorious for their massive energy usage and water consumption, which have raised significant concerns even in developed countries such as the U.S. [3–5]. More critically, the added pressure on local water resources is particularly acute in Africa, where many countries are already grappling with extended droughts and water scarcity challenges [6, 7]. Therefore, it is important to assess data centers' water consumption in Africa, supporting healthy development of the data center industry for essential economic growth while ensuring responsible utilization of limited freshwater resources.

Unfortunately, while recent studies have begun to address the growing water consumption of data centers and AI computing [5, 8–10], they have predominantly focused on regions with large data

Tackling Climate Change with Machine Learning: Workshop at NeurIPS 2024.

center concentrations, such as the U.S. and Europe, while leaving out Africa—despite its rapid expansion of data centers and pressing challenges of water scarcity.

In this paper, we address the critical gap in the literature and present a first-of-its-kind water efficiency dataset for data centers in 41 African countries across five distinct climate regions. The dataset includes hourly estimates of water usage efficiency (WUE) for both direct and indirect water consumption over one year. We obtain these estimates by combining weather data from across Africa with the corresponding fuel mix data (i.e., the composition of energy sources in each country).

To demonstrate the utility of this dataset, we consider two recent large language models (LLMs), i.e., Llama-3-70B and GPT-4, and compare their water consumption for inference in African countries with that in the U.S. and globally. Our results for 11 representative African countries show that writing a 10-page report using Llama-3-70B could consume around 0.7 liters of water while the water consumption by GPT-4 for the same task may go up to nearly 60 liters of water.

Interestingly, our results also suggest that 8 out of the 11 selected African countries (including South Africa and Egypt) have a lower water consumption than the global average for performing the same task. This is due in part to a less water-intensive fuel mix for electricity generation in these countries. Additionally, some countries such as Morocco are even less water-consuming than the U.S. Nonetheless, the actual impacts of running AI model inferences on local water resources in these countries are still significant in light of the high water stress levels in African countries. On the other hand, for the same task, some countries such as Namibia are more water-consuming than the global average. Further compounded by the enduring regional water stress, the higher water consumption may prompt more attention when deploying AI services in these countries.

2 Background and Methodology

Our dataset is primarily based on the methodology and modeling of [5, 8], which study WUE and AI model water consumption with a heavy emphasis on the U.S. data centers. Like [8], we do not model supply chain manufacturing because this aspect often relies on generalized, less accurate data that may not reflect the unique operational practices of individual data centers or computing workloads. The methodology in [8] provides equations for modeling *onsite WUE*, which refers to water directly consumed/evaporated to cool down the facility for each unit of server energy consumption, and *offsite WUE*, which is also called the electric water intensity factor and refers to indirect water consumption by the generation of electricity that supplies each unit of data center energy. Note that water *consumption* is defined as "water withdrawal minus water discharge", i.e., the evaporated portion of water withdrawal that may not be immediately available for reuse [11]. Data centers commonly consume 80% of their direct freshwater withdrawal (in many cases, potable water), while only about 10% of the water withdrawal is consumed by typical households and offices [12].

Onsite WUE. To assess onsite WUE, [8] presents an empirical model created from a commercial cooling tower by considering two configurations. The first configuration is called "*fixed approach*", which fixes the differential between wet-bulb and cold water temperatures, and the second one "*fixed cold water temperature*" sets a constant cold water temperature. The WUE formulas for these two configurations are as follows:

$$\gamma_{\text{Approach}} = \left[-0.0001896 \cdot T_w^2 + 0.03095 \cdot T_w + 0.4442\right]^+, \tag{1}$$

$$\gamma_{\text{ColdWater}} = \left[0.0005112 \cdot T_w^2 - 0.04982 \cdot T_w + 2.387\right]^+, \tag{2}$$

where T_w is the wet-bulb temperature in Fahrenheit and $[x]^+ = \max\{0, x\}$. Unless otherwise noted, we will focus on Equation (2) and simply refer it to as onsite WUE in this paper, because it is typically easier to set a fixed cold water temperature without adjustment in real systems. While the onsite WUE for a cooling tower can differ from other cooling methods such as air economization with water evaporation, we note that cooling towers are one of the most commonly adopted and efficient heat rejection mechanisms for data centers [12, 13], especially in hot regions like Africa.

Offsite WUE. Electricity generation is water-intensive and must respond to the demand in real time to maintain grid stability. Thus, similar to carbon emissions associated with electricity usage, data centers are also accountable for the electricity water consumption. Technology companies including Meta have recently begun to include indirect water consumption for electricity generation in their sustainability reports [14]. This is critical for holistically understanding the true water impact of data centers, especially in regions where the energy mix includes significant hydroelectric and/or thermal power generation with high water intensities [11, 15]. Based on [8, 11], we present the offsite WUE formula as follows:

$$\gamma_{\text{off}}(t) = \frac{\sum_k e_k(t) \cdot w_k}{\sum_k e_k(t)}, \tag{3}$$

where $e_k(t)$ is the amount of electricity produced by energy fuel type k (e.g., hydroelectric, geothermal, coal) at time t and w_k is the corresponding water consumption or intensity factor in L/kWh.

3 Data Collection

We describe how we collect the necessary data to compute the onsite and offsite WUEs, respectively.

Weather data. For weather data, we first identify five distinct climate regions in Africa: *Rainforest, Savanna, Desert, Steppe, and Mediterranean regions* [16]. We then collect weather data from the countries for each climate region, consisting of hourly wet-bulb temperature, humidity, precipitation over one year from August 23, 2023 to August 22, 2024. All the weather data is obtained from WeatherAPI [17], which is collected via ground-based weather stations and satellite imagery. We then pick the high and low extremes in terms of the average wet bulb temperature for each region to obtain a representative range. The selected 11 representative countries are summarized in Table 1.

Table 1: Climatic regions and representative countries

Climatic Region	Representative Countries
Rainforest	Republic of the Congo, Gabon, Rwanda
Savanna	Morocco, Tunisia
Desert	Egypt, Libya
Steppe	Namibia, Ethiopia
Mediterranean	Algeria, South Africa

Energy fuel mix. We next collect the energy fuel mix (i.e., the composition of energy fuels) for electricity generation in each selected country sourced from OurWorldInData [18]. Due to the lack of access to fine-grained data, we use annual granularity for estimating the offsite WUE as done in the prior literature [11]. Additionally, we need the water intensity of each fuel type in each selected country to compute offsite WUE in (3). While direct data on the water consumption of various energy fuel types for African countries is lacking, [19] studies water withdrawal and consumption throughout different stages of energy production in Africa. Thus, we use [19] to derive the average water intensity for each energy fuel type in Africa.

4 Dataset Evaluation

Our final dataset provides onsite and offsite (hourly) WUE for capital cities in 41 African countries. For clarity, Figure 1 illustrates the monthly averages for a few selected countries in the rainforest and desert regions. The plot clearly illustrates seasonal trends, as well as onsite WUE differences of up to about 40% between climate regions. Specifically, desert regions generally are more water-consuming than rainforest regions, which is consistent with the observations in other places [5, 20]. Due to space limitations, we omit the figures for the other regions and offsite WUE while referring the readers to our dataset for details.

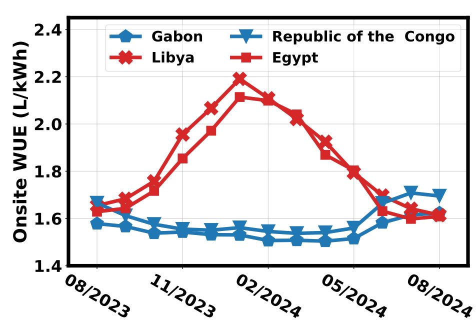
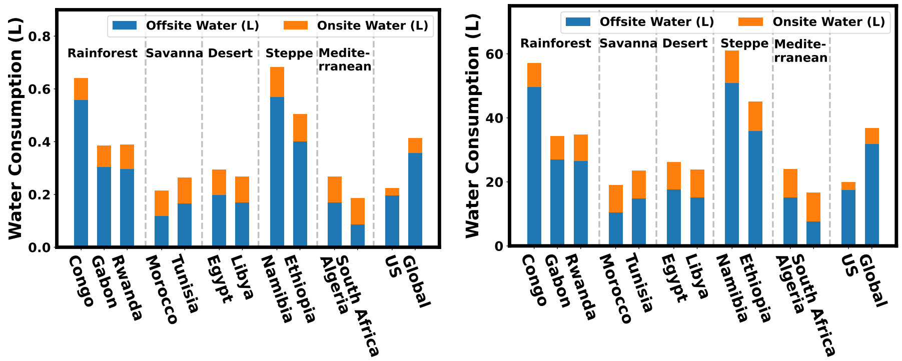


Figure 1: Average monthly onsite WUE for desert (red) and rainforest (blue) regions.

Estimating water consumption for AI models. To demonstrate the utility of our dataset, we use it to estimate the water consumption of two LLMs, i.e., Meta's Llama-3-70B and OpenAI's GPT-4, following the method in [5]. The tasks we evaluate are to write a comprehensive 10-page report and a medium-length email. *The details of estimating these AI models' water consumption and results for writing a medium-length email are available in the appendix.* Figures 2a and 2b indicate that writing a 10-page report using Llama-3-70B and GPT-4 in Africa could consume approximately **0.7 liters** and **60 liters** of water, respectively. In addition, we highlight the following points.



(a) Llama-3-70B

(b) GPT-4

Figure 2: Water consumption across 11 selected countries for writing a 10-page report (5,000 tokens) using Llama-3-70B and GPT-4, respectively.

First, 8 of the 11 selected African countries have a lower water consumption than the global average. In addition, Morocco and South Africa even have a lower water consumption than the U.S. average. This may be surprising, as Africa is commonly viewed as a water-scarce and dry continent. The main cause is that these countries generate electricity from energy fuels with relatively lower water intensities. Therefore, although the same AI model consumes more onsite water due to hotter weather, the significantly lower offsite water still makes these countries less water-consuming overall than the global average. Nonetheless, the actual impacts of running AI model inferences on local water resources in these countries are still significant in light of the high water stress levels in Africa.

Second, our results suggest a correlation between climate conditions with water consumption. Notably, countries categorized under the rainforest region (i.e., Republic of Congo, Rwanda, and Gabon) and the steppe climate (i.e., Ethiopia and Namibia) exhibit higher or roughly the same water consumption compared to the global average. We hypothesize several possible causes. First, the intrinsic hot and humid conditions of the rainforest climate and the dry, often hot conditions in steppe regions potentially degrade the onsite water efficiency for cooling compared to the global average onsite WUE. This effect is also observed from the onsite WUE differences among Microsoft's global data center locations. Second, the high offsite water consumption of these countries could suggest that countries in these regions rely more on water-intensive energy fuels like hydroelectric or thermo-electric power. Indeed, we observe this empirically — countries with high offsite WUE, such as the Republic of the Congo and Ethiopia rely almost entirely on hydroelectric power.

Third, we show in Figure 3 the water consumption and (scope-2) carbon emission across various African countries for writing a 10-page report using the Llama-3-70B model. We see a tradeoff between water consumption and carbon emission, which is consistent with the findings in prior studies [9]. This prompts further attention to strike a balance between water consumption and carbon emission to enable truly sustainable AI in African countries.

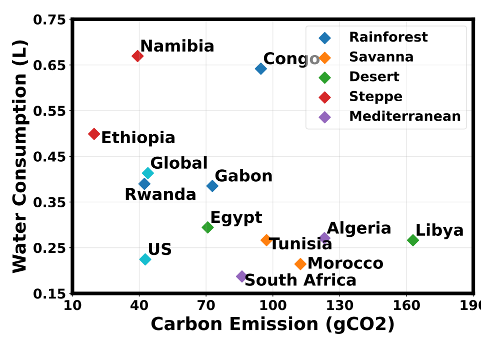


Figure 3: Water consumption and (scope-2) carbon emission across various African countries for writing a 10-page report using the Llama-3-70B model.

Finally, we emphasize the potential uncertainties in our quantitative results. For instance, it is challenging to obtain precise data on the energy fuel mix and the electricity water intensity in Africa. Moreover, the actual energy consumption of LLM inference may vary depending on the (possibly customized) optimization techniques used by real systems, particularly for the proprietary GPT-4 model. As such, our results should be regarded as first-order estimates rather than precise representations. We encourage AI model developers and data center operators to enhance transparency regarding their most recent water usage, especially in African countries facing water scarcity.

5 Conclusion

In this paper, we present the first-of-its-kind dataset of onsite WUE and offsite WUE for data centers in 41 African countries across five different climate regions. We also use our dataset to evaluate and estimate the water consumption of inference on Llama-3-70B and GPT-4 in 11 selected countries. Our findings underscore the need for region-specific adaptations in data centers, particularly in cooling systems that can operate efficiently under varying climatic conditions without substantially escalating water usage. Moreover, the reliance on water-intensive energy sources prompts a broader discussion on the water sustainability practices within the data center industry. By understanding the water usage efficiency in these different countries, we can make informed decisions that promote sustainable and responsible water use while supporting the growing demand for AI and computing services in Africa.

Acknowledgement

Noah Shumba, Opelo Tshekiso, and Giulia Fanti acknowledge the support of the Bill & Melinda Gates Foundation. Pengfei Li and Shaolei Ren were supported in part by the NSF under the grant CCF-2324916.

References

References

[1] H. C. Granade, J. Creyts, A. Derkach, P. Farese, S. Nyquist, and K. Ostrowski, "McKinsey global energy and materials: Unlocking energy efficiency in the U.S. economy," Jul. 2009.

[2] Africa Data Centres Association, "Data centres in Africa focus report 2024." `https://africadca.org/en/data-centres-in-africa-focus-report-2024`, 2024.

[3] K. Ahmed, M. A. Islam, S. Ren, and G. Quan, "Can data center become water {Self-Sufficient}?," in *6th Workshop on Power-Aware Computing and Systems (HotPower 14)*, 2014.

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

[5] P. Li, J. Yang, M. A. Islam, and S. Ren, "Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models," *Communications of the ACM*, 2024 (accepted).

[6] UNICEF, "Water crisis in the horn of africa." `https://www.unicef.org/documents/water-crisis-horn-africa`. (Accessed on 11/25/2024).

[7] The Water Project, "The water crisis: Poverty and water scarcity in africa." `https://thewaterproject.org/why-water/poverty`. (Accessed on 11/25/2024).

[8] P. S. Gupta, M. R. Hossen, P. Li, S. Ren, and M. A. Islam, "A dataset for research on water sustainability," in *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, pp. 442–446, 2024.

[9] M. A. Islam, K. Ahmed, H. Xu, N. H. Tran, G. Quan, and S. Ren, "Exploiting spatio-temporal diversity for water saving in geo-distributed data centers," *IEEE Transactions on Cloud Computing*, vol. 6, no. 3, pp. 734–746, 2018.

[10] W. E. Gnibga, A. A. Chien, A. Blavette, and A. C. Orgerie, "Flexcooldc: Datacenter cooling flexibility for harmonizing water, energy, carbon, and cost trade-offs," in *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, e-Energy '24, (New York, NY, USA), p. 108–122, Association for Computing Machinery, 2024.

[11] P. Reig, T. Luo, E. Christensen, and J. Sinistore, "Guidance for calculating water use embedded in purchased electricity," *World Resources Institute*, 2020.

[12] Google, "Environmental report." `https://sustainability.google/reports/`, 2024.

[13] Equinix, "Sustainability report." https://sustainability.equinix.com/wp-content/uploads/2024/07/Equinix-Inc_2023-Sustainability-Report.pdf, 2024.

[14] Meta, "Sustainability report." https://sustainability.atmeta.com/2024-sustainability-report/, 2024.

[15] M. A. B. Siddik, A. Shehabi, and L. Marston, "The environmental footprint of data centers in the United States," *Environmental Research Letters*, vol. 16, no. 6, p. 064017, 2021.

[16] A. Kröner, J. I. Clarke, R. W. Steel, D. N. McMaster, R. K. Gardiner, K. B. Dickson, A. L. Mabogunje, A. Smedley, J. F. Middleton, and D. S. Nicol, "Africa." Encyclopedia Britannica, Aug 2024. Accessed: 30 August 2024.

[17] "WeatherAPI." https://www.weatherapi.com/. Accessed: 30 August 2024.

[18] H. Ritchie and P. Rosado, "Energy mix," *Our World in Data*, 2020. https://ourworldindata.org/energy-mix.

[19] R. G. Sanchez, R. Seliger, F. Fahl, L. De Felice, T. B. Ouarda, and F. Farinosi, "Freshwater use of the energy sector in africa," *Applied Energy*, vol. 270, p. 115171, 2020.

[20] L. Karimi, L. Yacuel, J. Degraft-Johnson, J. Ashby, M. Green, M. Renner, A. Bergman, R. Norwood, and K. L. Hickenbottom, "Water-energy tradeoffs in data centers: A case study in hot-arid climates," *Resources, Conservation and Recycling*, vol. 181, p. 106194, 2022.

[21] "2023 equinix sustainability report." https://sustainability.equinix.com/wp-content/uploads/2024/07/Equinix-Inc_2023-Sustainability-Report.pdf, 2023. [Accessed 29-08-2024].

[22] N. Walsh, "How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability," *Microsoft Azure Blog*, April 2022.

[23] International Energy Agency, "Electricity 2024." https://www.iea.org/reports/electricity-2024, 2024.

[24] B. Tomlinson, R. W. Black, D. J. Patterson, and A. W. Torrance, "The carbon emissions of writing and illustrating are lower for AI than for humans," *Scientific Reports*, vol. 14, February 2024.

[25] P. Patel, E. Choukse, C. Zhang, A. Shah, I. Goiri, S. Maleki, and R. Bianchini, "Splitwise: Efficient generative LLM inference using phase splitting," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pp. 118–132, 2024.

[26] J. Stojkovic, C. Zhang, I. Goiri, J. Torrellas, and E. Choukse, "DynamoLLM: Designing LLM inference clusters for performance and energy efficiency," in *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.

[27] A. B. Samuel Rincé and V. Defour, "Ecologits calculator." https://huggingface.co/spaces/genai-impact/ecologits-calculator, 2024.

[28] P. Patel, E. Choukse, C. Zhang, I. n. Goiri, B. Warrier, N. Mahalingam, and R. Bianchini, "Characterizing power management opportunities for llms in the cloud," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS '24, (New York, NY, USA), p. 207–222, Association for Computing Machinery, 2024.

[29] OpenAI, "OpenAI API Pricing." https://openai.com/api/pricing/.

[30] Africa Data Centres Association, "African climate & datacenter PUE 2021." https://africadca.org/wp-content/uploads/2022/05/PUE-Climate-Hydrogen-in-African-DCs-White-paper.pdf, 2021.

A Estimating Water Consumption for LLM Inference

We now present the details of estimating waster consumption by two LLMs, i.e., Meta's Llama-3-70Angola B and OpenAI's GPT-4. We use the following equations to calculate the onsite and offsite water consumption, respectively:

$$W_{\text{on}} = \gamma_{\text{on}} \cdot E \quad \text{and} \quad W_{\text{off}} = \gamma_{\text{off}} \cdot \rho \cdot E,$$

where W is the water consumption, γ is the WUE, ρ is the power usage effectiveness (PUE), E is the server energy consumption for AI models, and the subscript "on" and "off" denote "onsite" and "offsite" wherever applicable, respectively. Thus, to estimate the LLMs' water consumption, we need their onsite and offsite WUEs, energy consumption, as well as the PUE.

A.1 WUE

For African countries, we use the average onsite and offsite WUEs from our dataset. For the U.S. and global references, their average onsite WUEs are obtained from publicly accessible reports based on Microsoft's U.S. average (0.55 L/kWh) and Equinix's global average (1.07 L/kWh), which represent efficient hyperscale and colocation data centers, respectively [21, 22]. Their average offsite WUEs are acquired from the World Resource Institute report [11].

A.2 Energy consumption

The exact LLM inference energy is often lacking in the public domain, especially for those powerful but proprietary LLMs such as GPT-4 deployed in real-world inference systems. To estimate LLM inference energy, some studies resort to a commonly cited claim that each request of the GPT model family underlying ChatGPT consumes about 10x the energy as a Google search [23], while others use GPUs' processing capability in tera operations per second (TOPS) and the power consumption reported by manufacturers [24]. In practice, however, the actual LLM inference energy consumption depends on a variety of factors, including the hardware, service level objectives (SLOs), and system optimization [25, 26].

To estimate LLM inference energy consumption for a user-facing application (used by, e.g., ChatGPT), we resort to an online calculator [27] and a recent study [26]. These two sources use different methods to calculate the LLM inference energy consumption, which we describe as follows.

A.2.1 LLM inference energy estimates by [26, 27]

The online calculator [27] estimates the inference energy consumption by various LLMs based on a transparent methodology. It first uses the energy measurements from a set of open-sourced models (mostly on Nvidia A100 GPUs) to fit an energy consumption curve in terms of the number of model parameters. For mixture-of-expert model architectures such as the one commonly believed to be used by GPT-4, a range of active model parameters are considered. The energy measurement takes into account a server's non-GPU power attributed to the model depending on the fraction of GPU resources the model utilizes. Nonetheless, [27] only considers the token generation phase, while neglecting the prompt processing phase (i.e., processing user prompts to generate the first output token) which is also energy-intensive [25, 26, 28]. In addition, it does not consider batching and essentially models a lightly-loaded system without request contention. This can be viewed as a reference system used by industries, e.g., [25, 28] measure LLM inference energy and power consumption without batching as a reference value for energy and power provisioning, while [25, 26] use the latency measurement in such a reference system to set real SLO targets.

On the other hand, [26] measures the actual GPU energy consumption for LLMs on enterprise-grade Nvidia DGX H100 servers. Its measurement also considers "state-of-the-practice" optimization techniques commonly used in real systems, including batching. Importantly, it considers the prompt processing phase and representative SLO targets, which are both crucial for real-world LLM deployment.

Energy estimates assuming a fully utilized system without accounting for SLOs may not reflect the industry practice, since a fully-utilized system can lead to significant SLO violations, which are not tolerable in real-world LLM deployment, especially for commercial LLM applications such as real-time conversations that have strict SLO targets to deliver good quality of experiences [25, 26]. As a result, server resources for LLM inference are typically provisioned based on the peak demand to ensure SLOs are met at all times. In other words, the LLM inference servers may not be highly utilized under non-peak loads, resulting in a high energy consumption per request. For example, the Llama-2-70B inference for a medium-length request on H100 GPUs consumes 9.4 Wh energy without batching [25], while the inference energy consumption is still over 4.0 Wh when batching is applied under various system loads using "state-of-the-practice" optimizations (the last column in Table II) [26].

The measurement in [26] only includes the GPU energy consumption for a small set of open LLMs. To account for the non-GPU server energy consumption, we need to multiply the energy consumption in [26] by a factor of $1.5 \sim 2.0$ based on the server power provisioning breakdown [28].

While [27] and [26] use different methodologies, we note that the server-level inference energy consumption estimated by [27] (as of November 20, 2024) is generally lower than that measured by [26] for the same model size, assuming the LLM inference system is optimized using state-of-the-practice techniques in [26]. For example, for Llama-3-70B to write a medium-length email with 250 tokens (or about 120-200 words), [27] estimates the inference energy consumption as 2.62 Wh (after removing the PUE of 1.2 for data center overheads), whereas [26] shows the server-level energy consumption is about 10 Wh (after multiplying the value in the last column of Table III by 1.6 to account for the non-GPU server energy).

This result might be surprising, as [27] does not consider system optimization or batching whereas [26] uses reasonable "state-of-the-practice" optimizations including batching. Nonetheless, [27] mostly uses A100 GPUs and does not consider prompt processing energy consumption, whereas [26] uses H100 GPUs (which may be more energy-consuming than A100 GPUs for LLM inference as shown by [25]) and considers both prompt processing and token generation energy consumption. Additionally, the strict SLOs in real-world deployment prohibits the LLM inference system from being fully utilized. Thus, the LLM inference energy consumption estimated by [27] without system optimization could be even lower and still serves as a good reference point.

A.2.2 Energy consumption for writing a 10-page report and a medium-length email

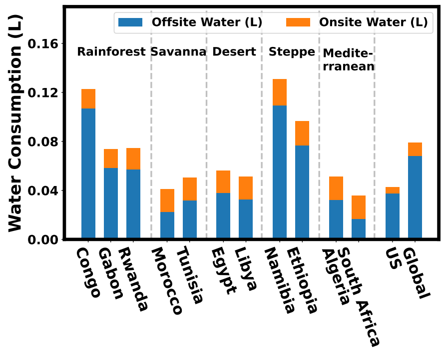
For the task of writing a 10-page report,[1] we assume the output is 5,000 tokens and use the estimates by [27], since the energy measurement results in [26] do not include generating such long outputs using Llama-3-70B. After removing the PUE of 1.2 for data center overheads, we estimate that the energy consumption to write a 5,000-token text by Llama-3-70B and GPT-4 are 52.25 Wh and 4.66 kWh, respectively, based on the results in [27] as of November 20, 2024. Note that, due to the proprietary nature of GPT-4, [27] assumes 1,760 billion parameters for GPT-4 with a mixture-of-expert architecture based on the best-known information from various public sources. Additionally, the energy consumption estimates for models with such large model sizes are based on extrapolation. As a result, without detailed information from model owners, the energy estimates for large proprietary models may have less accuracy than for small/medium open models.

For the task of writing a medium-length email, we assume the output is 250 tokens or about 120-200 words. For Llama-3-70B, by considering a medium-length prompt and a medium system load, we estimate the inference energy consumption as ~ 10 Wh after multiplying the value in the last column of Table III by 1.6 to account for the non-GPU server energy [26]. For GPT-4, we estimate the inference energy consumption as ~ 232 Wh [27].

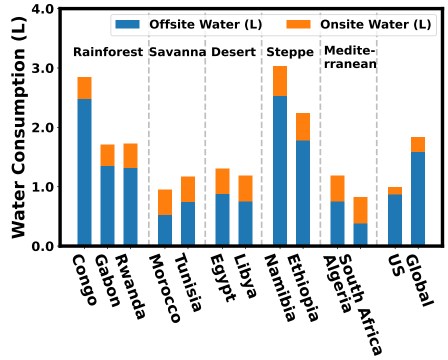
A.3 PUE

PUE is a metric that assesses the energy efficiency of a data center by comparing the total energy consumed by the facility to the energy used by the computing equipment. The ideal PUE is 1.0, indicating 100% energy efficiency in computing. The inference energy estimate provided by [27] assumes a default PUE of 1.2. The PUE overhead is not needed for calculating the onsite water consumption, but should be considered when assessing the offsite water consumption. For different African countries, we consider an average country-/region-wise PUE provided by [30]. By taking the lowest when multiple values are presented in [30], the PUE values for the 11 selected African countries are: 2.3 for Algeria, 2.3 for Egypt, 1.5 for Ethiopia, 1.9 for Gabon, 2.3 for Libya, 2.3 for Morocco, 2.1 for Namibia, 2.0 for Republic of the Congo, 1.4 for South Africa, 2.3 for Tunisia, and 1.7 for Rwanda. We consider Microsoft's U.S. average PUE of 1.17 [22] and Equinix's global average PUE of 1.42 [21] for the U.S. and global averages, respectively.

[1] The calculator [27] assumes 5,000 tokens for a 5-page report. Based on the token-to-word ratio [29], we consider 5,000 tokens as roughly a 10-page report.



(a) Llama-3-70B



(b) GPT-4

Figure 4: Water consumption across 11 selected countries for writing a medium-length email (250 tokens) using Llama-3-70B and GPT-4, respectively.

B Additional Results

In addition to the task of writing a 10-page report, we also examine the water consumption of Llama-3-70B and GPT-4 on a more common task: generating a medium-length email, averaging 250 tokens. We show the results in Figure 4.