Assessing Carbon Footprint Estimations of ChatGPT

Ithier d'Aramon[1*], Boris Ruf[2], and Marcin Detyniecki[2]

[1] Albert - Business & Data School, Paris, France idaramon@albertschool.com
[2] AXA Group Operations, AI Research, Paris, France
{boris.ruf,marcin.detyniecki}@axa.com

Abstract. ChatGPT takes the world by storm, while its environmental impact remains a serious concern. In the absence of official data, several researchers have tried to estimate the carbon emissions linked to the service, with very deviating results. We reproduce three popular calculations using an open data model for carbon footprint quantification. This enables a transparent comparison of the approaches and helps to identify their main differences and possible potential for improvement. Our work demonstrates how open data models can be used to lead the way to more robust carbon footprint estimation.

Keywords: carbon footprint quantification · large language models · digital sustainability.

1 Introduction

ChatGPT, created by OpenAI, is a publicly accessible tool that utilizes the large language model GPT-3[14]. ChatGPT's remarkable capability to produce language resembling that of humans and accomplish intricate tasks represents a noteworthy breakthrough within the realm of natural language processing and artificial intelligence. It is estimated to have reached 100 million monthly active users in January 2023, just two months after launch, making it the fastest-growing consumer application in history[17].

One drawback, however, is that training extensive neural networks like GPT-3 is known to entail considerable computational expenses, resulting in a significant demand for energy[12]. Recent accumulation of natural disasters has accelerated awareness of the unfolding climate crisis and increased the drive of policymakers and society to act[16,9]. While the important role of energy-related greenhouse gases (GHGs) in climate mitigation is well understood, quantifying the specific environmental impact of goods and services remains complex[5]. Measuring the "carbon footprint" involves collecting data on GHGs from various sources, calculating their carbon dioxide equivalent and aggregating the results.

In particular the incoming Artificial Intelligence Act in the European Union could give the matter a boost. In Article 28b 2. (d) of an amendment which was

* Work done during internship at AXA.

adopted in June 2023 by the European Parliament, providers of large language models (foundation models) are explicitly required to "mak[e] use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency, and the overall efficiency of the system. [...] They shall be designed with capabilities enabling the measurement and logging of the consumption of energy and resources, and, where technically feasible, other environmental impact the deployment and use of the systems may have over their entire lifecycle"[11].

OpenAI has not disclosed any specific information regarding the energy consumption or carbon emissions associated with its services. Therefore, various researchers have proposed estimations of their carbon footprint in order to narrow the information gap and support number driven decision making.

In the absence of official carbon emission reports by OpenAI, the carbon footprint of ChatGPT can only be estimated and extrapolated based on better studied systems with comparable features. However, carbon footprint modeling is a complex task which depends on many factors and includes several unknown variables. The proposed approximations take different approaches and their results deviate considerably. The presence of distinct underlying assumptions and variations in how the calculations are presented make it hard to evaluate and compare the propositions.

In this paper, we study and compare three widely considered estimations of ChatGPT. In the next section, we briefly introduce our methodology. Afterwards, we outline the approaches in detail. Finally, we compare the outcome and discuss the different estimates.

2 Methodology

We selected three carbon footprint studies for ChatGPT and reproduce each of them. For the in-depth analysis we use an open data model for carbon footprint quantification[15]. The advantage of formalizing the studies' calculations with such a methodology is to increase their accessibility. Using an open source data viewer[3] which interprets the data records, the carbon footprint scenarios can be interactively explored in a web browser. The viewer also automatically handles unit conversion and finds a common ground based on the availability of emission data by type of emission. Another benefit of the data model is that references to the source of information for all underlying assumptions upgrade the documentation of the estimate and thus improve transparency and traceability.

It is important to note that we only examine the inference costs, i.e. the impact on the environment that occurs during the operation of the model. The also very high energy consumption that occurs during the training process of the model is not considered in this paper.

[3] `https://github.com/borisruf/carbon-footprint-modeling-tool`

3 Carbon scenarios

In this section, we replicate the results of three popular carbon footprint estimates for GPT-3, the language model used by ChatGPT. For better comparability, we consider for all scenarios the period of January 2023, when ChatGPT is reported to have had about 100M unique users who made 590M queries[10]. All data records created in the course of this study have been published on GitHub and are linked at the appropriate place in the text.

3.1 Scenario 1

Raghavenda Selvan, Assistant Professor at the University of Copenhagen, has estimated the carbon footprint of ChatGPT for Süddeutsche Zeitung, a German daily[6]. He concludes that emissions associated with the service could have accounted for 24.24 kt CO2e in January 2023 (see interactive online scenario with reproduced data model for details[4]).

In the absence of reliable data on the energy consumption of GPT-3, Selvan used the freely available open source language model GPT-J as basis for his calculation[18]. To estimate the carbon footprint Selvan ran experiments on a local workstation. The model was initialized in half-precision (16 bit float instead of 32 bit float) to fit into GPU memory. The average length of text generated by the model was 230 words. The energy consumption got measured with CarbonTracker, a tool developed by researchers from University of Copenhagen[2]. As a result, one request to GPT-J consumed in average 0.01292 kWh.

Scaling to GPT-3, Selvan took the following factors into account: First, GPT-J only has 6 billion parameters, whereas GPT-3 has a much larger parameter count of 175 billion. Thus, Selvan assumes by a conservative estimate that the larger model could require 10x additional GPUs. Second, to fit the test environment, the GPT-J model was initialized in half-precision. The overhead of running a full precision model compared to half precision is considered about 1.5x more. In conclusion, Selvan estimates that submitting a query to GPT-3 requires 15x more energy compared to submitting a query to GPT-J.

Finally, Selvan uses the grid emission factor for Denmark in 2020 (0.212 kgC02e/kWh) to compute the GHGs.

3.2 Scenario 2

Chris Pointon also shared some consideration about the carbon footprint of ChatGPT[13]. Based on his calculation, the emissions linked to the service in January 2023 would amount to 225.64t CO2e (see interactive online scenario[5]).

Pointon bases his calculation on a statement made by Tom Goldstein, a Professor at the University of Maryland[4]. Based on experiments Goldstein had

[4] `https://borisruf.github.io/carbon-footprint-modeling-tool/index.html?id=gpt-selvan-0`

[5] `https://borisruf.github.io/carbon-footprint-modeling-tool/index.html?id=gpt-pointon-0`

conducted with a modified BERT model, he estimates that if GPT-3 were executed on a single Nvidia A100 GPU, it would require approximately 350 milliseconds to generate a single word[3]. Accordingly, Pointon takes the documented energy consumption of this processor under full load (407 Watts per hour) and concludes that one request to generate a single word consumes 0.03957 Watts[1]. Finally, he assumes that each query to GPT-3 produces in average 30 words.

Pointon expects ChatGPT to be hosted in an Azure datacenter in California and therefore uses the grid emission factor for Western USA in 2021 (0.322167 kgC02e/kWh) to compute the emissions.

3.3 Scenario 3

Kasper Ludvigsen approximated the carbon footprint of ChatGPT in a blog post on Towards Data Science[8]. Based on his estimation, the emissions caused by ChatGPT could have accounted for 752.84 t CO2e in January 2023 (see interactive online scenario[6]).

Ludvigsen bases his calculation on a study by Alexandra Luccioni and others who investigated the carbon footprint of the BLOOM model[7]. With 176 billion parameters, this large language model is of similar size as GPT-3. The researchers measured the energy demand while running the model during 18 days on 16 Nvidia A100 GPUs. In total, their system consumed 914 kWh while handling 230,768 requests. Based on this information, Ludvigsen derives the average energy consumption for one query and extrapolates the total demand.

Similar to Pointon, Ludvigsen uses the grid emission factor for Western USA in 2021 (0.322167 kgC02e/kWh) to finally estimate the emissions.

4 Comparison

A summary of key assumptions and results of the three scenarios can be found in Table 1. For a more interactive comparison, we refer to our online benchmark of the different approaches[7].

Overall, it can be observed that the estimates are very far apart. Scenario 1 is the most pessimistic, estimating a carbon footprint of 24.24 kt CO2e which is over 32 times higher than Scenario 3, which predicts 752.84 t CO2e, and even 107 times higher than Scenario 2 with 225.64 t CO2e.

We noticed that Selvan uses the grid emission factor for Denmark, while OpenAI's servers are more likely to be located in the US, as also assumed by the other two authors. However, since the Danish electricity grid is actually less carbon-intensive, adjusting for this would only widen the gap.

On the other hand, Selvan assumes a response to include 230 words in average, while Pointon only calculates with 30 words. Adjusting the data model

[6] `https://borisruf.github.io/carbon-footprint-modeling-tool/index.html?id=gpt-ludvigsen-0`

[7] `https://borisruf.github.io/carbon-footprint-modeling-tool/benchmark.html?ids[]=gpt-selvan-0&ids[]=gpt-pointon-0&ids[]=gpt-ludvigsen-0`

accordingly, Scenario 2 predicts emissions of 1.73 kt CO2e. Ludvigsen does not specify the number of words per response.

Table 1. Key assumptions and results of the different scenarios

	Reference model	Words	Electricity grid	Energy per query (kWh)	Estimated emission in 01-2023 (CO2e)
1	GPT-J	230	Denmark (2020)	0.01292	24.24 kt
2	BERT variation	30	Western USA (2021)	0.00119	225.64 t
3	BLOOM	-	Western USA (2021)	0.00396	752.84 t

As basis for their estimations, all authors take the measured energy consumption of openly accessible large language models. Only Selvan conducted his own experiments, while the other two authors rely on results published by others. The model used by Selvan has fewer parameters than GPT-3, he simply extrapolates his measurements by multiplying them by a factor. Pointon reconstructs the energy consumption via an estimated query duration by another researcher who extrapolated this value from a 3-billion parameter model. Ludvigsen takes a model of similar size as GPT-3 as reference.

All estimates calculate the average values of individual requests and then reuse these values to extrapolate the energy demand over a longer period of time. It worth to note that such a practice is prone to accumulating rounding errors.

Overall, it is obvious that due to the many unknowns in the equation, the range of possible outcomes is very wide. Therefore, more details about the architecture behind ChatGPT as well as usage reports are urgently needed for making more accurate predictions.

5 Conclusion

We utilized an open data model for carbon footprint quantification to reproduce three carbon footprint estimates for ChatGPT. We show that this approach allows for a more efficient comparison of different estimation scenarios, and also promotes discussions regarding their potential drawbacks. This finding emphasizes the importance of transparent and robust carbon footprint modeling.

References

1. Cloud Carbon Footprint, `https://www.cloudcarbonfootprint.org/docs/methodology/#appendix-iii-gpus-and-minmax-watts`, Accessed on 22-06-2023
2. Anthony, L.F.W., Kanding, B., Selvan, R.: Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems (July 2020)

3. Geiping, J., Goldstein, T.: Cramming: Training a language model on a single GPU in one day (2022)
4. Goldstein, T.: Twitter https://twitter.com/tomgoldsteincs/status/1600196995389366274, Accessed on 22-06-2023
5. IEA: CO2 Emissions in 2022 https://www.iea.org/reports/co2-emissions-in-2022, Accessed on 21-06-2023
6. Landwehr, T.: Der Energiehunger der KIs. Süddeutsche Zeitung https://www.sueddeutsche.de/wissen/chat-gpt-energieverbrauch-ki-1.5780744, Accessed on 21-06-2023
7. Luccioni, A.S., Viguier, S., Ligozat, A.L.: Estimating the carbon footprint of BLOOM, a 176b parameter language model (2022)
8. Ludvigsen, K.G.A.: The carbon footprint of (ChatGPT (March 2023), https://towardsdatascience.com/chatgpts-electricity-consumption-7873483feac4
9. Masson, T., Fritsche, I.: We need climate change mitigation and climate change mitigation needs the 'we': a state-of-the-art review of social identity effects motivating climate change action. Current Opinion in Behavioral Sciences **42**, 89–96 (2021), human Response to Climate Change: From Neurons to Collective Action
10. Milmo, D.: ChatGPT reaches 100 million users two months after launch. The Guardian https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app, Accessed on 22-06-2023
11. Parliament, E.: Report on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf, Accessed on 21-06-2023
12. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon emissions and large neural network training (2021)
13. Pointon, C.: The carbon footprint of chatgpt (Dec 2022), https://medium.com/@chrispointon/the-carbon-footprint-of-chatgpt-e1bc14e4cc2a
14. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
15. Ruf, B., Detyniecki, M.: Open and linked data model for carbon footprint scenarios. International Conference on Renewable Energy and Conservation (ICREC) (Nov 2022)
16. Stollberg, J., Jonas, E.: Existential threat as a challenge for individual and collective engagement: Climate change and the motivation to act. Current Opinion in Psychology **42**, 145–150 (2021), psychology of Climate Change (2021)
17. UBS: Has the AI rally gone too far? https://www.ubs.com/global/en/wealth-management/insights/chief-investment-office/house-view/daily/2023/latest-25052023.html, Accessed on 21-06-2023
18. Wang, B., Komatsuzaki, A.: GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax (May 2021)