Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 122 (2024) 707–712

www.elsevier.com/locate/procedia

31st CIRP Conference on Life Cycle Engineering (LCE 2024)

Estimating the environmental impact of Generative-AI services using an LCA-based methodology

Adrien Berthelot[a,b,*], Eddy Caron[a], Mathilde Jay[a,c], Laurent Lefèvre[a]

[a] *Univ. Lyon 1 UCBL, EnsL, UCBL, CNRS, Inria, LIP, Lyon, France*
[b] *OCTO Technology, Paris, France*
[c] *Univ. Grenoble Alpes, Grenoble INP, LIG, Grenoble, France*

* Corresponding author. *E-mail address:adrien.berthelot@ens-lyon.fr*

Abstract

As digital services are increasingly being deployed and used in a variety of domains, the environmental impact of Information and Communication Technologies (ICTs) is a matter of concern. Artificial intelligence is driving some of this growth but its environmental cost remains scarcely studied. A recent trend in large-scale generative models such as ChatGPT has especially drawn attention since their training requires intensive use of a massive number of specialized computing resources. The inference of those models is made accessible on the web as services, and using them additionally mobilizes end-user terminals, networks, and data centers. Therefore, those services contribute to global warming, worsen metal scarcity, and increase energy consumption. This work proposes an LCA-based methodology for a multi-criteria evaluation of the environmental impact of generative AI services, considering embodied and usage costs of all the resources required for training models, inferring from them, and hosting them online. We illustrate our methodology with Stable Diffusion as a service, an open-source text-to-image generative deep-learning model accessible online. This use case is based on an experimental observation of Stable Diffusion training and inference energy consumption. Through a sensitivity analysis, various scenarios estimating the influence of usage intensity on the impact sources are explored.

Keywords: Generative AI; Life Cycle Analysis; Digital services; Greenhouse Gas Emission; Energy; Methodology

1. Introduction

The successive emergence of image-creation software in the summer of 2022, followed by high-performance conversational agents like Chat-GPT a few months later, both based on Artificial Intelligence (AI), has popularized the term generative AI (Gen-AI). This term now designates a sector that claims to be growing as fast as it is significant [1], integrating its functionalities into numerous existing services, online research, software development, image editing, etc.

In the context of environmental challenges and considering the footprint of the digital sector, which in the EU already accounts for 9.3% of electricity consumption and over 4% of greenhouse gas emissions [2], many studies [3, 4, 5, 6, 7] have addressed the question of the environmental cost of AI. However, the majority of studies are limited to measuring the electricity consumed by creating these AI [6], and possibly deducing their contribution to global warming via the carbon intensity of the electricity mix used. By neglecting the conditions and resources required for deploying AI applications, such approaches miss out on a significant part of the environmental impact. Moreover, this tendency towards carbon tunnel vision does not take into account impact categories, such as metal scarcity, which may be prevalent in ICT [2].

Gen-AI services can be defined as a type of digital services, a system using digital goods and infrastructures [8], centered around the use of a Gen-AI model. As these model are particularly computationally intensive during the training and use phases, their growing use on a large scale is creating many unknowns about their footprint, both in terms of characterization and order of magnitude. To fill these gaps, this work proposes a methodology based on Life Cycle Assessment (LCA) and observes the use of these models as digital services. This method provides a multi-criteria evaluation of Gen-AI-based services, taking into account the life cycle of the involved equipment and the particular costs associated with training and inference, thanks to a combination of measurements and models derived from our experiments.

10.1016/j.procir.2024.01.098

Section 2 defines the principle of AI service and provides a review of existing literature. Section 3 describes the methodology: which equipment is included and how their embedded and electricity cost is accounted for. In Section 4 illustrates our methodology with Stable Diffusion [9], an open-source text-to-image generative deep-learning model released in 2022. Its service impact is analyzed for both a single use and one year of service. Complementary Sensitivity Analyses (SA) highlight the critical parameters influencing the environmental impact of Gen-AI. Section 5 questions the hypothesis made, shows the limitations of the methodology, and what would be needed to overcome them.

2. Environmental impact of Artificial Intelligence

2.1. Definitions

AI can be defined as the ability of a computer to automate a task that would usually require human discernment [10]. Machine Learning (ML) and Deep Learning (DL) refer to AIs that learn from existing data. Such AIs are generally referred to as models. Generative AI (Gen-AI) is a type of AI that produces new content, for example, human-like discussions and realistic images [11]. Developing ML models requires collecting data and learning from the data, which includes (1) selecting the best model and learning algorithm for the given task and (2) applying this algorithm to the model and the collected data. This second step is called training. Once the model has reached the targeted quality, it can be used on new data, which is referred to as the inference phase. Training and inferring DL models require more and more computations and Graphic Processing Units (GPUs) are today the most common processing unit for dealing with AI workloads.

2.2. Related work

Interest in the energy consumption of AI started in 2019 with the work of Strubell et al. [5]. Methodologies on how to measure or estimate it have been explored in several approaches [12, 13, 3]. However, they solely focus on training. A more recent study of the BLOOM model [4] reports the energy consumed by the training, an estimation of the global cost of inference, and an estimation of the manufacturing cost based on LCA. Wu *et al.* [7] explore the environmental impact of AI in various phases (Data, Experimentation, Training, and Inference). The authors study the impact from the carbon emission perspective and include the life cycle of hardware systems. However, they fail to provide evaluations of AI applications. In summary, none of the previous work integrates the whole life cycle cost of the involved ICT equipment on a multi-criteria assessment. To fill these gaps, this work proposes an evaluation of AI as a service, as described in the next section.

2.3. AI as a Service

The work focuses on AI as a service, which means that not only the specific costs of AI, i.e. training and inference phases, are integrated but also the infrastructure required to use it online, as a service, assuming that the AI is accessible through a

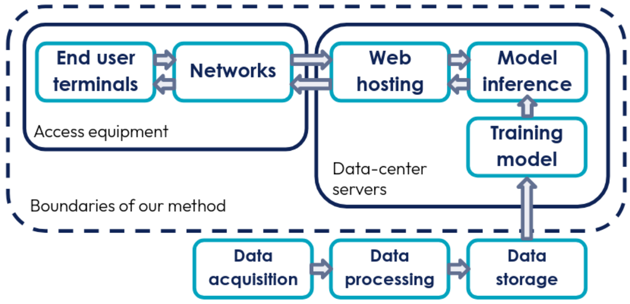

Fig. 1. Structure of Gen-AI service as considered in the work

website interface. As shown in Fig. 1, a Gen-AI service considered in this study concerns the sum of end-user terminals, networking, web hosting, model inference, model training, and data management.

3. Methodology

The methodology developed in this paper is based on LCA and a reproducible experimental observation of training and inference electricity consumption.

The environmental cost of the various equipment and processes used to operate a Gen-AI service can be evaluated thanks to LCA results for standard equipment such as smartphones, laptops, and server components, as well as impact factors linked to flows such as electricity consumption or online data transfer.

The impact is observed through 3 criteria, selected for their availability, quality, and relevance, considering the main impact categories known for digital services [2]. The first is Abiotic Depletion Potential (ADP) for minerals and metals. It represents the decrease in available resources that have limited reserves. The second, Global Warming Potential (GWP), evaluates the contribution to climate change. The third, Primary Energy (PE), expresses the total energy footprint.

Following the structure described in Fig. 1 and Table 1, this section presents how we evaluate the environmental impact of each service part we consider. The block separation was chosen for clarity and flexibility so that different Functional Units (FU) can be evaluated.

3.1. Access equipment - End-user terminals

End-user terminals are defined as the set of devices directly used by the user to satisfy the FU. It can be a smartphone, a screen and desktop computer setup, or another ICT equipment used to achieve the FU. each piece of equipment e, according to its share of use s_e required to achieve the FU, we calculate its footprint due to electricity consumption and that relative to the rest of its life cycle. The electricity consumption C_e is the average of its lifetime consumption for each type of equipment (laptop, smartphone, desktop computer, etc.). The use phase impact is the relation between the electricity consumption and the geographic area g where it is used. The function EGM_g represents the impacts of the local electricity mix. The rest of the life cycle impact F_e is the life cycle inventory(LCI) data of the equipment. All phase impacts then pass through the same allocation factor $a_e(t)$. A time-based allocation was chosen, cal-

Table 1. Definition of variables used for the different equations

Common Notations	
I	Impact
e	Equipment
C_e	Average consumption of electricity for e in its lifetime
F_e	Footprint for e: manufacture, transport, and end of life
s_e	Share of client using e for the FU
EGM_g	Electricity grid mix impact in geographic area g
PUE	Power usage effectiveness of Data Center
$a_e(t)$	Allocation for e's time of use t for the entire duration of its use (i.e. lifespan times percentage of active use)
End-user terminal notations	
t	Time of use of the equipment
Network notations	
n	Network type, mobile or fixed-line
d	Volume of data transferred through the network
IF_n	Impact factor of transferring data through n
Web hosting Notations	
v	Number of visit
$a_e(v)$	Allocation for e's number of visits for the total of visits
Model inference notations	
i	Inference done on a GPU
$C_{i,e}$	Consumption of electricity for i with e
Training model notations	
tr	Training of the AI model
$C_{tr,e}$	Consumption of electricity for tr with e

culated from the time of use of the equipment divided by its total duration of use. The total use life of the equipment is defined as the lifetime times its Active Utilization Rate (AUR). The AUR is the percentage of time in the equipment lifespan during which it is actively used, as opposed to when it is either idle or on standby. The notion of active utilization rate is discussed in Section 4.4.

$$I_{EndUser} = \sum_{e} s_e \times a_e(t) \times (F_e + C_e \times EGM_g) \tag{1}$$

3.2. Access equipment - Networks

The impact of the networks depends on the total volume of transferred data d required to produce the FU, and on the types of networks used n. The average impact factor IF is assumed to be derived from the impact of transferring one gigabyte through a fixed-line or mobile network. As before, s_n averages the impact according to the type of network used, e.g. European fixed-line network or French mobile network. IF_n takes into account the end-user's router.

$$I_{Networks} = \sum_{n} s_n \times d \times IF_n \tag{2}$$

3.3. Data center servers - Web hosting

This section focuses on the data center equipment used to access the inference functionality, i.e. the infrastructure receiving and processing user requests, or even providing them with a graphical interface via a website, for example.

The impact of this web hosting is calculated by the sum of each the equipment e usage s_e as in (1). However, here the allocation is no longer on a time basis, but on a usage basis. Equipment of this type is capable of handling several tasks in parallel, exposing a time-based allocation to the risk of over-allocation. Thus, the equipment is assumed to be dedicated solely to the FU, and the allocation is based on the total number of FUs that the equipment participated in, which for this study corresponds to the number v of visits to the web service. The Power Usage Effectiveness (PUE) is added to the electricity consumption to account for the energy efficiency of the data center.

$$I_{WebHosting} = \sum_{e} s_e \times a_e(v) \times (F_e + C_e \times EGM_g \times PUE) \tag{3}$$

3.4. Data center servers - Model Inference

Inference is supported by “on-demand” equipment, in this case a GPU e. Computing an inference i on a device e consumes $C_{i,e}$ electricity and takes t time. The electricity cost is then multiplied by the PUE and EGM_g.

The rest of the impact is calculated with a time-based allocation $a_e(t)$ on the equipment e, assuming that GPUs can not perform tasks in parallel. As for end-user terminals, the allocation is proportional to the total lifespan multiplied by its active utilization rate. It represents the fact that $on-demand$ devices provided by data centers are not always used.

$$I_{Inference} = \sum_{i} C_{i,e} \times EGM_g \times PUE + a_e(t) \times F_e \tag{4}$$

3.5. Data center servers - Model Training

This section focuses on the data center server needed to perform one or several training tr for the AI model. As previously, the time-based allocation is based on the lifespan of the equipment, the duration of the training, and its active utilization rate.

However, a special focus is given to the estimation of electricity consumption. As reproducing the training is too expensive and pure modeling unsatisfying, an in-between solution is required. As training is divided into steps constant in cost, measuring the electricity consumption of a few steps is enough to estimate the total cost of training [14], assuming knowledge of the original training.

Lastly, the training impact is a fixed cost needed to launch the service. When assessing the cost of a service using a model during its lifespan, the whole cost of training is considered. However, when assessing the cost of a single use of a service performing one inference using the model, the training cost needs to be allocated to each inference. An example of this allocation can be found in Section 4.1.

$$I_{Training} = \sum_{tr} C_{tr,e} \times EGM_g \times PUE + a_e(t) \times F_e \tag{5}$$

4. Use case: Stable Diffusion

In this section, the methodology is applied and validated on an AI service based on Stable Diffusion [9], an open-source text-to-image generative deep-learning model. Stable Diffusion was developed by researchers from the CompVis Group at Ludwig Maximilian University of Munich and Runway with a compute donation by Stability AI and training data from non-profit organizations. Stable Diffusion was selected because its model is open-sourced and its successive versions can be downloaded on Hugging Face [15]. The model is also freely available as a service [16] since August 2022. On the main web page, users fill out a prompt by describing the wanted image. In default settings, the service generates 4 pictures. Several versions of the model, created by successive trainings from v1-0 to v1-5, exist. The first model version used by the service at the start in 2022 was the v1-4, replaced by the v1-5 model 2 months later.

4.1. Functional units

Two different functional units (FU) were considered for this use case, following the IUT's recommendation for ICT services [8]. First, a *client* vision of the service, **FU1**, represents the average impact of a person visiting the website and submitting a prompt, generating 4 images. The allocation is calculated by dividing the training cost by the potential total number of inferences generated by this version of the model. This number is the sum of the average number of visitors on the website and the average lifespan of a model, i.e. the number of months the model stays online. Secondly, a *host* vision of the service, **FU2**, considers the cost of the service for one year, covering the activity periods of the v1-4 and v1-5 versions of the model (2 and 10 months) before a new one is proposed on the site.

4.2. Measuring tools and resources

We replicated inference and part of the training on nodes from the Sirius cluster of large-scale experimental platform called Grid'5000 [17]. This cluster was selected because of its similarity with the resources used by its developers for the training and inference of the Stable Diffusion model. Sirius is an Nvidia DGX A100 server (8 Nvidia A100-SXM4-40GB GPUs, 2 AMD EPYC 7742 CPUs). While many methods exist to estimate or measure the electricity consumption of a workload [18], we relied on hardware power meters from the Omegawatt company. We estimated the footprint for the GPUs based on the data from the Boavizta working group [19]. This estimate is a lower bound based on the methodologies usually applied to CPUs. We use other LCI databases like the ADEME database to estimate the footprint of the rest of the equipment described in Section 3. We also relied on general statistics on ICT and web usage [20] and web traffic measurement tools like Similarweb [21] or HypeStat [22] to obtain data like the monthly number of visits or average visit time of the website. We assume 1 Amazon Web Services c6gd.metal server was enough for the web-hosting part. We had to estimate the AUR since its value is not publicly shared by the cloud providers. We chose to set the training GPUs and the inference GPUs to 80% and

Table 2. Environmental impact of Stable Diffusion for FU1 and FU2

FU	Abiotic Depletion Potential (kgSb eq)	Warming Potential (kgCO² eq)	Primary energy (MJ)
FU1 - Single use of service	$6.72e^{-08}$	$7.84e^{-03}$	$2.02e^{-01}$
FU2 - A year of service	$4.64e^{+00}$	$3.60e^{+05}$	$8.93e^{+06}$

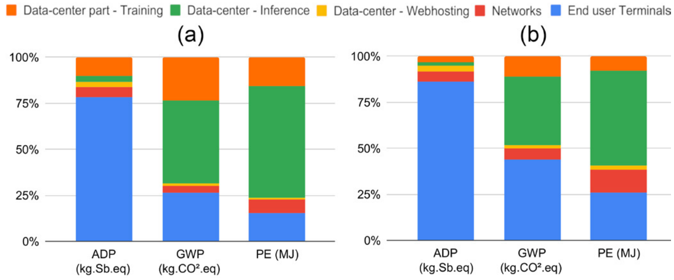

Fig. 2. Impact distributions for (a) FU1 and (b) FU2

40%, respectively. These AURs are high compared to the average AUR [23] but GPU are highly demanded resources, especially for intensive training. The choice is further discussed in Section 4.4.

4.3. Results

We were able to train and infer from the Stable Diffusion model with the same parameters and similar hardware as the original training [15] and the inference done on the website. From measurements of the first training steps and the number of steps provided by the developers of Stable Diffusion, we estimated that the v1-4 and v1-5 models consumed $1.28e^{+04}$ kWh and $3.39e^{+04}$ kWh to train, respectively. We measured that one inference consumed $1.38e^{-03}$ kWh.

Table 2 shows the total impact, for each category, of the Stable Diffusion service, for the FUs described in Section 4.1. We can acknowledge the significant environmental impact of this service for one year of run. With 360 tons of carbon equivalent emission, an impact on metal scarcity equivalent to the production of 5659 smartphones, and an energy footprint of 2.48 Gigawatt hours, it is clear that the impact of Gen-AI should be a matter of concern and not only for its carbon footprint.

Fig. 2(a) and Fig. 2(b) highlight the share of each part of the service in the total impact for the FU1 and the FU2. It can be noticed that the end-user terminals represent most of the impact in ADP, which was expected since such devices usually contain battery and screen which has a significant cost in manufacturing. The two other impact categories (GWP and PE) are dominated by the inference cost, which is coherent with reports from AI companies [7]. Another interesting observation is that the impacts of networks and end-user terminals are not negligible, which validates the need to include them in such evaluations.

The share of the training cost is decreasing from FU1 to FU2 since FU1 used the v1-5 model version and FU2 includes the training cost of two model versions (v1-4 and v1-5) where v1-4 consumed less electricity. These versions required more than their training for their development, potentially hiding an additional cost that we explore in sensitivity analysis.

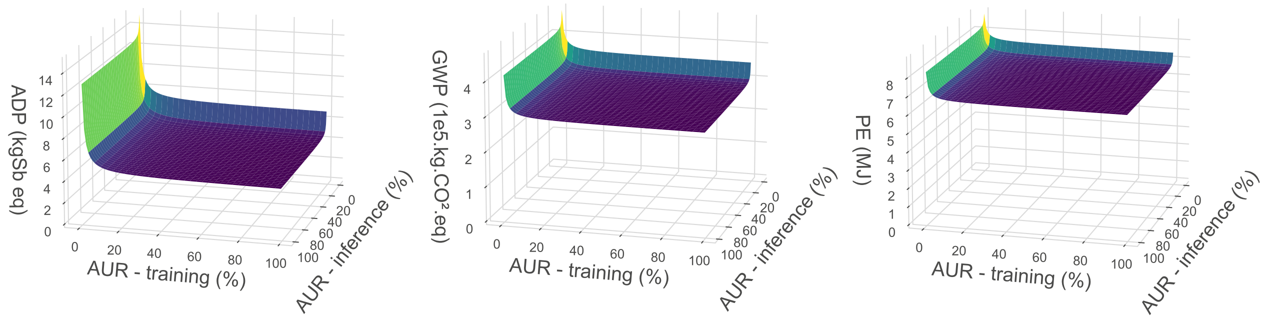

Fig. 3. Impact of the average active utilization rate (AUR) of data-center equipment

4.4. Sensitivity analysis

This sensitivity analysis explores the impact of parameters specific to the service LCA approach or to Gen-IA. We do not treat trivial cases, such as the migration to areas with a more favorable electricity mix or efficient *PUE*.

Firstly, let's focus on the average utilization rate of data center equipment. This parameter is crucial to the allocation of the footprint of the cost-heavy equipment needed for inference and training. In Fig. 3, the impact of the average utilization rate for inference and training GPUs can be observed. It can be seen that this impact is small for rates higher than 20%, which was the case for our choices of rate. Below 20%, the impact is more significant for the average utilization rate of training servers, which was expected since training requires 32 nodes with 8 GPUs. Unfortunately, the rare studies available on the subject are not that optimistic [23], pointing between 12% and 18% of the average utilization rate. For such rates, the impact is important, especially on the ADP impact. However, such studies are not recent, and considering the rapid improvement in virtualization and scheduling in data centers, it can be assumed that AURs are today higher than 20% on average.

Secondly, we can re-visit how we account for the training necessary to complete FU2. Two versions of the model, v1-4 and v1-5, were available during the assessed period, and so, in the standard scenario, the training part is the addition of both their training. However, these versions of the model needed the training of previous versions. They were necessary for FU2 to be completed and, as far as we know, were never used as a service on a large scale. Thus, we compare the standard scenario with a full legacy scenario consisting of including the training costs of v1-0, v1-1, and v1-2 but not v1-3, which were not directly necessary to complete the training of v1-4 and v1-5. In Fig. 4 we see that this parameter severely affects the impact on GWP and PE, even at a large case such as FU2. We interpret this as a warning not to neglect the impact of training in Gen-AI-based services.

5. Discussion

The current focus given to training and, more and more, to inference, in addressing the environmental impacts of AI can be seen as a reflex of studying only field-related subjects. Additionally, one could question the usefulness of assessing other

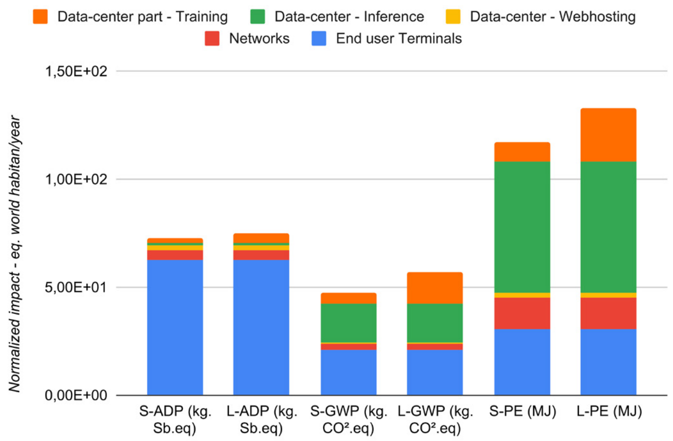

Fig. 4. Comparison between the standard scenario (S) and the full legacy scenario (L) on equivalent world habitant normalized impact for FU2

blocks of AI service development and deployment. For example, the network and end-user terminals could be overlooked because they are not influenced by AI and are not potential levers to reduce the total footprint. This study shows the relevance of evaluating the impact of Gen-AI as a service. The service approach shows that these parts have a significant impact but also gives some indication of the actual use intensity of inferences. Also, the impact of end-user terminals represents how Gen-AI contributes to the growth of IT usage and its unfortunately still growing footprint.

We chose an attributional approach for this assessment. However, a consequentialist approach would highlight critical subjects. The transition from classic digital services to Gen-AI-based services could severely increase the footprint of the ICT sector. Even more so if the claimed productivity gains [24] end up with a rebound effect [25]. This transition to ubiquitous AI services can lead to a growth in data center infrastructures. The unprecedented rise in the use of GPUs has the potential to increase the footprint of data centers, as GPUs have heavier environmental costs than CPUs. Additionally, the difficulties in mutualizing GPUs as well as other resources like CPUs or RAM will reduce the utilization rate of these types of equipment with the consequences detailed in Section 4.4.

This method is generalizable to other Gen-AI services. It represents the cost of access and use of the Gen-AI model and as we applied it to a *text-to-image* model, we will apply it to *text-to-text* model in future work.

6. Conclusion

This paper presents an LCA-based methodology to assess the direct environmental impact of digital services using Gen-AI. It details how to calculate these impacts for the different parts of the service on a multi-criteria basis, taking into account not only the inference and training cost but the service as a whole. We then apply our methodology to a popular Gen-AI available online as a service. Estimated impact values are significant. Supporting one year of Stable Diffusion service for the observed number of users can generate as much as 360 tons of CO^2 equivalent.

This application of the proposed methodology demonstrates not only its feasibility but also its value. We have broadened the scope of the study to better understand the environmental cost of these services. Indeed, like other digital services, Gen-AI services generate, in addition to their carbon footprint, a significant impact on mining and energy production. These impacts could be overlooked if focusing on carbon emissions, especially if we fail to include every equipment linked to the use of these services. User equipment, networks, and web servers, all essential to the existence of Gen-AI-based services, are in this use-case responsible for at least 30% of the environmental impact, and more than 90% in the case of depletion of mineral and metal resources.

Improvements can be made to this use case to consolidate the results. This applies in particular to GPU footprints, web server sizing, and server utilization rates, for which better transparency from both the equipment manufacturing industries and the hosting providers offering these services can be expected. However, we believe that this method can provide a low estimate of the potential cost of these services and their deployment. We invite researchers developing such models to apply this methodology to estimate the potential environmental impact of deploying such services to millions of clients. Transparency of these impacts can only contribute to the emergence of fair and ethical AI as well as awareness of the real cost of these technologies for our environment.

Acknowledgements

This work was funded by ANRT (CIFRE N° 2021/0576), MIAI (ANR19-P3IA-0003), and the BATE project (BATE-UGAREG21A87) of the Auvergne Rhône-Alpes French region. The authors used data from the open-source project Boavizta. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER, and several universities as well as other organizations.

References

[1] Alex de Vries. The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194, 2023.

[2] Bordage, F., de Montenay, L., Benqassem, S., Delmas-Orgelet, J., Domon, F., Prunel, D., Vateau, C. et Lees Perasso, E. Digital technologies in europe: an environmental life cycle approach.

[3] Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability*, 14:5172, 2022.

[4] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of BLOOM, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15, 2023.

[5] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019. arXiv: 1906.02243.

[6] Roberto Verdecchia, June Sallou, and Luís Cruz. A systematic review of green AI. *WIREs Data Mining and Knowledge Discovery*, 13(4):e1507, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1507.

[7] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, and Kim Hazelwood. *Sustainable AI: Environmental Implications, Challenges and Opportunities*. 2021.

[8] ITU. ITU l1410 : Methodology for environmental life cycle assessments of information and communication technology goods, networks and services.

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[10] Stuart J. Russell, Peter Norvig, and Ernest Davis. *Artificial intelligence : a modern approach*. Prentice Hall series in artificial intelligence. Prentice Hall Upper Saddle River, New Jersey, third edition edition, 2010.

[11] Francisco García-Peñalvo and Andrea Vázquez-Ingelmo. What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8:7–16, 08 2023.

[12] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

[13] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the Carbon Emissions of Machine Learning. Technical Report arXiv:1910.09700, arXiv, November 2019. arXiv:1910.09700.

[14] Lucia Bouza Heguerte, Aurélie Bugeau, and Loïc Lannelongue. How to estimate carbon footprint when training deep learning models? a guide and review. *Environmental Research Communications*, 2023.

[15] HF. Hugging Face runwayml stable diffusion repository. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2016.

[16] SD. Stable diffusion service. https://stablediffusionweb.com/#demo, 2022.

[17] Franck Cappello, Eddy Caron, Michel Dayde, Frederic Desprez, Emmanuel Jeannot, Yvon Jegou, Stephane Lanteri, Julien Leduc, Nouredine Melab, Guillaume Mornet, Raymond Namyst, Pascale Primet, and Olivier Richard. Grid'5000: a large scale, reconfigurable, controlable and monitorable Grid platform. In *SC'05: Proc. The 6th IEEE/ACM International Workshop on Grid Computing Grid'2005*, pages 99–106, Seattle, USA, November 2005. IEEE/ACM. hal number inria-00000284.

[18] Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel. An experimental comparison of software-based power meters: focus on CPU and GPU. In *CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing*, page 1. IEEE, 2023.

[19] B. Boavizta. https://boavizta.org/en.

[20] Simon Kemp. Digital 2023: Global overview report. https://datareportal.com/reports/digital-2023-global-overview-report.

[21] S. Similarweb. https://www.similarweb.com/.

[22] H. Hypestat. https://hypestat.com/.

[23] Josh Whitney Pierre Delforge. America's data centers consuming and wasting growing amounts of energy. Technical report, NRDC, 2015.

[24] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. National Bureau of Economic Research, April 2023.

[25] Vlad C. Coroamă and Friedemann Mattern. Digital rebound - why digitalization will not redeem us our environmental sins. In *ICT for Sustainability*, page 10, 2019.