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

Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefèvre

ESTIMATING THE ENVIRONMENTAL IMPACT OF GENERATIVE-AI SERVICES USING AN LCA-BASED METHODOLOGY *

Adrien Berthelot
ENS de Lyon, OCTO Technology, Inria
Lyon, France
{adrien.berthelot@ens-lyon.fr

Mathilde Jay
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG
Grenoble, France
mathilde.jay@univ-grenoble-alpes.fr

Laurent Lefevre
Univ. Lyon, EnsL, UCBL, CNRS, Inria, LIP
Lyon, France
laurent.lefevre@inria.fr

Eddy Caron
Univ. Lyon1 (ISFA), Inria, CNRS, LIP
Lyon, France
Eddy.Caron@ens-lyon.fr

ABSTRACT

Generative AI (Gen-AI) represents a major growth potential for the digital industry, a new stage in digital transformation through its many applications. Unfortunately, by accelerating the growth of digital technology, Gen-AI is contributing to the significant and multiple environmental damage caused by its sector. The question of the sustainability of IT must include this new technology and its applications, by measuring its environmental impact. To best respond to this challenge, we propose various ways of improving the measurement of Gen-AI's environmental impact. Whether using life-cycle analysis methods or direct measurement experiments, we illustrate our methods by studying Stable Diffusion a Gen-AI image generation available as a service. By calculating the full environmental costs of this Gen-AI service from end to end, we broaden our view of the impact of these technologies. We show that Gen-AI, as a service, generates an impact through the use of numerous user terminals and networks. We also show that decarbonizing the sources of electricity for these services will not be enough to solve the problem of their sustainability, due to their consumption of energy and rare metals. This consumption will inevitably raise the question of feasibility in a world of finite resources. We therefore propose our methodology as a means of measuring the impact of Gen-AI in advance. Such estimates will provide valuable data for discussing the sustainability or otherwise of Gen-AI solutions in a more transparent and comprehensive way. We intend to help this discussion by differentiating in our approach between the embodied and operational impacts of Gen-AI. In this way, we can consider the sustainability of models, as we already do for equipment.

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 [22], 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 [14, 15, 20, 21, 23] 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 [21], and possibly deducing their contribution to global

*This article is a preprint and has not been certified by peer review

warming via the carbon intensity of the electricity mix used. By neglecting the conditions and resources required for 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]. To fill these gaps, we propose a methodology based on life cycle assessment (LCA) and observe the use of these models as digital services. Our 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.

Section 2 defines the principle of AI service and provides a review of existing literature. Section 3 describes our methodology: which equipment we include and how we account for their embedded and electricity cost. In Section 4, we illustrate our methodology with Stable Diffusion [17], an open-source text-to-image generative deep-learning model released in 2022. We analyze their cost for a single use and for one year of service. Complementary Sensitivity Analyses (SA) highlight the critical parameters influencing the environmental impact of Gen-AI. In Section 5, we question the hypothesis we made, show the limitations of our methodology, and what we would need 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. 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. 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 up to the point where Central Processing Units (CPUs) are no longer sufficient. Graphic Processing Units (GPUs) are today the most common processing unit for AI workloads (training and inference).

2.2 Related work

Interest in the energy consumption of AI started in 2019 with the work of Strubell et al. [20]. Methodologies on how to measure or estimate it have been explored in several approaches [8, 13, 14]. However, they solely focus on training. A more recent study of the BLOOM model [15] 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.* [23] explore the environmental impact of AI in various phases (Data, Experimentation, Training, and Inference). They 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, we propose an evaluation of AI as a service, as described in the next section.

2.3 AI as a Service

We have chosen to focus on AI as a service, which means that we integrated not only the specific costs of AI, i.e. training and inference phases, but also all the equipment and infrastructure required to use it online, as a service. We assume that the AI is accessible through a website interface. As shown on Figure 1, we consider a digital service to be the sum of end-user terminals, networking, web hosting, model inference, model training, and data management.

3 Methodology

We propose a methodology based on LCA and on a reproducible experimental observation of training and inference electricity consumption. To account for the environmental cost of the various equipment and processes used to operate the service, we use 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.

We observe the environmental cost through 3 criteria, selected for their availability, quality, and relevance, considering the main impact 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.

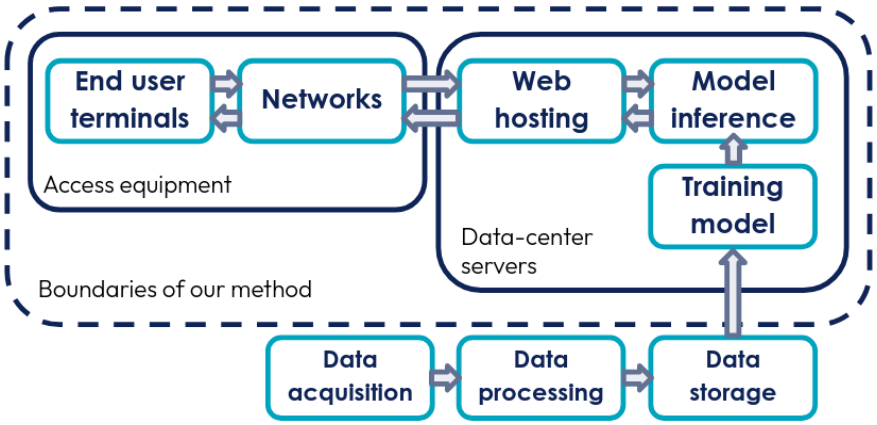

Figure 1: Structure of Gen-AI service considered by our methodology

Following the structure described in Figure 1 and Table 1, we now present how we evaluate the environmental impact of each service part we consider. We have chosen this block separation for the purpose of clarity, but also to adapt our method more easily to different types of Functional Units (FU).

3.1 Access equipment - End user terminals

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

Table 1: Definition of variables used for the different equations

We consider end-user terminals as the device or set of devices directly used by the user to satisfy the FU. It can be a smartphone or a screen and desktop computer setup. We only consider the ICT equipment. We take into account all

the user terminals used to achieve the FU. For 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 $a_e(t)$. We use a time-based allocation, calculated from the time of use of the equipment divided by its total duration of use. We consider the total use life of the equipment to be the lifetime times its Active Utilization Rate (AUR). The active utilization rate 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. We discuss more of the notion of active utilization rate 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. We assume that we have an average impact factor IF derived from the impact of transferring one gigabyte through a fixed-line or mobile network. As before, we use p_n to average the impact according to the type of network used, e.g. European fixed-line network or French mobile network. We assume that the IF_n takes into account end-users's router.

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

3.3 Data center servers - Web hosting

In this section, we look at the data center equipment used to access the inference function. This means the ability to receive and process user requests, or even to provide them with a graphical interface via a website, for example.

We calculate the impact of this part from the sum of the equipment used e in relation to its 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, we assume that this equipment is dedicated solely to the FU, and base allocation on the total number of FUs that the equipment participated to achieve. We count this as the number v of visits to the web service. We also add to the electricity consumption the Power Usage Effectiveness (PUE) parameter showing 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

To model the computations needed for inference, we consider that inference is supported by an "on-demand" equipment, 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. We use this allocation 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 - Training model

We consider here the data-center server setup needed to perform one or several training tr for the AI model. As previously, we use a time-based allocation based on the lifespan of the equipment, the duration of the training, and its active utilization rate.

However, we give a special focus to the estimation of electricity consumption. As reproducing the training is too expensive and pure modeling unsatisfying, we advise for an in-between solution. Training is divided into steps constant in cost, thus, measuring the electricity consumption of a few steps can be enough to estimate the total cost of training [3], assuming knowledge of the original training. We use this approach in our use case.

Lastly, the training impact is often a fixed cost needed to launch the service. When assessing the cost of a service using a model during its lifespan, we consider the whole cost of training. But when we assess the cost of a single use of a service performing one inference using the model, we need to allocate part of the training cost to this inference. We show an example of this allocation 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

To apply and validate our methodology, we have selected an AI service based on Stable Diffusion [17], 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. We selected Stable Diffusion because its model is open-sourced and its successive versions can be downloaded on Hugging Face [9]. The model is also freely available as a service [19] 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

We considered and applied 2 different functional units (FU) to our use case following the IUT's recommendation for ICT services [10].

First, we represent a $client$ vision of the service. **FU1**, represents the average impact of a person visiting the website and submitting a prompt, generating 4 images. For each inference, needing the trained model, we perform an allocation on the training cost. We calculate the allocation for one divided by the potential total number of inferences generated by this version of the model. We estimated it considering 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, we represent 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 respectively) 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 test beds for experimental research called Grid'5000 [5]. 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 in order to estimate or measure the electricity consumption of a workload [11], we relied on power meters from the Omegawatt company. We estimated the footprint for the GPUs based on the data from the Boavizta working group [1]. 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 use general statistics on ICT and web usage [12] and web traffic measurement tools like Similarweb [18] or HypeStat [7] to obtain data like the monthly number of visits or average visit time of the website. We had to estimate the active utilization rate 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 40%, respectively.

4.3 Results

We were able to train and infer from the Stable Diffusion model with the same parameters as the original training and the inference done on the website. We measured that one inference consumed $1.38e^{-03}$ kWh. 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.

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

FU	Abiotic Depletion Potential (kgSb eq)	Warming Potential ($kgCO^2$ eq)	Primary energy (MJ)
FU1 - Single use of service	$6.72e^{-08}$	$7.84e^{-03}$	$2.02e^{-01}$
FU2 - A year of service	$4.64e^{+00}$	$3.60e^{+05}$	$8.93e^{+06}$

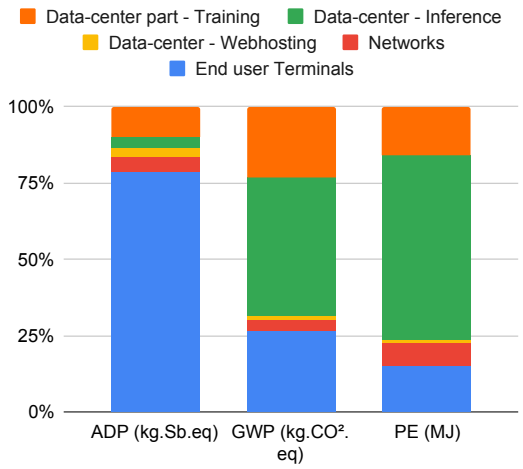

Figure 2: Impact distribution for FU1

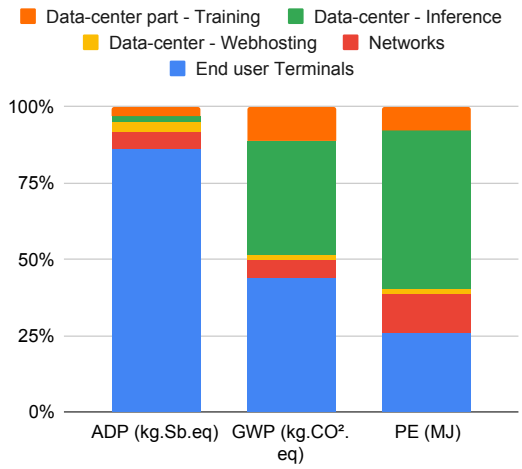

Figure 3: Impact distribution for FU2

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.

Figure 2 and Figure 3 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 [23]. 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 own training for their development, potentially hiding an additional cost that we explore in sensitivity analysis.

4.4 Sensitivity analysis

We would like to explore in our SA the impact of parameters specific to our service LCA approach or to Gen-IA. We don't dwell on trivial cases that have already been dealt with, such as the migration to areas with a more favorable electricity mix for their data centers.

Firstly, let's look at 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 Figure 5, we observe the impact of the average utilization rate for inference and training GPUs. We can see 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. However, it is hard to make an assumption about these parameters and the rare studies available on the subject are not that optimistic [16], pointing between 12% and 18% of the average utilization rate. For such rates, the impact is important, especially on the ADP impact.

Secondly, we can re-visit our SA with 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. But these versions of the model needed the training of previous versions like v1-0, v1-1 or v1-2. They were clearly necessary for FU2 to be complete and, as far as we know, were never used as a service at a large scale. We left this question open and decided to include in an SA the training cost 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 Figure 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. A vast hidden cost could be underestimated.

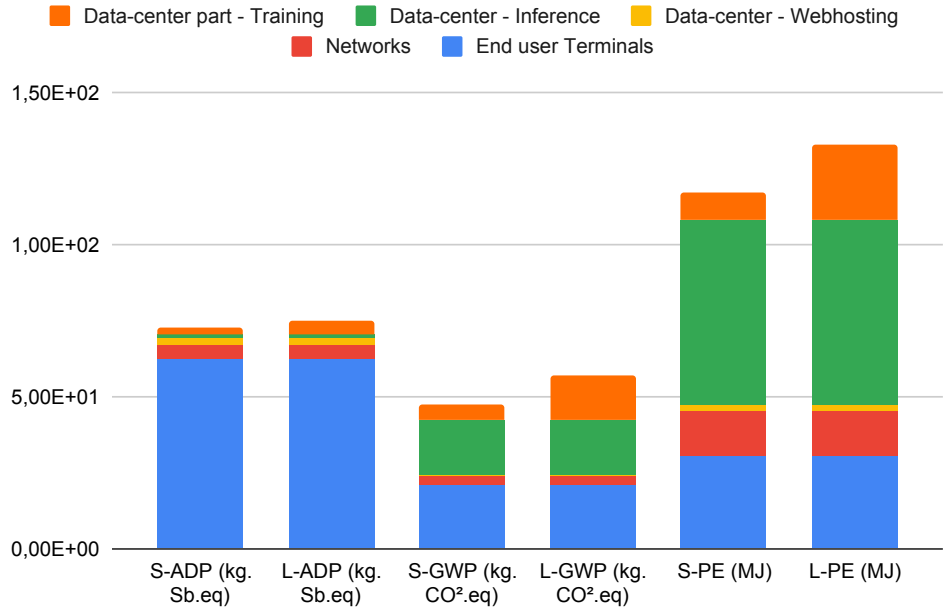

Figure 4: FU2 - comparison between S, standard scenario, and L, full legacy scenario, - normalized impact equivalent world habitant

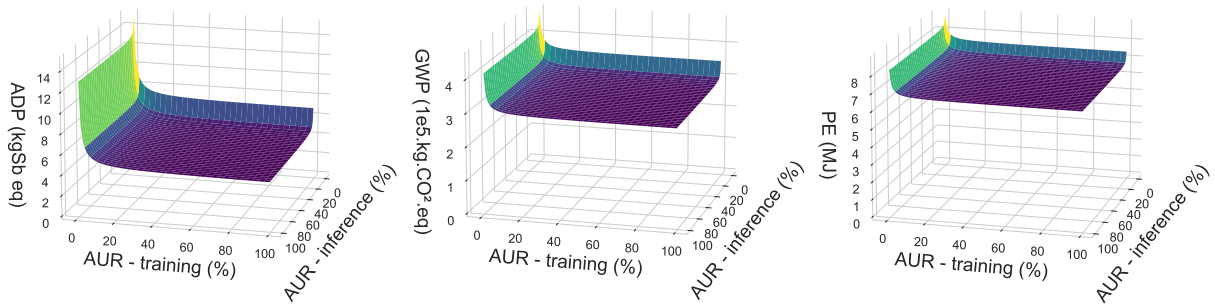

Figure 5: Impact of the average active utilization rate (AUR) of data-center equipment

5 Discussion

The current focus given to training and now, more and more, to inference, in addressing the environmental impacts of AI can be seen as a reflex of studying only field-related subjects. But there is also an approach neglecting the utility to assess the other part of an AI-based. For example, the network and end-user terminals should not be interesting because they are not influenced by AI and are not potential levers to reduce footprint. This study shows the relevance of evaluating the impact of Gen-AI as a service. This service approach shows that these parts have a significant impact but also gives some indication about the workload 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 conduct an assessment from an attributional point of view as we are, to our current knowledge, the first to perform such multi-criteria LCA of Gen-AI as a service. However, a consequentialist approach would highlight critical subjects. The transition from classic digital services to Gen-AI-based services would severely increase the footprint of the ICT sector. Moreover, if the claimed productivity gains [4] end up with a rebound effect [6]. Also, this transition to more AI requires an increase in data center infrastructures. The unprecedented massive use of GPUs would not only directly increase the footprint of data centers, as GPUs have heavier environmental costs than CPUs, but 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 we saw in our SA.

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

In this paper, we present an LCA-based methodology to assess the direct environmental impact of digital services using Gen-AI. We detail 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 impressive. 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 our methodology demonstrates not only its feasibility but also its value. We have broadened the scope of our studies to better answer questions about the environmental cost of these services. Indeed, like other digital services, these services generate, in addition to their carbon footprint, a significant impact on mining and energy production. These impacts could be overlooked by focusing too much on the carbon issue, especially if we forget the other sources of impact directly 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 our 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 in our use case application to consolidate the results. This applies in particular to GPU footprints, web server sizing, and server utilization rates, where we expect greater transparency from both the equipment manufacturing industries and the hosting providers offering these services. However, we believe that our 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

The authors used data from the open-source project Boavizta (`https://boavizta.org/en`). Experiments presented in this article 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 (`https://www.grid5000.fr/`). This work was funded by ANRT (CIFRE N° 2021/0576), MIAI (ANR19-P3IA-0003), and by the BATE project (BATE-UGAREG21A87) of the Auvergne Rhône-Alpes french region.

References

[1] B. *Boavizta*. URL: `https://boavizta.org/en` (visited on 09/29/2023).

[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*. Green IT. Dec. 7, 2021. URL: https://www.greenit.fr/wp-content/uploads/2021/12/EU-Study-LCA-7-DEC-EN.pdf (visited on 12/08/2021).

[3] Lucia Bouza Heguerte, Aurélie Bugeau, and Loïc Lannelongue. "How to estimate carbon footprint when training deep learning models? A guide and review". In: *Environmental Research Communications* (2023). ISSN: 2515-7620. DOI: 10.1088/2515-7620/acf81b. URL: http://iopscience.iop.org/article/10.1088/2515-7620/acf81b (visited on 10/23/2023).

[4] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. *Generative AI at Work*. Working Paper 31161. National Bureau of Economic Research, Apr. 2023. DOI: 10.3386/w31161. URL: http://www.nber.org/papers/w31161.

[5] Franck Cappello et al. "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*. hal number inria-00000284. IEEE/ACM. Seattle, USA, Nov. 2005, pp. 99–106. URL: https://hal.inria.fr/inria-00000284.

[6] Vlad C. Coroamă and Friedemann Mattern. "Digital Rebound - Why Digitalization Will not Redeem us our Environmental Sins". In: *ICT for Sustainability*. 2019, p. 10. URL: https://api.semanticscholar.org/CorpusID:140117969.

[7] H. *HypeStat*. URL: https://hypestat.com/ (visited on 10/29/2023).

[8] Peter Henderson et al. "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning". In: *Journal of Machine Learning Research* 21.248 (2020), pp. 1–43. ISSN: 1533-7928. URL: http://jmlr.org/papers/v21/20-312.html (visited on 10/30/2023).

[9] HF. *Hugging Face Runwayml Stable Diffusion repository*. 2016. URL: https://huggingface.co/runwayml/stable-diffusion-v1-5 (visited on 09/29/2023).

[10] ITU. *ITU L1410 : Methodology for environmental life cycle assessments of information and communication technology goods, networks and services*. Dec. 2014. URL: https://www.itu.int/rec/dologin_pub.asp?lang=f&id=T-REC-L.1410-201412-I!!PDF-E&type=items.

[11] Mathilde Jay et al. "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*. CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing. IEEE, May 1, 2023, p. 1. URL: https://hal.inria.fr/hal-04030223 (visited on 03/27/2023).

[12] Simon Kemp. *Digital 2023: Global Overview Report*. DataReportal – Global Digital Insights. Jan. 26, 2023. URL: https://datareportal.com/reports/digital-2023-global-overview-report (visited on 10/23/2023).

[13] Alexandre Lacoste et al. *Quantifying the Carbon Emissions of Machine Learning*. Tech. rep. arXiv:1910.09700. arXiv:1910.09700 [cs] type: article. arXiv, Nov. 2019. URL: http://arxiv.org/abs/1910.09700 (visited on 06/13/2022).

[14] Anne-Laure Ligozat et al. "Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions". In: *Sustainability* 14 (Apr. 25, 2022), p. 5172. DOI: 10.3390/su14095172.

[15] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model". In: *Journal of Machine Learning Research* 24.253 (2023), pp. 1–15. ISSN: 1533-7928. URL: http://jmlr.org/papers/v24/23-0069.html (visited on 09/28/2023).

[16] Josh Whitney Pierre Delforge. *America's Data Centers Consuming and Wasting Growing Amounts of Energy*. Tech. rep. NRDC, Feb. 6, 2015. URL: https://www.nrdc.org/resources/americas-data-centers-consuming-and-wasting-growing-amounts-energy (visited on 10/12/2023).

[17] Robin Rombach et al. "High-Resolution Image Synthesis With Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10684–10695.

[18] S. *Similarweb*. URL: https://www.similarweb.com/ (visited on 10/29/2023).

[19] SD. *Stable Diffusion service*. 2022. URL: https://stablediffusionweb.com/#demo (visited on 09/09/2023).

[20] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *arXiv:1906.02243 [cs]* (June 2019). arXiv: 1906.02243. URL: http://arxiv.org/abs/1906.02243 (visited on 11/02/2021).

[21] Roberto Verdecchia, June Sallou, and Luís Cruz. "A systematic review of Green AI". In: *WIREs Data Mining and Knowledge Discovery* 13.4 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1507, e1507. ISSN: 1942-4795. DOI: 10.1002/widm.1507. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1507 (visited on 10/25/2023).

[22] Alex de Vries. "The growing energy footprint of artificial intelligence". In: *Joule* 7.10 (Oct. 18, 2023), pp. 2191–2194. ISSN: 2542-4351. DOI: 10.1016/j.joule.2023.09.004. URL: https://www.sciencedirect.com/science/article/pii/S2542435123003653 (visited on 10/24/2023).

[23] Carole-Jean Wu et al. *Sustainable AI: Environmental Implications, Challenges and Opportunities*. Oct. 30, 2021.