Understanding the environmental impact of generative AI services

Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefevre

Understanding the environmental impact of generative AI services

Adrien Berthelot[1,2], Eddy Caron[1], Mathilde Jay[3] and Laurent Lefèvre[1*]

[1]École Normale Supérieur de Lyon, Université Lyon 1, CNRS, Inria, LIP
[2]Octo Technology
[3]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG

Abstract

Generative AI (Gen-AI) represents a new stage in digital transformation through its many applications. Unfortunately, by accelerating the growth of digital technology, Gen-AI is contributing to the multiple environmental damages caused by its sector. The question of the sustainability of IT must include this new technology and its applications, by estimating its environmental impact. 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 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. Our approach differentiates the embodied and operational impacts of Gen-AI in order to consider the sustainability of models and equipment. Such solution will provide valuable data for discussing the sustainability or otherwise of Gen-AI solutions in a more transparent and comprehensive way.

1 Introduction

The last few decades have been marked by the ever-increasing presence of digital technology in our societies. This growth, presented as digital transformation, comes naturally with the increasing weight of digital technology on our environment. We are now facing a potential new phase of digital transformation [1], represented by the emergence of generative AI (Gen-AI), a subfield of artificial intelligence where the objective is to generate new content, for example, human-like discussions and realistic images [2].

While we can hope that certain digital applications will help to meet sustainability challenges by reducing the impact of human activities on the environment, it remains difficult to measure the positive or negative environmental impact of digital technology [3, 4, 5]. The question of the sustainability of computing can not be addressed scientifically without means of evaluating the environmental damage. Moreover, we need to evaluate the real applications of computing and not just subparts. As it is, the deployment of Gen-AI as a service, available online from any device, like Chat-GPT or Stable-Diffusion, that are raising many questions, even beyond sustainability. This is why we present a methodology with its application, that makes an environmental assessment of Stable Diffusion as an end-to-end service. To better question the sustainability of Gen-AI, we assess not only the carbon impact but also the consumption of metals, in a life cycle assessment (LCA) approach. Finally, we also question the current methods used to estimate the electricity consumption used for training Gen-AI models. We propose an approach based on sampling through reproducible experiments. Transparency and reproducibility are necessary as best as possible, especially in the subject of environmental sustainability. We present in this article measurement tools and methods aiming at being more reliable and holistic to allow us to rethink sustainability challenges and improvements in the use of Gen-AI services. We base our impact methodology on a previous work [6], extending the scope to data storage, all training costs, and differentiating operational emissions, from the use phase of the hardware life-cycle, and embodied emissions, from the other phases.

We begin with a review of the current approach to environmental impact in the field of AI and Gen-AI (Section 2). Then, we present our tool to enhance the current way to assess the environmental impact of

*Cite as: Adrien Berthelot, Eddy Caron, Mathilde Jay, and Laurent Lefevre. 2025. Understanding the environmental impact of generative AI services. Communications of the ACM (CACM), Special Issue on Sustainability and Computing . ACM, New York, NY, USA, 14 pages.

Gen-AI (Section 3). Finally, we show how our contribution are helping to frame new obstacles and sustainability challenges around Gen-AI services (Section 4).

2 Overview and limits of environmental impact assessment for AI

2.1 Rise of generative AI, quick review

If, as [2] points out, the term "Generative AI" was highlighted in 2014 [7], it is since the end of 2022 that this term has been gaining notoriety and consequently great interest, well beyond AI research. However, the sudden and significant popularization of the notion of generative AI should not conceal the long research process these models are part of [8].

Developing Gen-AI models requires collecting data and learning from the data, which includes (1) selecting the best model structure and learning algorithm for the given task and (2) applying this algorithm to the model and the collected data. The first step hides an expensive development process, as Gen-AI models are usually composed of several already-developed models. The second step is called training. Once a model has reached the targeted quality, it can be used on new data, which is referred to as the inference phase. Due to the ever-increasing size of databases and models, training can require hundreds of Graphic Processing Units (GPUs) running in parallel. As an example, Stable Diffusion was trained with 258 GPUs and 64 Central Processing Units (CPUs) [9]. Data is grouped into batches so that each batch is processed in parallel. Training the model on one batch corresponds to a training step. Hundreds of thousands of steps were needed to train Stable Diffusion.

It is interesting to note that, while there were already powerful models in existence before 2022, it is the online availability of these models as a service, led by Chat-GPT, that is behind the popularity of Gen-AI. The massive new use of such services and the IT infrastructures that support them, with their high demand for electricity [10] and critical equipment [11], undoubtedly raises the question of sustainability. What is the environmental impact of this new digital usage, and in what order of magnitude?

2.2 Environmental impact of AI, an uncompleted work

Included in global studies on the environmental impact of digital technology, the growth in the size of machine learning (ML) models [12] has brought AI to the fore as a sustainability issue in its own right in 2019 [13, 14]. These initial studies focus on electricity consumption during the ML training phase as well as the greenhouse gas (GHG) emissions associated with this consumption. Building on this momentum, several other studies followed [15, 16, 17, 18], without departing from the scope of the measurement of electricity consumption for training ML models. This choice can be partly explained by the fact that ML models are seen as research projects rather than mass-market consumer products.

Several methods have been developed to measure or estimate the electricity consumption of the training phase. The most popular one is based on a manufacturing constant called Thermal Design Power (TDP), which is a good estimation of the maximal power of a component. The training phase is usually highly intensive, so it can be assumed that the computing components are running at maximal power. Therefore, a TDP-based estimation is the multiplication of the TDP of the computing components by the total training duration. This method is quite simple, as it only requires knowledge of the TDP and the training duration, but it has several limitations. First, the computing components (which are typically GPUs) are not the only components composing a server. CPUs, memory buses, switches, and fans should also be taken into account.

An alternative method is to measure the electricity consumption during the execution. Power meters placed outside of the computing node are the most accurate and complete tools, but they also require direct access to the hardware, which is, in most cases, difficult if not impossible. Workloads can also be monitored using software-based power meters (e.g. RAPL or NVML), which report the electricity consumption of a selection of components [19]. Many tools were developed based on these reports. These libraries are more accurate than a TDP-based estimation but have the same limitation in the completeness of the component taken into account. The main drawback of power meters is the necessity to use them during execution, with no possibility of estimating beforehand or afterhand. The literature lacks a methodology that is accessible, reproducible, reliable, and accurate for Gen-AI.

The popularization of Gen-AI has broadened the scope of such methodologies, encompassing more direct impacts. As a direct consequence of usage, the inferences made from models are studied [**luccioni_power_2023**, 20, 21, 10], even if, similarly as for the training phase, this often remains solely through the measurement of electricity consumption. Studies beginning to include the full life

cycle of equipment are still rare [22, 23], and are limited to the carbon cost of training and inference. This current perimeter does not fully address Gen-AI sustainability issues. As the digital sector has a large embodied footprint [24] including the full life-cycle impact is essential. The digital sector also has a strong environmental impact that goes beyond GHG emissions [25], e.g. the extraction of rare metals. Finally, Gen-AI relies on and operates through the existing digital ecosystem, on which it logically exerts pressure. It requires terminals for its users, i.e. smartphones and computers, as well as internet networks to be accessible from its data centers. All these resources are essential to the deployment of Gen-AI as it exists today, and are part of the AI sustainability issue. It is therefore expected and necessary to continue extending the scope of study on Gen-IA as an accessible service with multiple environmental impacts, to better discuss its sustainability.

2.3 LCA as an emerging tool for sustainability of computing

LCA is a multi-criteria evaluation method based on the ISO 14040[26] and 14044[27] standards and can be based on complementary standards depending on the sector studied, such as the ITU (International Telecom Union) L1410 standard for ICT (Information and Communication Technologies) goods, networks, and services [28]. It aims to produce an evaluation of the potential environmental impacts of a product or activity, considering all its life-cycle phases: manufacturing, usage and end-of-life. As pictured in Figure 1, an LCA is composed of four interdependent phases. The goal is to, for a defined purpose and perimeter (step 1), account for all the sub-products and elementary flows needed for the study's subject (step 2), and sum their environmental impacts given by life cycle inventory (LCI) data [29](step 3). Step 4 questions the potential conclusions regarding the initial goals of the study and the uncertainties regarding the hypotheses taken during the previous phases.

Although LCA has only recently been applied in the context of digital services [30, 31] compared to other sectors [32, 33], it is widely recognized for its specific qualities. It enables a more comprehensive assessment by taking into account the complete life cycle and the different impact categories, thus avoiding focusing solely on the carbon emissions of the use phase. Moreover, while its use of assumptions is criticized, it nonetheless enables relevant estimates to be made in a context such as that of digital technology, where the industry's lack of transparency [34, 35] could block the study of its damage to the environment. LCA therefore has the necessary qualities to question the sustainability of IT products and services, since it questions other sectors of activity according to the same standard.

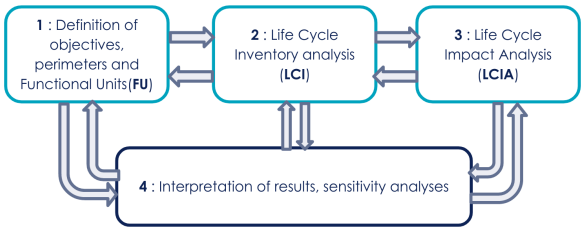

Figure 1: *Four main stages of LCA*

3 Enhanced tools for measuring Gen-AI environmental impact

Even if we consider only the environmental aspect of sustainability, the deployment of a service has many consequences on different levels [36]. In the ICT context, there are frameworks for assessing sustainability [37, 38, 39]. Referring to these frameworks, we propose contributions to better assess the direct environmental impacts of Gen-AI. As an example and validation, we apply our methodology to assess the environmental cost of Stable Diffusion [40], 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 it is popular, its model is open-sourced, and its successive versions can be downloaded on Hugging Face [41].

3.1 Generative AI as a service

To better assess the environmental impact of AI, we propose to study not only the impact of developing a model but also that of its deployment and use as a service. Extending the model of a previous study[6], Figure 2 shows what we consider the standard structure of a Gen-AI service. The arrows represent the data flows between the various parts.

Gen-AI users access the service from a personal terminal, sending a request that is transferred over networks and managed by a web server. Specific computation components are used to infer from a model. The results of the inference process come back to the users through networks. The model has been previously trained during a specific phase, with access to training data in addition to its computational resources. The environmental costs of producing these training data are not taken into account in this work. Although this is a critical issue in the creation of

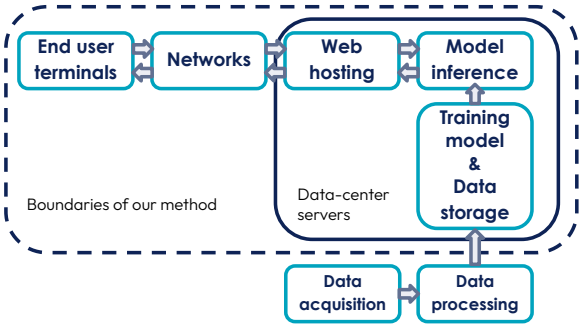

Figure 2: *Structure of a considered Gen-AI service*

Gen-AI services, the process is still too opaque to be analyzed from an environmental perspective.

As an example, the Stable Diffusion model is freely available as a service [42] since August 2022. On the main web page, users fill out a prompt by describing the wanted image.

The activity of a Gen-AI service and the resulting environmental impact was developed so that it is modular and can be adapted to more specific user paths. For example, a service hosted and used on a personal machine could remove the "Networks" and "Web hosting" sections.

Following the recommendations of the industry standard for ICT services [28], we seek to assess in our study the environmental impact of running the service for a full year.

3.2 Estimating the electricity consumption of training through training step replication

We propose a new approach to estimate the electricity consumption of the training phase. Existing methodologies presented in Section 2.2 are either unsatisfactory or require replication of the training, which would be too expensive in the case of Gen-AI model training. Our approach consists of replicating a fraction of the training while monitoring the electricity consumption and doing an estimation of the total training based on those observations. [43] proved that the electricity consumption of epochs is constant. We show that this characteristic can be used to estimate the total training electricity cost by replication, assuming sufficient information from the original training.

In this section, we illustrate our approach to Stable Diffusion. Several versions of the model, created by successive training phases from v1-0 to v1-5, exist. We executed experiments on nodes from the Sirius cluster (Table 1) of the large-scale experimental Grid'5000 platform [44]. This cluster was selected because of its similarity with the resources used by developers for the training and inference of the Stable Diffusion model.

Cluster	Sirius	Gemini
System	Nvidia DGX A100	Nvidia DGX-1
CPU model	AMD EPYC 7742 (Zen 2, 64 cores/CPU)	Intel Xeon E5-2698 v4 (Broadwell, 64 cores/CPU)
# CPUs	2	2
GPU model	Nvidia A100-SXM4-40GB	Nvidia Tesla V100-SXM2-32GB
# GPUs	8	8
Memory	1 TB	512 GiB

Table 1: *Experimental setup*

For all experiments, we used Ubuntu 20.04 and we installed an Nvidia GPU driver with the default power management configuration. The power consumption of the Sirius cluster is monitored by an Omegawatt [45] power meter, which has a precision of 0.1 watts (W). We used it with a sampling frequency of 1 Hz. Additionally, we used a software-based power meter called ALUMET [46] that gathers power metrics from Nvidia NVML and Intel RAPL at a sampling frequency of 2 Hz. To ensure reproducibility, all results are averaged from seven experiments. The code and data we used are publicly available [47]. We based our experiments on the Diffusers library and the Accelerate optimizer framework.

We were able to train the v1-1 Stable Diffusion model on Sirius with the same gradient accumulation, batch size, and optimizer as originally. The learning rate was kept constant. The original training was distributed across 32 nodes. Assuming that the energy consumed by each node is equivalent, we carried out the experiments on a single node. We used the Pokemon BLIP captions dataset [48] which contains 833 images with captions. A linear regression was trained on data points gathered from 61 training experiments with 7 to 3500 training steps, and we tested it on 6 experiments with 5000 to 6500 training steps. Two resolutions of images were used for the original training, 256x256 and 512x512, so we conducted the experiments and built a regression for each resolution (Equation 1 and 2, respectively). Those regressions were validated with a score higher than 99%.

$$\text{Energy (kWh)} = 5.26e^{-04} \times N + 2.01e^{-02} \quad (1)$$

$$\text{Energy (kWh)} = 1.78e^{-03} \times N + 1.64e^{-02} \quad (2)$$

Where N is the number of training steps.

Table 2 presents the estimated energy consumed by the model versions that are pertinent for this work, based on the number of steps provided by the developers of Stable Diffusion and our regression models.

The estimated energy was multiplied by the number of nodes originally used (32). The obtained values are close to existing studies on similar models [49]. We discuss this methodology in Section 4.3.

Version	Image size	# steps	Estimated energy (kWh)	
			1 node	32 nodes
v1-1	256	$2.37e^{+05}$	$4.70e^{+02}$	$1.50e^{+04}$
	512	$1.94e^{+03}$		
v1-4	512	$2.25e^{+05}$	$4.01e^{+02}$	$1.28e^{+04}$
v1-5	512	$5.95e^{+05}$	$1.06e^{+03}$	$3.39e^{+04}$

Table 2: *Estimated energy consumption of training Stable Diffusion (number of steps provided by the developers)*

3.3 LCA-based modeling

Metrics and methodology We use LCA to calculate the environmental costs of the service. In this way, we can obtain potential impacts for the entire life cycle of the resources employed, and for several impact categories. The choice of which environmental impact categories to measure is often constrained by the lack of data available on certain categories. However, we recommend a minimum of 3 impact categories for AI that are available in the different databases used in this study. The first is Abiotic Depletion Potential (ADP) which represents a decrease of minerals and metals resources [50]. The second, Global Warming Potential (GWP) [51], evaluates the contribution to climate change. The third, Primary Energy (PE), expresses the cumulative energy demand [52].

These 3 categories cover the most significant environmental impacts of digital technologies [25]. Water is an important issue in AI [53], but we don't include a water consumption indicator as we currently lack reliable data. Above all, water consumption tends to be a contextual issue, with one liter of water withdrawn having a different impact depending on the region and time of year, "When and Where matter" [53].

To evaluate the service according to these categories, we assess the cost of the average user journey for each module described in Section 3.1, in terms of electricity consumption and use of IT equipment. To obtain the footprint of this equipment and its electricity consumption, depending on the country of use, we rely on life cycle inventory (LCI) data: environmental databases from public agencies [54], consortiums such as NegaOctet [55], and open-source projects such as Boavizta [56]. The first [54] provides full life-cycle impacts for the electricity mix, the second [55] for networks and terminals parts, also in full life-cycle, and the third [56] for the datacenters parts but only for manufacturing and usage phases of the life-cycle.

These databases enable us to translate the electricity consumption and device usage, we measure or estimate, into our 3 impact categories. For example, the Equation 3 from [6] calculates the impact of the inference part. Electricity consumption $C_{i,e}$, PUE [57] and EGM_g are giving us the operational impact for the 3 impact categories. The same goes for embodied impacts, where each use of a device costs a share of the equipment's total footprint F_e. This share is determined by an allocation $a_e(t)$, proportional to the duration during which the equipment is used over the total time the equipment is used over its lifetime. Other allocations can be used to share the embodied footprint, e.g. the volume of data transferred as part of a network device. The other parts of the service, as described in Figure 2, are calculated respectively by their own equation described in [6]. Notations are summaries in Table 3.

$$I_{Inference} = \sum_i C_{i,e} \times EGM_g \times PUE + a_e(t) \times F_e \tag{3}$$

$I_{Inference}$:	Inference Impact
i :	Inference done on a GPU
e :	Equipment used for inference
$C_{i,e}$:	Consumption of electricity for i with e
EGM_g :	Electricity grid mix impact in a geographic area g
PUE :	Power usage effectiveness of the site
$a_e(t)$:	Allocation for e's time of use t for the entire duration of its use (i.e. lifespan times percentage of use)
F_e :	Footprint for e : manufacture, transport, and end of life

Table 3: *Equation and notation for calculating the inference impact*

Computing the impacts Based on our previous study [6], we have improved our model and, as an example, assessed the environmental impact of Gen-IA Stable Diffusion's service [42]. The user journey evaluated is that of a user visiting the site and submitting a request to the image generation service with standard settings. During the site's observation period, from August 2022 to August 2023, the standard settings returned 4 images in 512x512 format for a written request. For the one-year evaluation, we estimated the number of users from the measured request traffic [58, 59]. We assumed that half the visits led to one request for image generation.

We used the national average of the US electricity mix in our calculations for the "Inference" and "Training" sections. For the "End User Terminals" and "Network" sections, we calculated an average electricity mix based on the countries most represented in the user population.

Our assessment resulted in Figure 3 which shows, for one year of service, the distribution of impacts between the different parts for each impact. For more readability, the "Web hosting" part is included in the "Inference" part, and the "Data storage" part in the "Training" part. These parts have negligible impacts and are strongly correlated with the part in which they were merged. To produce our results, we based ourselves on the number of visits measured and the characteristics of all the different training phases (or versions) of the model, i.e. not only the last training. We also reworked our data to be able to separate operational impacts from embodied impacts, as can be seen in Figure 4.

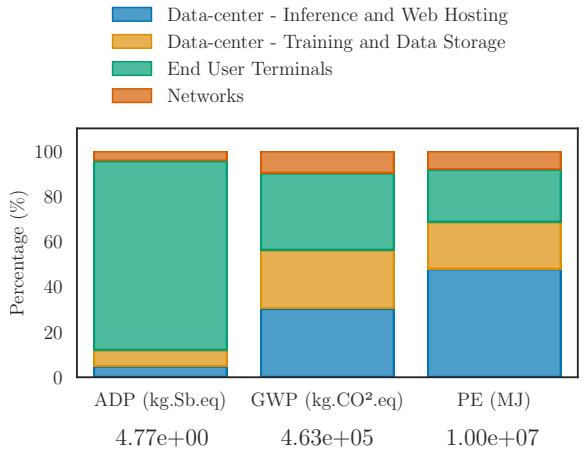

Figure 3: *Impact distribution for one year of Stable Diffusion as a service with 75M visits and 150M pictures generated.*

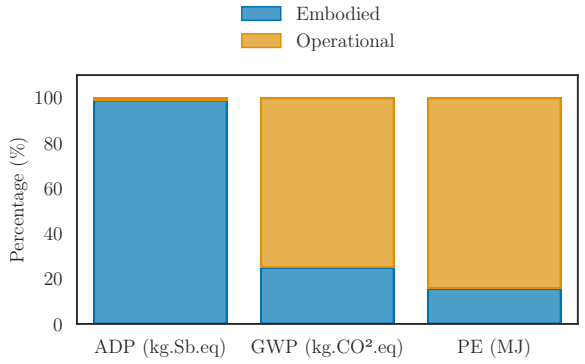

Figure 4: *Impact distribution between operational and embodied footprint for one year of Stable Diffusion as a service.*

Carrying out an LCA of this service enables having a model that could be parameterized, both in terms of volume and characteristics of the activity. In the next section, we discuss the significant impact of the service.

4 Rethink environmental sustainability for Gen-AI

Measuring the environmental impact of Gen-AI is a complex problem to which we have made the following contributions: service level modelization, estimation by regression, and LCA-based assessment. Once these contributions have been integrated, how do they modify our understanding of the sustainability of Gen-AI and its measurement?

4.1 LCA of digital services

Through the use of the LCA of this service, beyond the important impact, i.e. 463 tonnes of CO_2 eq., we can draw some information on the distribution of the environmental impact for one year of a Gen-AI service.

Firstly, Figure 3 shows that terminals and networks represent a significant share in the operation impact of a Gen-AI service: more than 85% of the ADP impact, more than 30% of the energy footprint, and 45% of the carbon footprint. It validates the need to take them into account, all the more so that the footprint of networks and terminals grows with the number of users, even if users do not use an online service, i.e. ignoring the cost of the "network" part.

Secondly, the multi-impact vision provided by LCA shows us that while decarbonizing the electricity consumed by data centers reduces the impact on climate change, embodied carbon emissions and those produced on the user side remain significant. The energy footprint is also a concern to consider. Reducing it would need important efficiency gains, but could trigger a rebound effect [60]. Any progress on the energy efficiency of Gen-AI could well lead to an increase in its total usage, as it is common in the digital sector [61].

Lastly, the issue of metal extraction poses a problem that is difficult to solve. Moreover, Gen-AI services are boosting demand for GPUs, critical resources whose footprint is still difficult to estimate. In our study, we are using underestimated values for the footprint of GPUs, based on a method dedicated to CPUs detailed in [62].

There are various ways of reducing the footprint associated with the manufacture of equipment: the use of low-carbon energy in the manufacturing process, optimization of the manufacturing or transport process, or the use of recycled materials tending towards a form of circularity, etc. But these necessary transformations of industry are levers that are difficult for service users and designers to access. The next section will therefore focus on the issue of lifespan.

4.2 Obsolescence of hardware and software

Ultimately, one of the most direct levers for reducing both ADP's footprint and embodied emissions remains to extend the life of the equipment. Moreover, if we also consider the general trend of the embodied carbon emission in IT [24], Gen-AI, with its high electricity consumption, as shown in Figure 4, could indicate a reversal of tendencies. However, there is nothing obvious about this conclusion and it depends on how we consider the separation between operational impacts and embodied impacts.

Figure 5, represents previously presented results in new categories.

Taking Figure 4, we have represented the service's network and training impacts in new, separate categories. In our previous representation, the training phase was mainly included as an operational impact due to its high electricity consumption. However, in the case of Gen-AI, it can be considered an embodied impact. Training once expended a large quantity of resources, similarly to the manufacture of equipment. It is, therefore, logical to consider training as a device that enables the Gen-AI service to function and as part of its embodied footprint, like a development cost [63]. We can then more naturally conclude that it is important to include Gen-AI models in the action of extending the life cycle of equipment. We need to extend the life of equipment, hardware, and software.

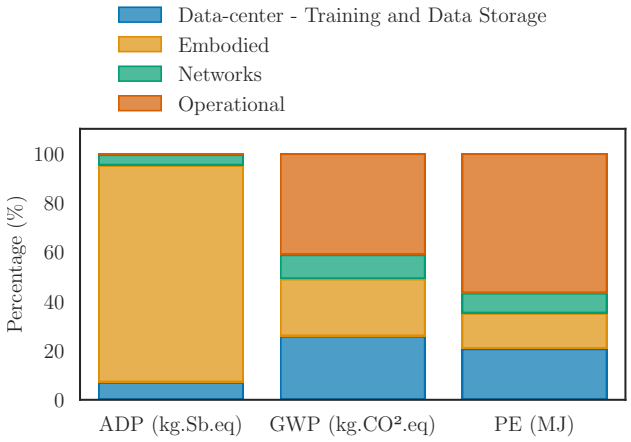

Figure 5: *Impact distribution between operational, embodied, networks and model training footprint for one year of stable diffusion as an online service. The Embodied and Operational categories cover the "Data-center - Inference" and "Web Hosting" service parts.*

We can also question the characterization of the network footprint. Admittedly, most of their footprint comes from their electricity consumption. However, as [64] explained, networks are a perpetually powered infrastructure where the transfer of data for a service does not in itself directly generate additional consumption. Especially for fixed networks, the infrastructure has a basic cost that provides for needs as long as they remain within its capacity. Modifying the volume of data transferred by the service may in fact have little effect on electricity consumption [65], unless it is on a scale that would require the infrastructure to be upscaled [64] or unless the majority of the network is mobile.

These last two points are not intended to show that a reduction strategy based on reducing data and energy flows could be ineffective. However, it would run the risk of being insufficient at best, or worse, of contributing to a rebound effect. On the other hand, an approach based on the parsimonious and efficient use of existing resources and infrastructures could lead to more sustainable gains. It is perhaps more interesting to optimize existing resources than to seek to create new, supposedly more efficient resources.

4.3 Balancing between accessibility and reliability of the electricity consumption estimation

In this section, we question the methodology we proposed to estimate the electricity consumption of the training phase and discuss the feasibility of deploying it at a large scale.

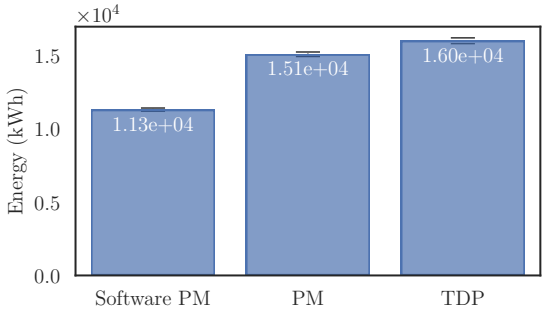

Figure 6: *Estimations of the electricity consumed by the v1-1 Stable Diffusion model training from various existing methods. PM: Estimation based on power meter; TDP: TDP-based estimation; Software PM: Estimation based on software-based power meter.*

We start by comparing our result based on power meter (PM) observations with two other estimations. First, a TDP-based estimation and secondly, an estimation based on the results of software-based power meters (software PM) (Figure 6). Both rely on a linear regression too, but in cases where a power meter is not available. Figure 6 shows the results of those methods applied to the training phase of the v1-1 version of Stable Diffusion. We see that a

TDP-based estimation is almost 6% higher than a PM-based estimation which is highly surprising considering the discussion made in Section 2.2, where we assumed a TDP-based estimation underestimates the consumption. To better understand the results, figures 7 and 8 compare the power consumed as measured by the power meters and the Thermal Design Power (TDP). Figure 7 shows the evolution of the

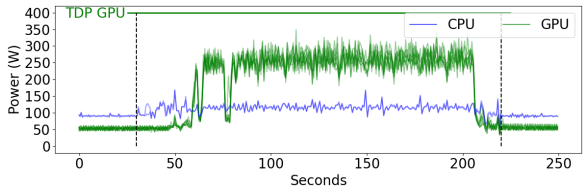

Figure 7: *Evolution of the power consumed by each GPU and CPU for 10 steps compared with the TDP of a GPU, on the Sirius cluster.*

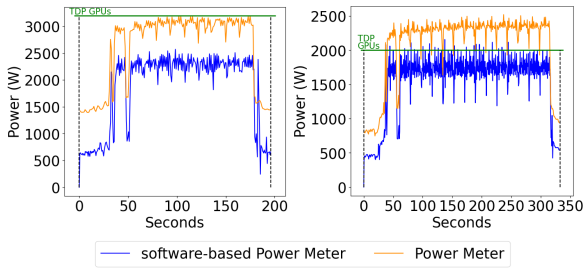

Figure 8: *Evolution of the server power according to different power meters during training steps compared with the sum of the GPU TDP values, on the Sirius cluster (left) and the Gemini cluster (right).*

power consumed during the execution of 10 steps of training Stable Diffusion on the cluster Sirius by every computing component (CPUs and GPUs) as reported by software-based power meters, compared with the TDP. It can be seen that the average power is largely lower than the TDP, even though the GPU average utilization is above 95% when the GPUs are in their intense phase. A TDP-based estimation overestimates the electricity consumed by the GPUs in this case. The left graph of Figure 8 shows the consumption of the server, as reported by the power meter, of the same training execution. We can notice that the sum of the GPU TDP values is very close to the power consumed by the server. We conclude that the underestimation of a TDP-based approach is compensated by the power consumed by the other components of the server in this case. However, this would not be the case in every workload or with other equipment. We executed the same code for 150 training steps on the Gemini cluster, whose specifications can be found in table 1, and the power observed can be seen in the right graph of Figure 8. This time the power consumed by the server is higher than the TDP, which lets us think that a TDP-based estimation is unreliable and depends on the workload and the server used. The conclusion would have been the same if we had included other components (CPUs) in the TDP-based estimation.

Figure 8 also shows that the difference between the power measures power of the software PM and the PM is quite significant (around 20%), which results in a 25% difference when scaling to hundreds of thousands of training steps in figure 6.

To conclude, the best solution to estimate the electricity consumption of the training phase of a Gen-AI model is to be able to replicate the training step while monitoring with an accurate power meter. Without access to a power meter, software-based power meters give an accurate measure of the computing components, but with a significant difference with power meters which is exaggerated by the number of training steps.

If it is not possible to replicate the training steps, but information like the duration of the execution is available, using the TDP can provide an estimation, but without any guarantee of its accuracy.

4.4 Current limits

The first limitation of our evaluation work is the distance we still have to cover to best represent the operation of existing Gen-AI services. Without internal information on how the services operate, it is difficult to model the impacts correctly. In our case, the Stable Diffusion service was open-source, as was its main dataset, and details on the training sessions, such as the number of steps, were public. We still had to make some assumptions, for example about the sizing of the Webhosting part. Although we had indicators of site traffic, we couldn't find any model-based way of sizing the server we needed. This is one example of a possible improvement. These gaps need to be filled, whether by better modeling of services based on open-source information or by greater transparency on the part of service providers.

The second limitation is that, even though we are presenting one of the first multi-criteria approaches to Gen-AI, it is still important to consider the environmental impact categories that we were unable to cover. To avoid transferring impacts from measured categories to unmeasured categories, we need to be exhaustive. For example, a predominantly nuclear electricity mix could transfer part of the "global warming" impact to the "ionizing radiation" impact. The transfer of impact is not essentially bad, but it should remain transparent. Similarly, for water, a data center powered by renewable energy could be a major consumer of water in a water-stressed area.

Increasing the number of environmental impact categories observed could lead to more informed and legitimate choices but will depend on the availability of more LCI data. Similarly, it is still difficult to include completely the impact of the end of life of equipment, as LCI data are still scarce [66].

The final limitation is the issue of allocating the footprint of the training. We made our assumptions based on the traffic to the site offering the service. In our case, as the model is freely available online, it can be used outside the site and further trained on a specific dataset thus we cannot determine the full scope of training and inference made thanks to it. As the models have only been available as a service on the site for a year, we remain confident that the site has concentrated a major part of the use of these models, even more so today when some might consider them obsolete. The issue of footprint allocation is non-trivial. If an organization hosted a Gen-AI service using a model it had not trained, who would be responsible for the model's footprint? Some might even think that the most prominent models would have a negligible training cost compared to the countless inferences they enable across multiple services. But it is not supported by existing studies [20, 67].

4.5 Future work

One of the Gen-AI service mechanisms that we have not yet been able to integrate is fine-tuning. Common to Gen-AI services [68, 8], fine-tuning corresponds to any additional training of the model between its main training phase and its use by the final client. Fine-tuning can be carried out at the level of a company deciding to integrate a Gen-AI service, which would then train the model on its own data. It can also be carried out at end-user level, or even at the level of an end-user usage session. These actions mobilize resources similar to those used for training. It will be important to integrate these costs into our service evaluation model and find a way of measuring or estimating the frequency and intensity of these additional training sessions. These costs could change the way we divide the Gen-AI footprint between inference and training.

Another consequence of the growth of Gen-AI services is the transformation they are bringing about in data centers. Gen-AI services require massive use of specific resources, i.e. GPUs. This new demand for access to a large number of GPUs as a service generates numerous organizational challenges [69, 70, 71], due in part to the difficulty of pooling GPUs as easily as CPUs. As a result, data centers are likely to become less efficient at sharing resources. This drop in efficiency will have an environmental cost, as it will probably have to be compensated for by quantity in order to meet demand. It would then be interesting, in a consequentialist approach [72], to calculate the cost of this growth burst largely attributable to Gen-AI services.

We believe our methodology is not restricted to Gen AI models and could be applied to any AI models deployed as a service. They are applications with different characteristics as a model only deployed to a restricted set of users or on the contrary a model integrated into a daily-used service, such as a webmail. In those use cases, the repartition of impact between service phases would be different than our use case and could bring more insights.

5 Conclusion

Measuring the environmental impact of Gen-AI is the basis for assessing sustainability. By adding service scale assessment, full life cycle multi-category cost, and enhanced energy estimation by sampling, we significantly improved our knowledge of Gen-AI environmental impact from the narrowness of mono-category assessment and uncertainties of TDP estimates. More than underlining the significant cost of this technology, we have highlighted the different sources and types of environmental impact. It is this detailed knowledge of Gen-AI's footprint that will enable it to be reduced. Our method gives us the means to better predict the potential impact of both training of the model and use of the service. From a sustainability perspective, this a priori assessment capability is an essential decision-making tool. To even more enhance our knowledge of Gen-AI sustainability, we invite the community to be more transparent, not only in terms of accessible code and electricity consumption reports. Serious evaluation of the impact of a Gen-AI technology requires data about the concrete means used to deploy and run this as a service. An informed opinion on the sustainability of Gen-AI in our society will require reliable knowledge of the environmental impact of Gen-AI services.

Acknowledgements

The authors used data from the open-source project Boavizta (`https://boavizta.org/en`), especially the work of Samuel Rince on GPU. Experiments presented in this article were carried out using the Grid'5000 / Slices 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 the BATE project

(BATE-UGAREG21A87) of the Auvergne Rhône-Alpes French region.

References

[1] E. Brynjolfsson, D. Li, and L. R. Raymond, *Generative AI at work*, Apr. 2023. DOI: 10.3386/w31161. [Online]. Available: https://www.nber.org/papers/w31161 (visited on 10/02/2023).

[2] F. García-Peñalvo and A. 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*, vol. 8, pp. 7–16, Jul. 2023, Accepted: 2023-08-28T12:17:56Z Publisher: International Journal of Interactive Multimedia and Artificial Intelligence, ISSN: 1989-1660. DOI: 10.9781/ijimai.2023.07.006. [Online]. Available: https://reunir.unir.net/handle/123456789/15134 (visited on 11/30/2023).

[3] J. Pohl, V. Frick, M. Finkbeiner, and T. Santarius, "Assessing the environmental performance of ICT-based services: Does user behaviour make all the difference?" *Sustainable Production and Consumption*, vol. 31, pp. 828–838, May 1, 2022, ISSN: 2352-5509. DOI: 10.1016/j.spc.2022.04.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352550922000914 (visited on 07/26/2022).

[4] A. Rasoldier, J. Combaz, A. Girault, K. Marquet, and S. Quinton, "How realistic are claims about the benefits of using digital technologies for GHG emissions mitigation?" In *Eighth Workshop on Computing within Limits 2022*, LIMITS, 245 Main St, Cambridge MA: PubPub, 2022, p. 14.

[5] J. C. T. Bieser, R. Hintemann, L. M. Hilty, and S. Beucker, "A review of assessments of the greenhouse gas footprint and abatement potential of information and communication technology," *Environmental Impact Assessment Review*, vol. 99, p. 107 033, Mar. 2023, ISSN: 0195-9255. DOI: 10.1016/j.eiar.2022.107033. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0195925522002992 (visited on 05/19/2023).

[6] A. Berthelot, E. Caron, M. Jay, and L. Lefèvre, "Estimating the environmental impact of Generative-AI services using an LCA-based methodology," in *CIRP LCE 2024 - 31st Conference on Life Cycle Engineering*, Turin, Italy: Elsevier, Jun. 2024, pp. 1–10. [Online]. Available: https://inria.hal.science/hal-04346102.

[7] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (visited on 11/30/2023).

[8] A. D. Cheok and E. Y. Zhang, *From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models*, Oct. 2023. DOI: 10.32388/3NTOLQ.2. [Online]. Available: https://www.qeios.com/read/3NTOLQ.2 (visited on 11/30/2023).

[9] S. CompVis and LAION, *Stable diffusion v1-4 model card*, https://huggingface.co/CompVis/stable-diffusion-v1-4#training.

[10] A. de Vries, "The growing energy footprint of artificial intelligence," *Joule*, vol. 7, no. 10, pp. 2191–2194, Oct. 18, 2023, ISSN: 2542-4351. DOI: 10.1016/j.joule.2023.09.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542435123003653 (visited on 10/24/2023).

[11] M.-J. Chen and J. Leong, *Nvidia and the great east–west semiconductor game*, Rochester, NY, Apr. 2022. DOI: 10.2139/ssrn.4085010. [Online]. Available: https://papers.ssrn.com/abstract=4085010 (visited on 11/30/2023).

[12] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, "Compute trends across three eras of machine learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-440, IEEE, Jul. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9891914. [Online]. Available: https://ieeexplore.ieee.org/document/9891914 (visited on 12/13/2023).

[13] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. [Online]. Available: https://aclanthology.org/P19-1355.

[14] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the Carbon Emissions of Machine Learning," arXiv, Tech. Rep. arXiv:1910.09700, Nov. 2019, arXiv:1910.09700 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1910.09700 (visited on 06/13/2022).

[15] Q. Cao, A. Balasubramanian, and N. Balasubramanian, "Towards accurate and reliable energy measurement of NLP models," in *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, Online: Association for Computational Linguistics, 2020, pp. 141–148. DOI: 10.18653/v1/2020.sustainlp-1.19. [Online]. Available: https://www.aclweb.org/anthology/2020.sustainlp-1.19 (visited on 10/15/2021).

[16] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *Journal of Machine Learning Research*, vol. 21, no. 248, pp. 1–43, 2020, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v21/20-312.html (visited on 10/30/2023).

[17] D. Patterson *et al.*, "Carbon Emissions and Large Neural Network Training," en, *ArXiv*, vol. abs/2104.10350, p. 22, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233324338.

[18] D. Patterson *et al.*, "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, Jul. 2022, Conference Name: Computer.

[19] M. Jay, V. Ostapenco, L. Lefèvre, D. Trystram, A.-C. Orgerie, and B. 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*, IEEE, May 1, 2023, pp. 106–118. [Online]. Available: https://hal.inria.fr/hal-04030223 (visited on 03/27/2023).

[20] C.-J. Wu *et al.*, "Sustainable ai: Environmental implications, challenges and opportunities," *Proceedings of Machine Learning and Systems*, vol. 4, pp. 795–813, 2022.

[21] A. Das and A. Modak, "The carbon footprint of machine learning models," *IJERA*, vol. 3, pp. 246–249, May 24, 2023. DOI: 10.5281/zenodo.8012383. [Online]. Available: https://zenodo.org/record/8012383 (visited on 09/01/2023).

[22] A.-L. Ligozat, J. Lefevre, A. Bugeau, and J. Combaz, "Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions," *Sustainability*, vol. 14, p. 5172, Apr. 25, 2022. DOI: 10.3390/su14095172.

[23] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176b parameter language model," *Journal of Machine Learning Research*, vol. 24, no. 253, pp. 1–15, 2023, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v24/23-0069.html (visited on 09/28/2023).

[24] U. Gupta *et al.*, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Los Alamitos, CA, USA: IEEE Computer Society, Feb. 1, 2021, pp. 854–867, ISBN: 978-1-66542-235-2. DOI: 10.1109/HPCA51647.2021.00076. [Online]. Available: https://www.computer.org/csdl/proceedings-article/hpca/2021/223500a854/1t0HVFxaAFy (visited on 02/05/2023).

[25] 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. 2021. [Online]. Available: https://www.greenit.fr/wp-content/uploads/2021/12/EU-Study-LCA-7-DEC-EN.pdf (visited on 12/08/2021).

[26] ISO/TC 207/SC 5, "Environmental management — life cycle assessment — principles and framework," International Organization for Standardization, Geneva, CH, Standard ISO 14040:2006, Jul. 2006, https://www.iso.org/standard/37456.html.

[27] ISO/TC 207/SC 5, "Environmental management — life cycle assessment — requirements and guidelines," International Organization for Standardization, Geneva, CH, Standard ISO 14044:2006, Jul. 2006, https://www.iso.org/standard/38498.html.

[28] ITU, *ITU l1410 : Methodology for environmental life cycle assessments of information and communication technology goods, networks and services*, Dec. 2014. [Online]. Available: https://www.itu.int/rec/dologin_pub.asp?lang=f&id=T-REC-L.1410-201412-I!!PDF-E&type=items.

[29] European Commission. Joint Research Centre. Institute for Environment and Sustainability., *International Reference Life Cycle Data System (ILCD) Handbook :specific guide for Life Cycle Inventory (LCI) data sets.* LU: Publications Office, 2010. [Online]. Available: https://data.europa.eu/doi/10.2788/39726 (visited on 04/15/2022).

[30] B. Whitehead *et al.*, "The environmental burden of data centres – a screening LCA methodology," in *CIBSE ASHRAE Technical Symposium, Imperial College, London UK*, 222 Balham High Road, London: CIBSE, May 2012, p. 17.

[31] J. Malmodin, D. Lundén, M. Nilsson, and G. Andersson, "LCA of data transmission and IP core networks," *2012 Electronics Goes Green 2012+*, pp. 1–6, Sep. 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:20354449.

[32] R. G. Hunt, W. E. Franklin, and R. G. Hunt, "LCA — how it came about: — personal reflections on the origin and the development of LCA in the USA," *The International Journal of Life Cycle Assessment*, vol. 1, no. 1, pp. 4–7, Mar. 1996, ISSN: 0948-3349, 1614-7502. DOI: 10.1007/BF02978624. [Online]. Available: http://link.springer.com/10.1007/BF02978624 (visited on 11/25/2021).

[33] I. Boustead, "LCA — how it came about: — the beginning in the u.k.," *The International Journal of Life Cycle Assessment*, vol. 1, no. 3, pp. 147–150, Sep. 1996, ISSN: 1614-7502. DOI: 10.1007/BF02978943. [Online]. Available: https://doi.org/10.1007/BF02978943 (visited on 12/07/2021).

[34] T. Billstein, A. Björklund, and T. Rydberg, "Life cycle assessment of network traffic: A review of challenges and possible solutions," *Sustainability*, vol. 13, no. 20, p. 11 155, Jan. 2021, Number: 20 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/su132011155. [Online]. Available: https://www.mdpi.com/2071-1050/13/20/11155 (visited on 02/04/2022).

[35] S. Pauliuk, G. Majeau-Bettez, C. Mutel, B. Steubing, and K. Stadler, "Lifting industrial ecology modeling to a new level of quality and transparency: A call for more transparent publications and a collaborative open source software framework," *Journal of Industrial Ecology*, vol. 19, n/a–n/a, Jun. 1, 2015. DOI: 10.1111/jiec.12316.

[36] N. C. Horner, A. Shehabi, and I. L. Azevedo, "Known unknowns: Indirect energy effects of information and communication technology," *Environmental Research Letters*, vol. 11, no. 10, p. 103 001, Oct. 5, 2016, Institution: Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States) Publisher: IOP Publishing, ISSN: 1748-9326. DOI: 10.1088/1748-9326/11/10/103001. [Online]. Available: https://www.osti.gov/pages/biblio/1377527 (visited on 12/07/2021).

[37] L. M. Hilty and B. Aebischer, "ICT for sustainability: An emerging research field," in *ICT Innovations for Sustainability*, L. M. Hilty and B. Aebischer, Eds., ser. Advances in Intelligent Systems and Computing, Cham: Springer International Publishing, 2015, pp. 3–36, ISBN: 978-3-319-09228-7. DOI: 10.1007/978-3-319-09228-7_1.

[38] A. Lundström and D. Pargman, "Developing a framework for evaluating the sustainability of computing projects," in *Proceedings of the 2017 Workshop on Computing Within Limits*, ser. LIMITS '17, New York, NY, USA: Association for Computing Machinery, Jun. 22, 2017, pp. 111–117, ISBN: 978-1-4503-4950-5. DOI: 10.1145/3080556.3080562. [Online]. Available: https://dl.acm.org/doi/10.1145/3080556.3080562 (visited on 12/05/2023).

[39] L. Grimal, I. Di Loreto, N. Burger, and N. Troussier, "Design of an interdisciplinary evaluation method for multi-scaled sustainability of computer-based projects. a workbased on the sustainable computing evaluation framework (SCEF)," *LIMITS Workshop on Computing within Limits*, 2021. DOI: 10.21428/bf6fb269.2ee80cf1. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-03616569 (visited on 04/20/2022).

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 10 684–10 695.

[41] HF, *Hugging Face runwayml stable diffusion repository*, 2016. [Online]. Available: https://huggingface.co/runwayml/stable-diffusion-v1-5 (visited on 09/29/2023).

[42] SD, *Stable diffusion service*, 2022. [Online]. Available: https://stablediffusionweb.com/#demo (visited on 09/09/2023).

[43] L. F. W. Anthony, B. Kanding, and R. Selvan, *Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models*, en, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2007.03051 (visited on 10/15/2021).

[44] F. 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: IEEE/ACM, Nov. 2005, pp. 99–106. [Online]. Available: https://hal.inria.fr/inria-00000284.

[45] OmegaWatt, *Omegawatt*, 2018. [Online]. Available: http://www.omegawatt.fr/.

[46] G. Raffin and M. Jay, *Alumet preliminary version: NVML sensor (CPU+GPU)*, 2023. [Online]. Available: https://github.com/TheElectronWill/nvml-sensor.

[47] M. Jay, *Measuring the electricity consumption of training and inferring from stable diffusion*, https://github.com/mjay42/Assessing-the-electricity-consumption-of-ML-training/tree/main/StableDiffusion, 2023.

[48] J. N. M. Pinkney, *Pokemon blip captions*, https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

[49] S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of ai deployment?" In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24, Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, pp. 85–99, ISBN: 9798400704505. DOI: 10.1145/3630106.3658542. [Online]. Available: https://doi.org/10.1145/3630106.3658542.

[50] L. van Oers, J. B. Guinée, and R. Heijungs, "Abiotic resource depletion potentials (ADPs) for elements revisited—updating ultimate reserve estimates and introducing time series for production data," *The International Journal of Life Cycle Assessment*, vol. 25, no. 2, pp. 294–308, Feb. 1, 2020, ISSN: 1614-7502. DOI: 10.1007/s11367-019-01683-x. [Online]. Available: https://doi.org/10.1007/s11367-019-01683-x (visited on 12/15/2023).

[51] I. P. on Climate Change (IPCC), "The earth's energy budget, climate feedbacks and climate sensitivity," in *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2023, pp. 923–1054. DOI: 10.1017/9781009157896.009.

[52] R. Frischknecht, F. Wyss, S. Büsser Knöpfel, T. Lützkendorf, and M. Balouktsi, "Cumulative energy demand in LCA: The energy harvested approach," *The International Journal of Life Cycle Assessment*, vol. 20, no. 7, pp. 957–969, 2015, ISSN: 1614-7502. DOI: 10.1007/s11367-015-0897-4. (visited on 05/03/2024).

[53] P. Li, J. Yang, M. A. Islam, and S. Ren, *Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models*, Apr. 6, 2023. DOI: 10.48550/arXiv.2304.03271. arXiv: 2304.03271[cs]. [Online]. Available: http://arxiv.org/abs/2304.03271 (visited on 04/14/2023).

[54] ADEME, *Base impact®v2.02*. [Online]. Available: https://base-empreinte.ademe.fr/ (visited on 12/07/2023).

[55] N, *Negaoctet*, 2023. [Online]. Available: https://negaoctet.org/ (visited on 09/29/2023).

[56] T. Simon, D. Ekchajzer, A. Berthelot, E. Fourboul, S. Rince, and R. Rouvoy, "BoaviztAPI: a bottom-up model to assess the environmental impacts of cloud services," in *HotCarbon'24. Workshop on Sustainable Computer Systems*, Santa Cruz, United States, Jul. 2024, p. 7. [Online]. Available: https://hal.science/hal-04621947.

[57] C. L. Belady and C. G. Malone, "Metrics and an infrastructure model to evaluate data center efficiency," vol. ASME 2007 InterPACK Conference, Volume 1, Jul. 2007, pp. 751–755. DOI: 10.1115/IPACK2007-33338. eprint: https://asmedigitalcollection.asme.org/InterPACK/proceedings-pdf/InterPACK2007/42770/751/2666558/751_1.pdf. [Online]. Available: https://doi.org/10.1115/IPACK2007-33338.

[58] H, *Hypestat*, 2023. [Online]. Available: https://hypestat.com/ (visited on 10/29/2023).

[59] S, *Similarweb*, 2023. [Online]. Available: https://www.similarweb.com/ (visited on 10/29/2023).

[60] S. Sorrell, "Jevons' paradox revisited: The evidence for backfire from improved energy efficiency," *Energy Policy*, vol. 37, no. 4, pp. 1456–1469, Apr. 2009, ISSN: 0301-4215. DOI: 10.1016/j.enpol.2008.12.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301421508007428 (visited on 12/08/2023).

[61] V. C. Coroama and F. Mattern, "Digital rebound - why digitalization will not redeem us our environmental sins," in *Proceedings of the 6th International Conference on ICT for Sustainability, ICT4S 2019, Lappeenranta, Finland, June 10-14, 2019*, Aachen, Germany: CEUR-WS.org, 2019.

[62] S. Rince. "Gpu component manufacture impacts," Boavizta. (2023), [Online]. Available: https://github.com/Boavizta/boaviztapi/issues/65 (visited on 12/14/2023).

[63] T. Simon, P. Rust, R. Rouvoy, and J. Penhoat, "Uncovering the environmental impact of software life cycle," in *International Conference on Information and Communications Technology for Sustainability*, IEEE, Jun. 5, 2023, pp. 176–187. [Online]. Available: https://inria.hal.science/hal-04082263 (visited on 06/23/2023).

[64] G. Guennebaud, A. Bugeau, and A. Dudouit, "Assessing VoD pressure on network power consumption," in *ICT4S - International Conference on Information and Communications Technology for Sustainability*, Rennes, France: IEEE, Jun. 2023, pp. 76–86. [Online]. Available: https://hal.science/hal-04059523.

[65] D. Schien, P. J. S. Shabajee, H. B. Akyol, L. Benson, and A. Katsenou, *Help, i shrunk my savings! assessing the carbon reduction potential for video streaming from short-term coding changes*, 2023.

[66] M. Ficher, T. Bauer, and A.-L. Ligozat, "A comprehensive review of the end-of-life modeling in LCAs of digital equipment," *The International Journal of Life Cycle Assessment*, 2024, ISSN: 1614-7502. DOI: 10.1007/s11367-024-02367-x. [Online]. Available: https://doi.org/10.1007/s11367-024-02367-x (visited on 09/30/2024).

[67] C.-J. Wu, B. Acun, R. Raghavendra, and K. Hazelwood, "Beyond efficiency: Scaling ai sustainably," *IEEE Micro*, pp. 1–8, 2024. DOI: 10.1109/MM.2024.3409275.

[68] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges," *Future Internet*, vol. 15, no. 8, p. 260, Aug. 2023, Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1999-5903. DOI: 10.3390/fi15080260. [Online]. Available: https://www.mdpi.com/1999-5903/15/8/260 (visited on 11/30/2023).

[69] B. He *et al.*, "Dxpu: Large-scale disaggregated gpu pools in the datacenter," *ACM Trans. Archit. Code Optim.*, vol. 20, no. 4, Dec. 2023, ISSN: 1544-3566. DOI: 10.1145/3617995. [Online]. Available: https://doi.org/10.1145/3617995.

[70] F. Filippini, J. Anselmi, D. Ardagna, and B. Gaujal, "A Stochastic Approach for Scheduling AI Training Jobs in GPU-based Systems," *IEEE Transactions on Cloud Computing*, no. 01, pp. 1–17, 2023. DOI: 10.1109/TCC.2023.3336540. [Online]. Available: https://hal.science/hal-04337856.

[71] Z. Ye *et al.*, "Deep learning workload scheduling in gpu datacenters: A survey," *ACM Comput. Surv.*, vol. 56, no. 6, Jan. 2024, ISSN: 0360-0300. DOI: 10.1145/3638757. [Online]. Available: https://doi.org/10.1145/3638757.

[72] T. Schaubroeck, "Relevance of attributional and consequential life cycle assessment for society and decision support," *Frontiers in Sustainability*, vol. 4, p. 1 063 583, 2023.