Environmental Impact of Artificial Intelligence

Etienne Delort, Laura Riou, Anukriti Srivastava

Artificial Intelligence and Eco-responsibility Internship

Bibliographic Report

Etienne Delort, Laura Riou & Anukriti Srivastava

ENVIRONMENTAL IMPACTS OF ARTIFICIAL INTELLIGENCE

September 2023

Contents

Abbreviations

ADEME Agence Française de l'environnement et de la maîtrise de l'énergie

AI Artificial Intelligence

API Application Programming Interface

BMC Baseboard Management Controller

CI Carbon Intensity

CNN Convolutional Neural Networks

CORSE Compiler Optimization and Run-time Systems

CPU Central Processing Unit

DRAM Dynamic Random Access Memory

FLOP Floating Point Operation

GHG Greenhouse Gas

GPU Graphical Processing Unit

ICT Information and Communication Technology

Inria Institut National de Recherche en Informatique et Automatique

IoT Internet of Things

IPCC Intergovernmental Panel on Climate Change

LCA Life Cycle Analysis

LCI Life Cycle Inventory

LLM Large Language Models

MACs Multiply Accumulate Operations

NLP Natural Language Processing

nvidia-smi NVIDIA System Management Interface

PDU Power Distribution Unit

PUE Power Usage Effectiveness

R&D Research and Development

RAPL Running Average Power Limit

SIMD Single Instruction Multiple Data

TDP Thermal Design Power

TPU Tensor Processing Unit

WUE Water Usage Effectiveness

Context of this Bibliographic Work

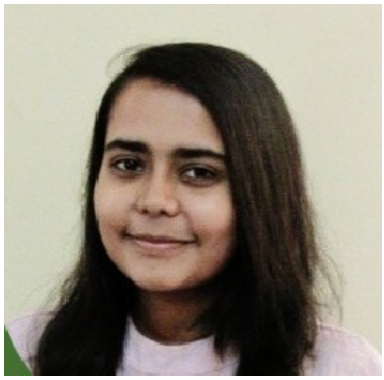
This bibliographic report was written as part of the *AI & Eco-design* internship, which was the result of a collaboration between Inria and CEA-LETI Eco-innovation. The internship was funded by the CEA-LETI Carnot programme [1] and the DeepGreen project[2]. It was supervised by

- Dr. Fabrice RASTELLO, Research Director, Inria - CORSE
- Dr. Guillaume IOOSS, Research Associate, Inria - CORSE
- Ms. Audrey VIDAL, Research Engineer, CEA-Leti/DSys
- Mr. Josua GUERID, Research Engineer, CEA-Leti/DSys

The internship took place between 3rd April and 29th September 2023 and was carried out by three trainees:

Laura Riou

Etienne Delort



Anukriti Srivastava

Laura Riou is a graduate engineer in Information Systems Engineering from Grenoble INP - Ensimag. She is completing a double degree in Ecological Transition Management with Science Po Grenoble. Etienne Delort is doing his final year internship in the Computer Science, Artificial Intelligence, and Data Science department at IMT Mines Alès. Finally, Anukriti Srivastava is a graduate engineer in Computer Science specializing in Image and Multimedia from Toulouse INP - ENSEEIHT.

The aim of the *AI & Eco-design* internship was to assess the environmental footprint of AI algorithms and raise awareness among AI practitioners of the impact of their work. This began with an in-depth study of the state of the art on the environmental impact of AI that is presented in this bibliographic report. This study continued for the rest of the internship to help provide a scientific basis for the other tasks we worked on such as (i) measuring and predicting the energy consumed while running inferences on Large Language Models on the Grid'5000 testbed, (ii) exploring solutions that enable users of CEA's clusters to evaluate the environmental impact of their jobs and (iii) proposing directions for CEA to take into account the environmental impact of their AI projects.

This report aims to structure our knowledge of the state of the art on the environmental impact of AI and its measurement that we have acquired from our readings.

[1] The CEA-LETI Carnot Institute: https://www.leti-cea.fr/cea-tech/leti/Pages/Leti/a-propos-du-Leti/Leti-Institut-Carnot.aspx

[2] Collective and open and sovereign deep learning platform for embedded systems: https://anr.fr/fileadmin/aap/2022/france2030-ami-IA-deepgreen.pdf

Introduction

Large AI models have high training costs in terms of energy and carbon emissions generated due to computation and the manufacture of specialized hardware accelerators on which such models are trained. Hence, the impact of such models on our environment is significant. However, despite the pervasiveness of such models and AI services based on them, some of which have billions of users (for example, ChatGPT has been visited 1.5 billion times in August 2023 according to SimilarWeb[3]), the environmental impact of Artificial Intelligence (AI) as a domain has not been studied or reported. Moreover, quantification of the environmental impact of the latter involves studying the collective impacts of myriad AI models and their applications. The assessments made for the environmental impact of recent models, which we shall detail later in this report, are often partial since they only study carbon footprint which is insufficient for understanding their overall environmental impact.

This bibliographic work focuses on gathering literary resources (papers, articles, reports, etc.) that provide context for and explain techniques for the reliable measurement of the environmental impact of AI algorithms.

This report is organized in three parts. First, we present the environmental impacts of ICT and then of AI which is a part of the former (Section 1). Second, we tackle the ways of quantifying the environmental impact of AI (Section 2). Third, we review some case studies on environmental impact assessment for AI (Section 3).

[3]Number of visits on ChatGPT login page in August 2023: https://www.similarweb.com/website/chat.openai.com/#overview

Important Definitions

For clarity, in this section, we define some of the key notions around AI and environmental impact assessment necessary for understanding this bibliographic work.

The AI models that are discussed here have a common algorithmic and conceptual basis. It is a diverse group that includes algorithms used to create systems such as the ChatGPT chatbot, robot vacuum cleaner, or computer vision analysis software for facilitating recruitment processes. Importantly, for us, AI algorithms are inseparable from the services in which they are used and are often very hardware-based.

1. Artificial Intelligence There is no unique definition of AI and it is an ever-evolving field. For example, Russell et al. [4] define AI as the *"the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment."* More concretely, this view of AI brings together several distinct sub-fields of computer vision, speech processing, natural language understanding, reasoning, knowledge representation, learning, and robotics, to have the machine achieve an outcome.

2. Deep Learning Deep learning is part of a broader family of AI methods based on artificial neural networks. These algorithms can execute a wide variety of tasks, by first learning information from a lot of examples (*training* of the model) and then producing new information on unseen cases (*inference*). Deep learning models can be used for a variety of tasks including image classification, text generation, or even extension of musical tracks. **Note:** In Section 2, while talking about AI, we are mainly focusing on Deep learning models which are the type of AI that are used the most these days.

3. Natural Language Processing Natural Language Processing (NLP) is the field of study that deals with human language, combining the knowledge of linguistics, computer science, and AI. Deep learning is bringing massive performance improvements to this domain, notably since the introduction of the Transformer model described in the root paper Vaswani et al. [9].

4. Carbon Dioxide Equivalent Abbreviated as CO_2-eq [5]or CO_2e, it is a metric measure used to compare the emissions from various greenhouse gases on the basis of their global warming potential (GWP) by converting amounts of other gases to the equivalent amount of carbon dioxide with the same global warming potential.

[4]Artificial Intelligence: A Modern Approach, Stuart Russell and Peter Norvig,1995

[5]Carbon Dioxide Equivalent Definition https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Carbon_dioxide_equivalent

1 Environmental Impact of ICT and AI Among It

According to the Intergovernmental Panel on Climate Change (IPCC) [6], from 1870 to 2010, human activities emitted 2100 $GtCO_2e$, leading to a global climate warming of 1.3°C within those 140 years. The IPCC warns about the high risks of going above 1.5°C of global warming. Keeping global warming under this limit was fixed as an objective by 195 countries through the Paris Agreement in 2015. To meet this objective, global Greenhouse Gas (GHG) emissions, estimated at 49 $GtCO_2e$ in 2010, must be reduced each year by 9% and reach zero by 2050. Thus, all sectors of human activities must reduce their emissions.

Information and Communication Technology (ICT) sector is one of them, in which AI is included. According to Wikipedia [6],

> "ICT is an umbrella term that includes any communication device, encompassing radio, television, cell phones, computer and network hardware, satellite systems, and so on, as well as the various services and appliances with them such as video conferencing and distance learning."

Here, we discuss ICT's contribution to global warming in 2020 (Section 1.1.3), its expected contribution to global warming for the next decades (Section 1.1.4), multi-criteria evaluation of ICT's environmental impacts (Section 1.1.2) and will finally discuss environmental impacts of AI with respect to those of ICT (Section 1.2).

1.1 Environmental Impacts of ICT

In this first section, different orders of environmental impacts will be shown (Section 1.1.1), then the importance and looking at multiple criteria and categories of impact will be proven (Section 1.1.2). Finally based mainly on the important study from Freitag et al. [26] we will present ICT's contribution to global warming in 2020 (Section 1.1.3) and then its contribution to the latter in the coming decades (Section 1.1.4).

1.1.1 Orders of ICT's Environmental Impacts

Hilty et al. [2] introduce a conceptual framework that classifies the effects of ICT on the environment into three orders or levels as presented in Table 1.

Orders of Environmental Impact	Perimeter	Positive impacts	Negative impacts
First order - direct effects of ICT	Technology	/	Environmental impacts of production, transport, use and disposal of ICT
Second order - effects of use	Application	Optimisation and substitution	Induction and obsolescence
Third order - systemic effect of ICT	Structural change	Structural and lifestyle transitions	Emerging risks and rebound effect

Table 1: Conceptual Framework for the Environmental Impacts of ICT (Hilty et al. [2])

Here, *first-order* effects are caused by the physical existence of ICT and include environmental impacts of production, use, recycling, and disposal of ICT hardware) while *second-order* effects are the indirect environmental effects of ICT caused as a result of its power to transform processes such as transport, production, or consumption, resulting in a reduction or increase in the environmental impacts of the latter. The *third-order* effects are caused as a result of medium- or long-term adaptation of behavior (E.g. consumption patterns) and economic structures to the availability of and the services provided by ICT.

[6] Wikipedia - Information and Communications Technology https://en.wikipedia.org/w/index.php?title=Information_and_communications_technology

Understanding these three categories of orders helps with the recognition of the overall impact the introduction of a technology such as AI can have. In the rest of the bibliography, it is primarily first-order impacts that are discussed because those are the ones being *partially* estimated or measured.

1.1.2 Multi-criteria Analysis of ICT's Environmental Impacts

The earth system has physical boundaries of different natures. The concept of planetary boundaries, introduced in 2009 aimed to define the environmental limits within which humanity can safely operate (Rockström et al. [3]). These 9 planetary boundaries of course include climate change.

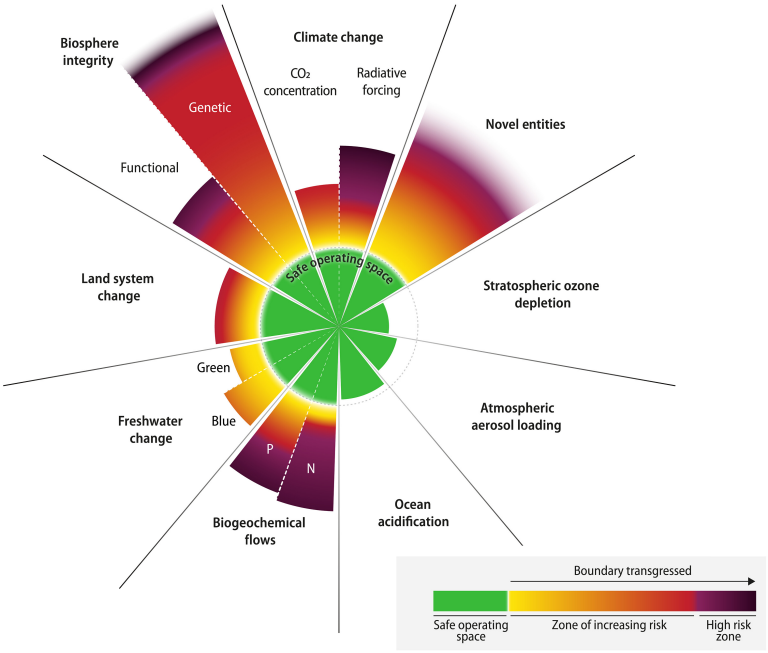


Figure 1: Current status of control variables for all nine planetary boundaries (Richardson et al. [47]).

A recent study from Richardson et al., re-assessed these planetary boundaries and out of the 9 of them 6 have been transgressed and one is close as shown in Figure 1. The green zone below the boundary is the safe operating space. The yellow to red zones represent increasing risk. Purple means a high-risk zone where inter-glacial Earth system conditions are transgressed with high confidence.

For each of the planetary boundaries, a choice is made of one of several control variables to capture the most important anthropogenic influence at the planetary level of the boundary in focus. These values for control variables are normalized in Figure 1 in an attempt to be able to compare the levels of risks between the different boundaries.

The introduction of this concept is important for grasping the necessity of measuring not only GHG emissions that primarily impact climate change but also other indicators that account for other environmental impacts and planetary boundary crossing. Moreover, by looking at only one indicator, there is a real risk of moving the environmental impacts from one boundary to another

A recent ADEME-ARCEP [30] study mandated Agence Française de l'environnement et de la maîtrise de l'énergie (ADEME) and France's "Electronic Communications, Postal and Print Media Distribution Regulatory Authority" (ARCEP) to evaluate the environmental impact of ICT in France. The study follows a rigorous methodology called Life Cycle Analysis (LCA) (described later in this bibliographic report) for evaluating multi-criteria impacts and shows that the depletion of natural abiotic resources (minerals and metals) accounts for around a quarter of ICT's overall normalized pollution in their assessment.

1.1.3 ICT's Contribution to Global Warming in 2020

Gaining growing importance in our societies, the GHG emissions of the ICT sector have become a subject of interest for researchers in the past few years. GHG emissions produced by this sector can be divided into three phases of the life cycle of the devices it is composed of :

(i) embodied emissions, which are a result of ICT object manufacture

(ii) usage or operational emissions that are caused during the use phase

(iii) end-of-life emissions, that are emitted as a result of the disposal of the object after it ceases to be used.

Since ICT devices are electrically powered, the use phase emissions are closely linked to the Carbon Intensity (CI) of the country in which the ICT object is located. The papers studied here obtain the CO_2e emissions by multiplying the energy consumption by the CI. Thus, as an example, for the same power consumption, a data center is polluting less in France than in the USA because the French energy mix is less carbonized than the American one (International Energy Agency [40]).

In a review paper about ICT's carbon footprint Freitag et al. [26], the authors have systematically studied peer-reviewed papers post-2015 assessing the environmental impacts of ICT to compare the existing models and results for the same. They have mainly focused on the work of 3 research groups: Andrae and Elder, Belkir and Elmeligi, and Malmodin and Lundén, relying both on their papers and contacting them to get additional information. The three research groups based their assessments on hybrid top-down and bottom-up approaches, following LCA methodologies. These three groups assess the first-order emissions of ICT, while Freitag et al. broadens the discussions to orders 2 and 3.

While Freitag et al. are only considering the carbon footprint of ICT and hence its contribution to global warming, one has to keep in mind that climate change contribution is just a part of ICT's environmental impacts as seen in Section 1.1.2.

Based on these works, Freitag et al. have found predicted ICT emissions for 2020 between 1.0 and $1.7GtCO_2e$, which represent between $1.8\% - 2.8\%$ of worldwide GHG emissions for this year. Freitag et al. have split the emissions into 3 domains or tiers: user devices, data centers, and networks. For the three groups of researchers, it appears that the three domains occupy a similar GHG emissions part (see Figure 2). The differences are due to different sets of scopes, databases, and hypotheses, summarized in Table 2. Freitag et al. argue that the three sets of results are questionable because of possible conflicts of interest, the outdatedness of the databases used, or the lack of transparency within the databases.

	Andrae and Elder	Belkhir and Elmeligi	Malmodin and Lundén
Estimation of ICT's impact for 2020 (% of global GHG emissions)	1.5-6.3	1.9 - 2.3	1.9
Year of the data used for predictions	2011	2008 - 2011	2015
Conflicts of interests	Working at Huawei	No obvious one	Working at Ericsson
Data transparency	Yes	Yes	No: non-reviewable confidential data, from ICT companies

Table 2: Comparison of Andrae & Elder, Belkir & Elmeligi and Malmodin & Lundén works

They also believe that these groups of researchers have underestimated the true impact due to the following reasons. First, due to the truncation error linked to the LCA methodology, it is impossible to consider all the industrial pathways of a product. Second, rebound effects (third order) are hard to measure so they are not taken into account in the reviewed papers. Third, some booming sectors such as AI, Internet of Things (IoT), and block-chain are poorly considered by the three research groups as of now. Fourth, massive investment in ICT might lead to unpredictable growth of the sector.

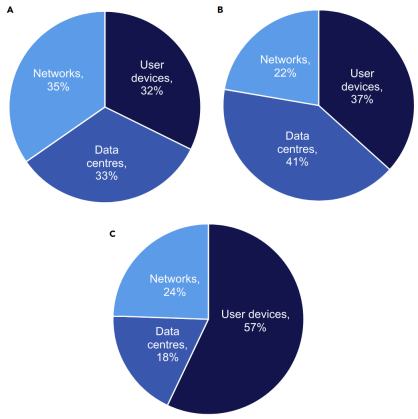


Figure 2: Repartition of ICT's 2020 GHG emissions by tier, from Freitag et al. A, B, and C are the repartitions of Andrae and Elder, Belkir and Elmeligi, and Malmodin and Lundén, respectively.

This paper also demonstrates the difficulty of setting a perimeter for the study of ICT emissions. The precise frontiers of this sector are discussed by the researchers. They all include computers, smartphones, tablets, data centers, networks and servers. IoT, blockchain, or intelligent systems in cars are not taken into account yet they enter into the definition of ICT as they rely on the same base. Since these examples have high growth potential, they can also grow ICT's emissions.

1.1.4 ICT's Contribution to Global Warming Projected for the Coming Decades

If the ICT sector followed the necessary decarbonization trajectory identical to the rest of the economy, it would have to reduce its footprint by 42% by 2030, 72% by 2040, and 91% by 2050, then reach net zero or have an equivalent effect on other sectors according to (International Telecommunication Union [19]). Meanwhile, the sector is expected to grow according to 2 out of 3 research groups. Andrae and Elder expect an exponential increase in the GHG emissions of the sector, more realistically an increase by 50% by 2030, and by 500% in their worst scenario. Belkir and Elmeligi also believe that an exponential increase is more likely, causing ICT's GHG emissions to increase 2.6-fold by 2030 and 5.1-fold by 2040. Their predictions are displayed in Figure 3.

On the contrary, Malmodin and Lundén forecast a plateau for the sector because of a saturation phenomenon. They justify that prediction with the example of smartphones: within the coming years, the majority of people will have a smartphone. Since smartphone usage is limited by the time available during the day, the emissions linked to smartphone usage should stabilize. Malmodin and Lunden argue that this trend might be true for ICT as a whole. Freitag et al. [26], argue against that by saying that ICT companies generally have a strong incentive to prevent saturation from happening. Indeed, saturation would cut their income growth. Moreover, IoT devices, that require little engagement time on the user's behalf and can operate in the background are a counterexample. IoT drives both embodied and use-phase emissions because of the production of billions of IoT devices, the networks allowing them to communicate, and the data centers that analyze the IoT data.

Also, Freitag et al. prompt the readers to carefully explore little-known ICT domains. They especially urge the scientific community to assess the impact of IoT since the number of IoT devices is expected to explode from 15 billion internet-connected devices in 2015 to 75 billion of such devices in 2025 according to Statista [53]. In parallel, AI is also expected to grow massively. Freitag et al. argue that the world's data is doubling every 2 years, that it was already 59 ZetaBytes in 2020, and that AI computational training cost is doubling every 3 months. **Closely related, AI and IoT create a risk of accelerating the environmental impacts of ICT dramatically.** Therefore, measuring AI is crucial, and more insight into the same is given in Section 2.

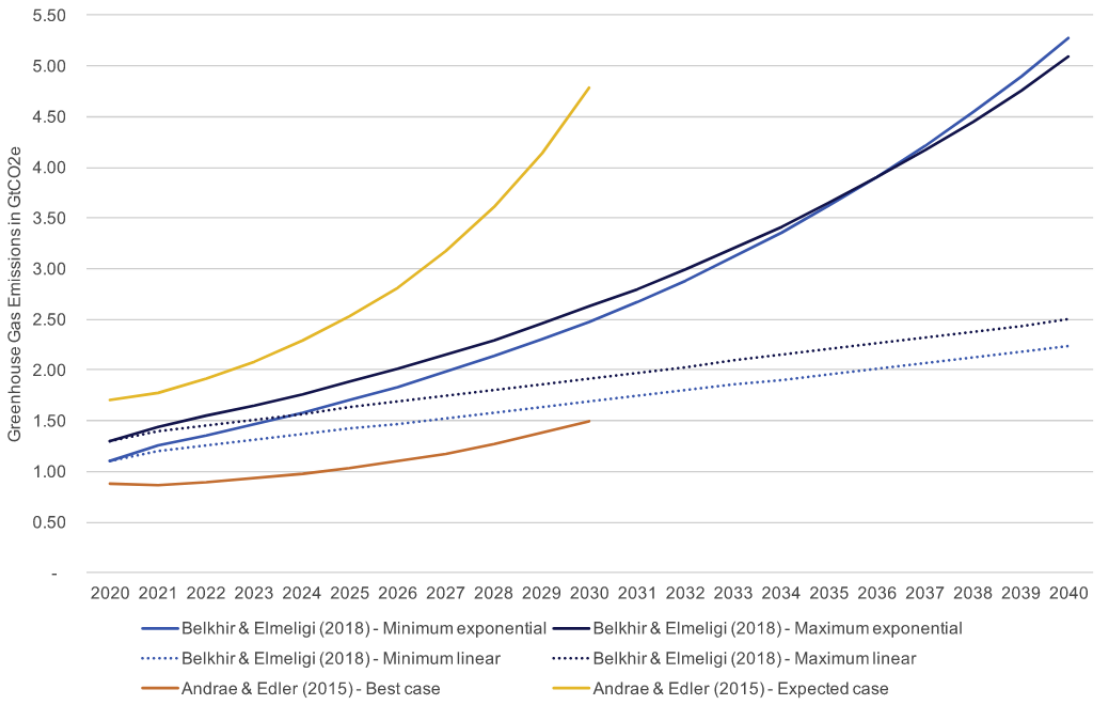


Figure 3: Projections of ICT's GHG emissions from 2020 studied by Freitag et al.

1.2 Artificial Intelligence Among ICT

To our knowledge, no environmental assessments have been made for AI as a domain yet. However, since AI is a part of the ICT sector, its impacts are included in the impacts of the latter.

Artificial Intelligence is raising environmental problems of the same nature as the rest of the ICT. It relies on more or less similar hardware (with the exception of some specialized accelerators used for training large models), with multiple significant impacts during the manufacturing and end-of-life phases, and the main concern up to now is the GHG emissions due to energy consumption during the use phase.

Nevertheless, AI is remarkable among ICT for several reasons, as it will be shown in this section: it is noticeably data-, hardware- and compute-intensive (Section 1.2.1), poorly assessed (Section 1.2.2), and getting a lot of political attention (Section 1.2.3).

1.2.1 AI Is Compute-, Data- and Hardware-Intensive

Nowadays, AI solutions are becoming more and more widespread in society. With the aim of always scoring higher on performance metrics, the AI models are growing exponentially in several respects: (i) the number of operations performed during training, (ii) the amount of data used, (iii) the energy consumption and (iv) the monetary cost. Thus, deep learning outgrows classical Machine Learning regarding all those indicators. And more recent techniques like continuous learning are even more costly (Ligozat et al. [34]).

Computation-intensiveness AI is increasingly compute-intensive. According to an article by OpenAI [11], the computational power and in turn, the energy required for deep learning research and development has been doubling every 3-4 months resulting in an approximately $300,000$ times increase between 2012 and 2018 as shown in Figure 4.

Data-intensiveness Improvements in specialized hardware and algorithms have given rise to AI models that are trained on very large data sets which, despite optimizations, have a large carbon footprint due to computation, network, and storage costs (Strubell et al. [16]).

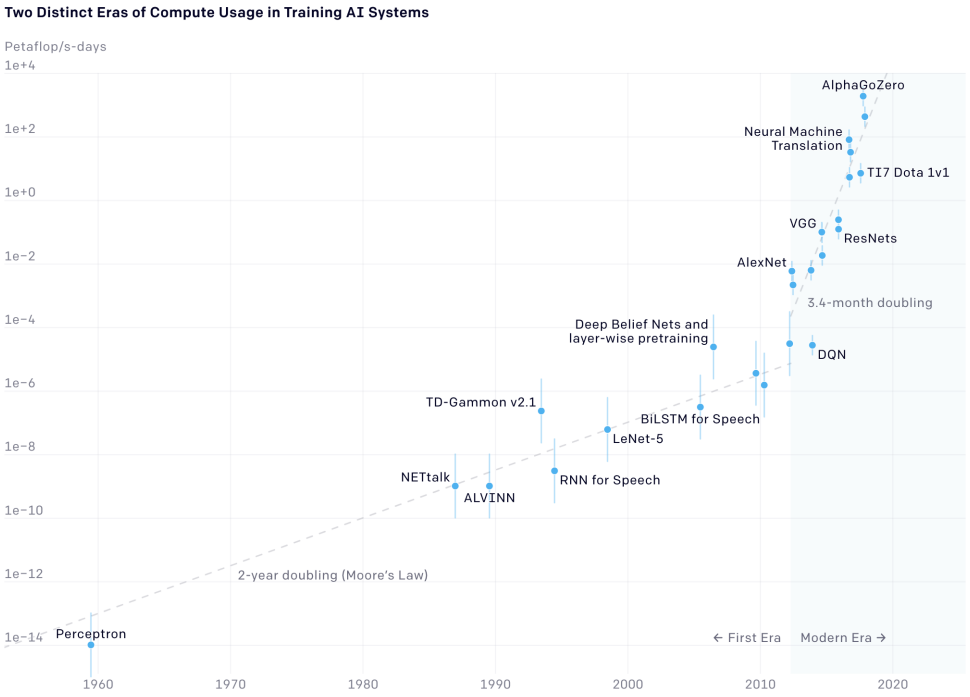


Figure 4: The total amount of compute, in petaflop/s-days used to train big relatively well-known AI models over the years, OpenAI [11]

Acquisition and transfer of such vast amounts of data have an added cost. According to a report commissioned by the European Parliamentary group of the Greens/EFA, AI relies more on hardware and data than the rest of the ICT (Benqassem et al. [25]) and this can be attributed to the above-mentioned reasons.

Hardware-intensiveness AI relies on specialized hardware such as GPUs and TPUs which are responsible for many embodied emissions. To demonstrate this point, we can consider the fact that AI is the main driver of growth in Meta's data centers (Wu et al. [38]) and Nvidia, an American multinational technology company that is a software and fabless chip-making company that designs graphics processing units, application programming interfaces for data science and high-performance computing, is experiencing huge growth this year thanks to the boom in generative AI and its underlying need of those specialized processors.

1.2.2 AI Is Poorly Assessed

Despite the efforts by a few to measure the environmental impact of AI, the state of reporting in AI research publications on carbon footprint is low, and is way lower when it comes to other environmental metrics.

Environmental reporting for AI is uncommon. In 2019, the rise of large NLP models trained on large amounts of data caught the attention of researchers on the topic which led to a publication regarding their energetic, financial, and environmental training cost by (Strubell et al. [16]). Then the introduction of the Green AI concept (Schwartz et al. [22]) and tools to measure its energy consumption and related GHG emissions. Henderson et al. [32] randomly sampled 100 NeurIPS papers issued in 2019 and found that out of these, 1 measured energy in some way, 45 measured run-time in some way, 46 provided the hardware used, 17 provided some measure of computational complexity, while none provided carbon metrics.

R&D impact is underestimated. Big AI models are usually associated with a considerable R&D cost, especially in terms of energy consumption and hence carbon footprint. Before the training of the final model, a lot of experimentation is led. Smaller models are trained to benchmark the several model components that will be used in the final version. When the final model's structure is decided, during a phase called *hyperparameter tuning*, short training sessions are performed to select a definitive set of hyperparameters ([35]). However, usually, only the efficiency of the final model is published. This leads to an underestimation of the true impact of AI research as pointed out by Strubell et al. [16].

1.2.3 Political Interest in AI

AI is becoming a priority in innovation policies worldwide. For example, AI has been getting a lot of attention from the French government in the last few years. In 2017, the French deputy and mathematician Cédric Villani led an initiative regarding the creation of a France and Europe-wide AI strategy. Out of the propositions made in this report, with the aim of pioneering innovation by 2030, France has invested 1.5 billion euros in AI between 2018 and 2021, and 2 billion euros more were scheduled for 2021-2025. In 2025, AI should produce an income of $90B for France, while it produced only $7B in 2020.

1.2.4 Paths for Taking AI's Impact into Account

As said previously, many software tools for measuring the energy consumption of AI training were introduced in 2019 following the rise of awareness of its high energy demand among ML practitioners. These tools are part of a measurement approach, which is indeed a lever for action to reduce environmental impacts. It is important to note that the carbon assessment of AI services is the most common environmental assessment, partly because it is the easiest to carry out. However, as shown in Section 1.1, merely measuring the carbon footprint of the training phase of an AI lacks completeness. So, other ways of taking into account the environmental impacts of AI need to be explored.

Ligozat et al. [28] add proposals for reducing the environmental footprint of AI. Their paper is targeted towards the community of machine learning practitioners. It suggests going beyond simply measuring the impact of training AI models. Their suggestions are divided into two categories: (i) measures to be taken as a practitioner and (ii) measures to be taken as an institution. These measures are listed below as they are in the paper :

As a practitioner (ordered by impact) :

- ***Reduce your I/O and redundant computation/data copying/storage:*** *Start with smaller datasets to debug your model, and use shared data storage with members of your team so you don't need to have individual copies.*
- ***Choose a low-carbon data center:*** *When running models on the cloud, consult a tool like Electricity-Map to choose the least carbon-intensive data center.*
- ***Avoid wasted resources:*** *by steering clear of grid search and by reusing or fine-tuning previously trained models when possible. Also, strive towards designing your training and experimentation to minimize discarded computing time and resources in case of failure.*
- ***Quantify and disclose your emissions:*** *use packages like CodeCarbon, Carbon tracker, and Experiment impact tracker, which can be included in your code at runtime or online tools like Green algorithms and ML CO2 Impact that can allow you to estimate your emissions afterward. In both cases, share these figures with your community to help establish benchmarks and track progress!*

As an institution :

- ***Deploy your computation in low-carbon regions*** *when possible.*
- ***Provide institutional tools*** *for tracking emissions and enable them by default on your computing infrastructure.*
- ***Cap computational usage*** *at say a maximum of 72 hours per process, in order to reduce wasted resources.*

- ***Carry out awareness campaigns*** *regarding the environmental impact of ML.*
- ***Facilitate institutional offsets*** *for those emissions that cannot be avoided, such as commuting and building construction.*

In addition, some researchers want to integrate environmental considerations into the design of AI projects. For example, Lefèvre et al. [44] propose an environmental assessment framework document intended for responses to calls for projects involving Artificial Intelligence (AI) methods. It helps to take into account various environmental criteria, in particular, the general impacts of digital services but also the specificities of the AI field (impacts of the learning and inference phases, data collection, etc.). Moreover, the former is the result of an explicit request from the French Ministry of Ecological Transition to the EcoInfo research group, which brings together engineers and researchers from the research and higher education sectors in France to work towards the common goal of reducing the (negative) environmental and societal impacts of AI.

The document contains a questionnaire that asks the respondents to consider in the proposal the impact of the digital equipment over its entire life cycle, the justification for the proposed method along with the environmental impact of the behavioral, economic, and societal changes brought about by the proposal. The questionnaire is accompanied by an explanatory note on these topics. This framework document provides an excellent summary of the aspects to be taken into account when assessing the various types of impacts of the proposal for a new AI service.

For example, measuring and publishing the carbon emissions of training AI models is a first step towards taking account of the environmental effects of AI. Indeed, energy measurement methods and tools for estimating carbon emissions are important. However, in order to be exhaustive, the impacts of AI must be considered in a multi-criteria and multi-order manner (direct, indirect, and societal impacts), but this is not the most widespread paradigm as of now.

2 Quantifying the Environmental Impact of AI

The booming evolution of AI described in Section 1.2 makes AI environmental impacts seminal to study and quantify. Currently, existing methods for measuring the impact of AI mainly focus on measuring energy consumption and the carbon footprint of machine learning algorithms during the training as well as inference stages. However, the carbon footprints of training and inference respectively only make up a small part of AI's overall impact on the environment. As explained for ICT in Section 1.1.2, the environmental impact of AI is not restricted to carbon emissions. Metrics such as abiotic resource depletion, water consumption, as well as the impact of biodiversity, need to be studied alongside the carbon footprint even though these are harder to measure due to the lack of specific methodology and data.

In this section, an LCA methodology for AI is presented (Section 2.1), followed by ways of measuring (Section 2.2) and predicting carbon footprint (Section 2.3). Finally, some insight is given on the water footprint of AI (Section 2.4).

2.1 Life Cycle Analysis for AI

The LCA methodology built by ETSI/ITU is a way to exhaustively assess the environmental impacts of ICT(International Telecommunication Union [7]).

LCA is the most widely recognized methodology for environmental impact assessment. It is a standardized method with ISO standards: 14040 and 14044. This methodology covers different steps of the life cycle of the target system (product or service) and quantifies multiple environmental criteria. Since a single impact category cannot result in the proper evaluation of a product, multiple impact categories need to be studied. ADEME, the French national environmental agency, counts 13 environmental indicators [7], including impacts on air, water, earth's resources, as well as human health, and emphasizes the importance of being exhaustive about these indicators for environmental impact assessment.

The ITU standard states that out of the various existing impact categories, climate change resulting from high energy consumption is an important category that is hence mandatory. Nevertheless, they mention that it is important to consider other impact categories such as ozone depletion, human toxicity, ionizing radiation, eutrophication, acidification, land use, and resource depletion (water, mineral, or fossil). However, there is no consensus on the criteria to take into account within the LCA community. So, it is up to practitioners to decide which impact categories are relevant based on the ICT product system being studied and its purpose.

Ligozat et al. [34] propose a methodology for applying LCA to AI services that relies on the specific methodology standard of International Telecommunication Union [7]. They illustrate the insufficiency of the current reporting with Figure 5. A proper LCA of an AI service should include production, use, and end-of-life impacts, and those impacts should cover not only global warming potential but also water usage, human toxicity, and abiotic depletion potential. In the literature, some papers reporting the carbon footprint of training AI models have been published lately. Using carbon footprint as the only metric is however far from exhaustive for measuring the entire environmental impact of an AI.

A big mistake that is made when it comes AI is the narrative around its *immateriality*. As Ligozat et al. show in their paper, each task of an AI service relies on dedicated hardware, and each of them should be the subject of a LCA if one wants to exhaustively assess the impacts of an AI service. This is illustrated in Figure 6.

Nonetheless, it is hard today to find publicly available data regarding the environmental impacts of hardware used for AI such as GPUs. Popular Life Cycle Inventory (LCI) databases such as NegaOctet for ICT or EcoInvent that cover a lot of general domains cost several thousands of euros. Thus, Ligozat et al. suggest that the AI community could lobby companies to open more of their data so that AI practitioners can start reporting the environmental impacts of their models in an exhaustive manner.

[7] Indicators suggested by ADEME for LCA: https://expertises.ademe.fr/economie-circulaire/consommer-autrement/elements-contexte/dossier/impacts-indicateurs/indicateurs-dimpacts-couramment-utilises-lanalyse-cycle-vie-lair

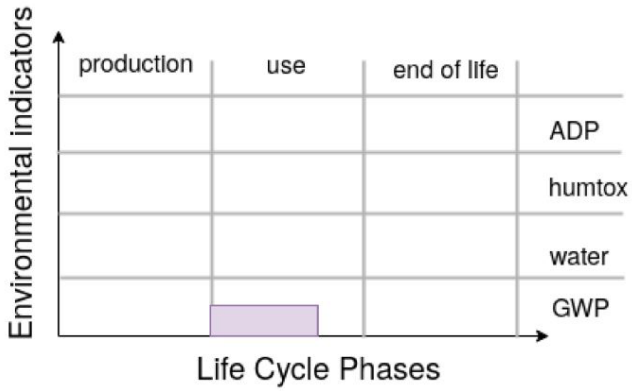


Figure 5: LCA dimensions: the first dimension corresponds to the phases of the life cycle and the second one to the environmental impacts (Ligozat et al., 2021). The purple rectangle shows what is typically assessed for AI services when something is assessed at all

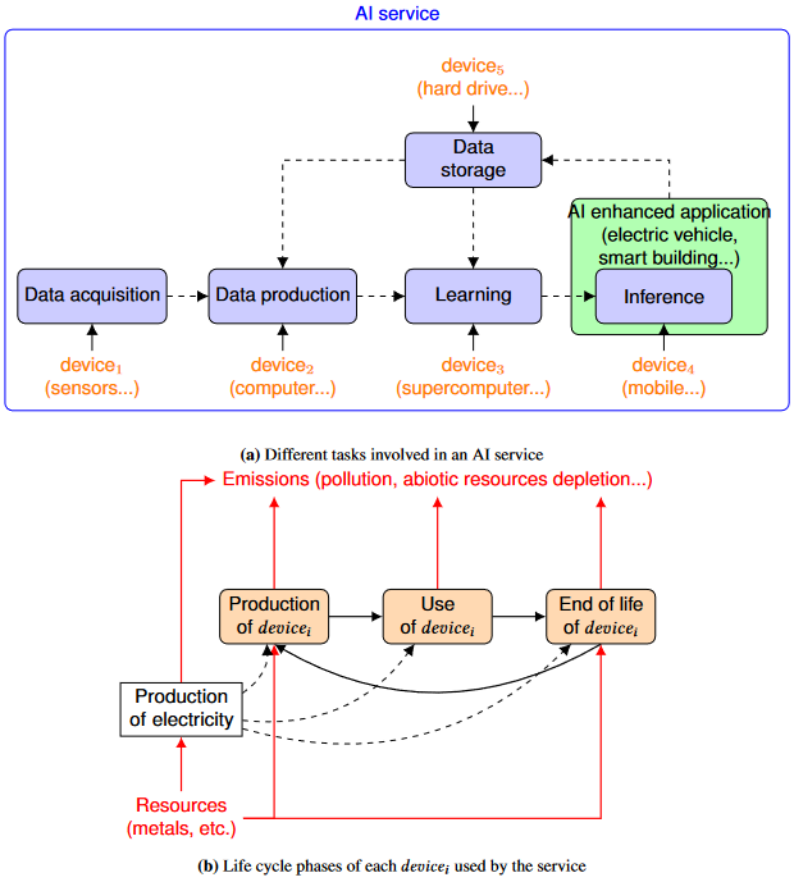


Figure 6: Life Cycle Inventory of an AI service

2.2 Measuring Energy Consumption & Carbon Footprint of AI Computation

Energy consumption is usually what has been looked at in papers when it comes to AI's environmental impact since the key paper from Strubell et al. [16] raised awareness on this topic. AI services can have a high environmental impact in terms of GHG emissions because of the substantial energy consumption of the computational facilities used to develop and train them.

Calculating and estimating the carbon footprint of AI projects systematically can help raise awareness, encourage the development of energy-efficient software, and limit the waste of resources. (Lannelongue et al. [43]).

In this section, we first explain the link between energy consumption and GHG emissions for an AI task (Section 2.2.1), and finally we present two categories of tools to measure AI energy consumption: hardware power meters (Section 2.2.2) and software power meters (Section 2.2.3).

2.2.1 Relationship Between Energy Consumption and Greenhouse Gas Emissions

Most of the biggest AI models are trained in data centers. On the other hand, smaller AI models can be trained on personal computers. This section focuses on how training an AI in a data center leads to GHG emissions. The training is done on hardware specialized in parallel computation (when available) -GPUs or TPUs - benefiting from the fact that training computations are highly parallelizable. Such types of hardware consume significant amounts of electricity. For example, the NVIDIA Tesla A100 GPU power consumption can reach 300 W, which is about the power consumption of a plasma television. Additionally, training requires a CPU (or several) and some Dynamic Random Access Memory (DRAM), though their consumption is often way lower than the one of GPUs.

Performing computation in data centers involves an energetic overhead in addition to the energy consumed for the computation itself. For example, cooling the server on which the computation is performed is energy-consuming. To take this into account, Power Usage Effectiveness (PUE) is generally used.

The PUE, which is an efficiency metric for data centers, describes how efficiently a computer data center uses energy. It compares the whole energy consumption to the energy consumption dedicated to computation. Its formula is given in Equation 1.

Note: Contrary to what its name suggests, this metric is actually a ratio between two energies and not powers.

$$PUE = \frac{TotalFacilityEnergy}{ITEquipementEnergy} \tag{1}$$

According to Brady et al. [5], the energies used in Equation 1 should be obtained by tracking the consumption of the data center over a year. This is because the consumption of the total facility varies over time. For example, it is supposedly lower during winter, during which the cooling can be slowed down.

The closer the PUE tends to 1, the more optimized the use of the computing resources. Nevertheless, as stated in Brady et al. [5], it is important to note that a good PUE does not necessarily depict eco-friendliness. Indeed, increasing the computation load while keeping other energetic costs fixed diminishes the PUE while increasing the data center's energy consumption. In other words, even if the data center's computing consumption is very high if the surrounding costs are optimized, the data center may have a PUE value close to 1.

For the rest, PUE on being multiplied with the energy of computation gives the energy consumption of all the devices in a data center that enable a model's training. This can be done like in Equation 2 :

$$E_{training} = PUE \times (E_{DRAM} + E_{CPU} + E_{GPU}) \tag{2}$$

Finally, the energy consumed during training is linked to GHG emissions through an indicator called carbon intensity (CI). It describes how polluting the electricity powering the data center is in terms of GHG emissions with coal-based electricity being more polluting than wind energy, for example.

GHG emissions, energy consumption, and CI can be linked as in Equation 3. This equation was introduced in a paper by Strubell et al. [16].

$$m_{GHG\_training} = E_{training} \times CI \tag{3}$$

2.2.2 Methods for Measuring the Energy Consumption of Computer Nodes

As described by Jay et al. [41], the energy consumption of a running program can be measured in four different ways:

- external devices (Wattmeters and Power Distribution Unit (PDU)),
- intra-node devices (Baseboard Management Controller (BMC))
- hardware sensors and software interfaces (Running Average Power Limit (RAPL), NVIDIA System Management Interface (NVIDIA-SMI), and
- power and energy modeling (based on power with the Thermal Design Power (TDP) or based on usage).

One way to access power information is through physical power meters which are external devices that are hence not embedded in computational nodes and measure the entire power consumption of the node. They are usually placed at the interface between the power socket and the power supply unit. These power meters can be wattmeters or PDU with measuring capabilities. Some energy measuring devices can also be placed inside computing nodes such as a BMC, which is a small and specialized processor used for remote monitoring and management of the host system it is placed on. It is placed between a computing node's power supply and the main board. It can give information at the component level(CPU, GPU, DRAM, etc.).

These two above-mentioned categories of devices are known to have little overhead on the computation, but they require a financial investment, and depending on the model they can have different performances regarding accuracy and measuring rate (Jay et al. [41]).

Some CPU and GPU vendors provide tools to track energy consumption. Intel and AMD for instance provide an RAPL interface each (intel RAPL, AMD RAPL). This interface, among other things, enables access to the value of energy consumed since the processor was started. It gives this information for different power domains that are physically meaningful like DRAM or the entire CPU socket. On the other hand, NVIDIA GPUs provide information about its energy consumption through the NVIDIA-SMI. Strubell et al. query these interfaces to obtain the average power draw from the hardware used for training the four NLP models being studied.

Lastly, when these previously cited measurements are imprecise or unavailable, one can resort to approximations of energy consumption. This can, for example, be done with a constant average power value or a proportion of usage of the measured device.

2.2.3 Software Power Meters

The energy model used by Strubell et al. [16] (also described in Section 2.2.1) for estimating the energy needed to train large NLP models is also implemented in some open-source tools available as Python packages:

- CodeCarbon (Sasha et al. [21])
- Carbon Tracker (Anthony et al. [17])
- Experiment Impact Tracker (Henderson et al. [32])
- Cumulator (Trebaol et al. [23])
- Eco2AI (Budennyy et al. [31]).

These tools use mainly the RAPL and NVIDIA's NVML library or TDP to make estimates.

Although the above-mentioned tools were developed especially for AI workloads (machine learning, NLP, deep learning), they work for other types of computation too.

Since these tools offer software-based ways to measure the energy consumed during the execution of a code snippet, they are referred to as "software power meters" in the paper by Jay et al. [42].

Works like that of Heguerte et al. [39], Jay et al. [41], and Bannour et al. [24] have looked into these tools and tried to compare them. Heguerte et al. [39] analyze seven tools for estimating energy

consumption for training a deep learning model. They have tested this list of tools on neural networks for image processing. Bannour et al. [24], on the other hand, systematically reviewed tools that were freely available, usable in their programming environment (Mac/Linux terminal), documented in a scientific publication, and suitable for measuring the impact of NLP experiments while providing a CO_2e measure for experiments.

Even though these tools have the same goal, there are significant differences between them that are highlighted in the previously presented papers. These differences include:

- the OS and hardware compatibility
- the release date
- the open source license
- the carbon intensity used
- the multiplication by a PUE factor or not, the value of PUE used and if it's configurable or not
- the value of carbon intensity used and if it is real-time or not
- the way the results are presented
- the use of usage factors
- the frequency of measurements
- if the code has errors that require code modification
- the tooling used to measure CPU GPU, and Memory consumption.

Tracarbon and Cloud Carbon Footprint can also be cited but they aren't documented in a scientific publication. They may have specificities that have been overlooked here.

Other tools measure energy without supplying a CO_2e measure. It is what Jay et al. [41] call in a broader sense *software power meters*. The previous tools of this section are included in this study among others.

Figure 7 is an imperfect way to show all of the software tools that enable energy consumption monitoring of computer nodes in different settings including AI.

2.3 Predicting Energy Consumption & Carbon Footprint of AI Compute

In order to have some control over one's GHG emissions, one may want to predict how much an AI task will emit before running it. Several methods to do so have been developed, focusing on different targets of the AI workflow.

Prediction from partial measurement Carbon Tracker (Anthony et al. [17]) simply predicts total energy consumption by averaging the energy consumption of previous epochs.

TDP approximation Some simple tools available online like ML CO2 Impact (Alexandre et al. [14]) and Green Algorithms (Lannelongue et al. [27]) fulfil this goal. They use TDP as an approximation for the power consumption of hardware. It is a hardware-specific value provided by the manufacturer and that represents the maximum amount of heat generated by the component under a steady workload. If the amount of time required for the job is known, the total processor energy consumption can be estimated as follows:

$$E_{total} = TDP \times t \tag{4}$$

There are 2 issues with this equation. One, in most cases one doesn't know the execution time before running the program, so it limits the predictive capacity of those tools. Two, it doesn't grasp the effects of the executed code's nature of power consumption. Some steps of the execution might be computationally intensive while others might let the processor almost idle, so there is no guarantee on how good the TDP approximation is.

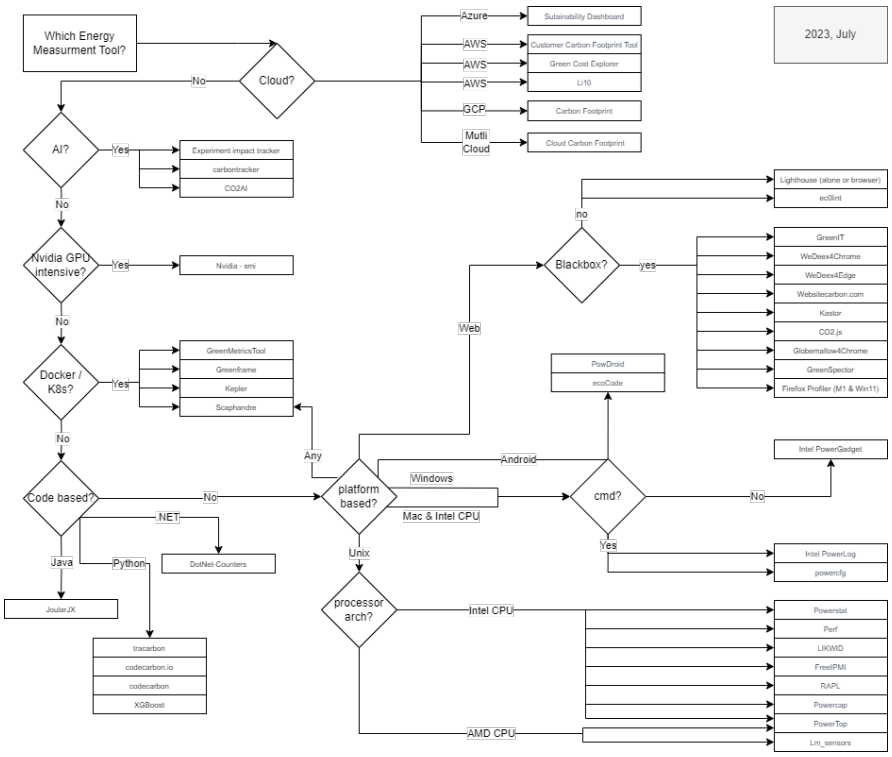


Figure 7: Decision tree for choosing an energy consumption measurement software (schaDev Github)

FLOPs as a metric of complexity There is a need to find a better metric than time to describe the complexity of a piece of code. Schwartz et al. [22] propose instead the Floating Point Operation (FLOP)s. Roughly, they are the number of operations that need to be performed by the hardware to execute a piece of code. It is computed analytically by counting two base operations, `ADD` and `MUL`. They have some interesting advantages. They can be computed by parsing the code analytically, without executing it. For example, in Equation 5, the FLOPs count is equal to 2, due to one $+$ and one $\times$ operator. A variation of FLOPs that can be seen in the literature is called Multiply Accumulate Operations (MACs). It has the same idea of operations counting but in Equation 5, the MACs count would be 1. It fits the behavior of some hardware, that is able to perform both the multiplication and the accumulation at once.

$$x \leftarrow x + y \times z \tag{5}$$

Also, the number of FLOPs is agnostic to the hardware on which the code is used. Thus they are a metric of choice to describe the workload of a piece of code. In particular, it is possible to compute the number of FLOPs necessary for a AI model inference.

Predicting Inference Consumption With FLOPs Rodrigues et al. [12] lead experiments to link FLOPs to the energy consumption of Convolutional Neural Networks (CNN) for inference. They pick a set of 9 CNN models. Within this set, they run inferences on an Nvidia Jetson TX1 while measuring the energy consumption, the number of bus accesses, the number of Single Instruction Multiple Data (SIMD), and the number of MACs.

Afterward, they try to find relationships between those parameters. They first find a linear relationship between the number of bus accesses and the energy consumption for an inference. Then, they find a second linear relationship between the number of bus accesses and the number of SIMD. Next, another linear relationship is found between SIMD and MACs. Thanks to these linear relationships, Rodrigues et al. create a linear predictor from the 3 linear relationships, using solely the number of MACs as input, to predict the energy consumption. Their linear regression shows satisfying results, with a relative error of $7.08 \pm 5.05\%$.

Nonetheless, it is worth noting that the relationships and predictor depend on the hardware used: running the same model on different hardware doesn't consume equally. Moreover, they didn't try to see if those results were generalizable to other models. Provided that the results can be generalized, their conclusion is that it is possible to predict the energy consumption of an unknown CNN inference on the NVIDIA Jetson TX1 architecture.

Predicting training consumption with FLOPs Mehta et al. [20] focus on predicting the energy consumption of training a Deep Neural Network. To predict the energy of training, they make the following assumption:

$$Energy^{Training} \approx T.(Energy_{FP} + Energy_{BP}) \tag{6}$$

where:

- T is the number of epochs
- $Energy_{FP}$ is the energy required for a full forward pass
- $Energy_{BP}$ is the energy required for a full backward pass

They characterize the energy consumption of a forward pass and backward pass as follows:

$$Energy_{FP}^{comp} \approx b_s(\alpha_{flop} \sum_{s=1}^{S-1} n^{(s)} n^{(s+1)} + \alpha_{act} \sum_{s=1}^{S} n^{(s)}) \tag{7}$$

$$Energy_{BP}^{comp} \approx b_s(2\alpha_{flop} \sum_{s=1}^{S-1} n^{(s)} n^{(s+1)} + \alpha_{err} \sum_{s=1}^{S} n^{(s)}) \tag{8}$$

where:

- b_s is the batch size used for training
- α_{flop} is the energy cost of Floating Point Operations on this hardware, which can be obtained by running a micro-benchmark program
- α_{act} is the energy cost of the activation functions used (ReLU, Tanh, etc.), also obtained with a micro-benchmark program
- α_{err} is the energy cost of computing the propagation error in a neuron (micro-benchmark program)
- $n^{(s)}$ is the number of neurons in the layer s.

They run small programs on the NVIDIA Tegra K1 development kit to measure the energy consumption of the different kinds of operations of training. They count the number of each operation, weigh them with the associated energetic cost, and sum the energetic costs together. Similar to Rodrigues et al. [12], they propose a method to evaluate energy consumption without running the whole training. A limitation of this work is that the model created to predict the consumption only works on a given hardware.

2.4 Measuring Water Consumption & Water Footprint of AI

Water consumption is an important environmental impact of AI which is rarely reported in peer-reviewed papers. According to the United Nations World Water Development Report (UN-Water & United Nations World Water Assessment Programme [13]), nearly 6 billion people will suffer from clean water scarcity by 2050. Boretti et al. [15] suggest that this number may be higher owing to not only the drivers of water scarcity like population (economic growth and water demand), resources, and pollution but also unequal growth accessibility and needs, which are underrated.

Water is often used in data centers for cooling servers and for maintaining humidity levels. According to Holzle [33], data centers that are water-cooled use approximately 10% less energy. Hence, they have

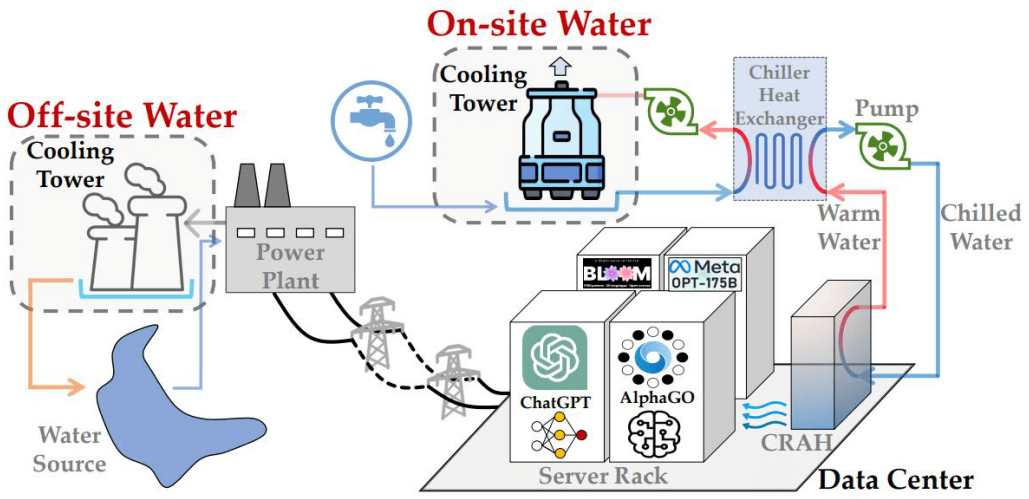


Figure 8: Data center water footprint: on-site water consumption for data center cooling, and off-site water consumption for electricity generation. Li et al. [45]

approximately 10% less carbon emissions as compared to many air-cooled data centers. This explains in part the trend of switching from air-based to water-based cooling in data centers.

Earlier works by Google (Holzle [33]), Microsoft (Nakagawa [46]), AT&T [49], and Meta [50] address methods for improving on-site water efficiency (including the efficiency of cooling towers) without taking into account the spatial and temporal variation of cooling demands for training or usage phases of AI models.

While methods like LCA take into account the embodied water footprint of AI models (for example, that for manufacturing hardware on which an AI model is trained), Li et al. [45] take into account the operational water footprint of AI models associated with only training and inference. The water footprint of AI models, as estimated by the latter, depends on the type of cooling system used in a data center, the water consumed in producing the electricity used to run the AI programs on the servers, the time of the day during which one runs the AI programs as well as the physical location of the data center. For example, according to Li et al., training GPT-3 (Brown et al. [18]) in Microsoft's U.S. data centers can consume 700,000 liters of clean freshwater, while training in Microsoft's Asian data centers means triple the water consumption.[8]

The concept of Water Usage Effectiveness (WUE) used by Li et al. [45] considers both the

- **On-site water footprint, W_{on} :** The authors considered the following factors for on-site water footprint calculation for a cooling tower which is a type of cooling method in data centers where water is circulated to extract waste heat from a system and then it is dissipated into the atmosphere mainly through evaporation:
  - temperature approach settings (i.e., the difference between the cold water temperature and entering wet bulb temperature)
  - cycles of concentrations (i.e., water recirculation times before *blowndown* [9])
  - wet bulb temperature (lowest possible temperature attained by adiabatic evaporation of water into the air until it is saturated)
  - water flow rate

[8] Without data from the model developer, Li et al. took $WUE = 0.55L/kWh$ for training in Microsoft's U.S. data centers, and $WUE = 1.65L/kWh$ for its Asian data centers based on Microsoft-reported average WUE. [37]

[9] Blowdown: When water evaporates from a cooling tower, the concentration of the solids in the remaining water becomes too high, so some of this water is removed and replaced with fresh water. This is called blowdown

  - air pressure
  - humidity, etc.

- **Off-site water footprint, W_{off} :** It is related to water consumed in electricity production. It varies with time due to variations in energy fuel mixes [8] [4]

The authors consider a time-slotted model $t = 1, 2, , T$ where each time slot can be from 10 minutes depending on the frequency at which the water footprint is to be assessed. The *total water footprint* of the AI model is given by Equation 9.

$$W = W_{on} + W_{off} = \sum_{t=1}^{T} e_t \cdot WUE_{on,t} + \sum_{t=1}^{T} e_t \cdot PUE_t \cdot WUE_{off,t} \tag{9}$$

Where

- e_t is the AI model's energy consumption at time t as measured by power meters or by the server's built-in tools
- $WUE_{on,t}$ is the on-site WUE at time t
- $WUE_{off,t}$ is the off-site WUE at time t
- PUE_t is the power usage effectiveness of the data center hosting the AI model at time t
- T is the total length of interest (e.g., training phase, total inference phase, or their combination)

Li et al. use Google's large language model, LaMDA (Thoppilan et al. [36]), to demonstrate their methodology for water footprint estimation. They also underline the need to address the water footprint along with the carbon footprint for creating more sustainable AI. In addition, they suggest paying attention to the trade-off between carbon intensity and WUE, both varying with time and location. For example, by only considering carbon efficiency for scheduling a training session for a model, AI developers may want to choose a slot around noon due to the abundance of solar energy at that time. However, high temperatures during such time slots lead to the worst water efficiency (Islam et al. [10]).

While the approach by Li et al. [45] provides us a fine-grained assessment of the water consumption of cooling towers, it is important to remember that while cooling towers are common, more water-efficient cooling solutions need to be implemented on a large scale across data centers to reduce water-wastage and help reduce AI's contribution to water crises in the future.

3 Reviewing Some Case Studies of AI Environmental Impact Assessment

In this section, the environmental impacts of some popular AI case studies are discussed: LLMs in Section 3.1 and *AI for Green* in Section 3.2. The aim is to help the readers grasp some orders of magnitude of the impact of such models on the environment, as well as gain an insight into how are they assessed in the literature.

3.1 Large Language Models

BLOOM Luccioni et al. [35] contains the most complete model assessment we could find in the literature. It tries to take into account all the dimensions of environmental impacts that have been described earlier with an LCA approach. Sadly, only GHG emissions are looked at. The steps of the LCA that were accounted for were equipment manufacturing, model training, and model deployment. Due to the lack of information, the impacts of steps like raw material extraction, hardware disposal, and end-of-life could not be taken into account. The model assessed in this article is BLOOM, an LLM especially used for text summarization. BLOOM was trained on the Jean Zay computing cluster.

Although the authors were not able to measure the real-time consumption for training BLOOM, they empirically noticed that GPU consumption was very high during the training, so they used the TDP approximation, mentioned in Section 2.3. They also neglected the CPU consumption whose value is typically 40 times lower than that of a GPU. By multiplying the 1 million hours GPU-time of the training by the TDP and the number of GPUs, they obtain 25 tCO_2eq. This consumption purely linked to the training computation is called dynamic consumption. For an exhaustive assessment, this has to be summed with idle consumption (which is the power consumed by servers that are powered on but not in use) and embodied consumption (which are the emissions associated with the materials and processes involved in manufacturing a given product, such as the computing equipment needed to train and deploy ML models). The authors also estimate the idle consumption by running experiments directly on the Jean Zay cluster. Finally, while the NVIDIA A100 GPUs used for BLOOM training have not been assessed in a LCA by the manufacturers, they perform an LCA for the closest possible hardware to include embodied emissions. The results are presented in Figure 9.

Process	**CO_2emissions (CO_2eq)**	**Percentage of total emissions**
Embodied emissions	11.2 tonnes	22.2 %
Dynamic consumption	24.69 tonnes	48.9 %
Idle consumption	14.6 tonnes	28.9 %
Total	**50.5 tonnes**	**100.00 %**

Figure 9: Breakdown of CO_2eq emissions from different sources of the BLOOM model life cycle (Luccioni et al., 2022)

Luccioni et al. also ran an experiment to calculate the carbon footprint of running inferences on BLOOM. They booked an instance composed of 16 NVIDIA A100 40GB GPUs for 18 days on the Google Cloud Platform with an inference Application Programming Interface (API). They tracked the consumption on the instance with the tool CodeCarbon, presented in Section 2.2.3. Over the 18 days, for this use case, they found GHG emissions of 340 $kgCO_2eq$. It is negligible compared to the training consumption, but this 18-day experiment is a single-use case of BLOOM deployment. During the duration of the experiment, the API was requested 230k times, but that amount might be different in another use case. Moreover, it excludes the consumption of the equipment used to query this API and that of the network.

Thus, the information provided by the above-mentioned paper is insufficient for determining whether it is training or inference that consumes the most energy since in the case of large-scale applications, the number of requests is much higher. As an example, ChatGPT has been visited 1.5 billion times in August 2023 according to SimilarWeb[10]. Unfortunately, OpenAI has neither disclosed the energetic

[10]Number of visits on ChatGPT login page in August 2023: https://www.similarweb.com/website/chat.openai.com/#overview

cost of a request to their model nor aggregated values of how much their whole service is consuming. In many cases, for widely used models, the question of their exact environmental impact is impossible to answer due to the lack of data.

Google's LLMs In Patterson et al. [29], researchers from Google and Berkeley estimate the energy costs and related GHG emissions for some of the latest large language models from 2021: T5, Meena, Gshard, Switch Transformer and GPT3. This is what they understand as the environmental impact of a large neural network: emissions resulting from the electricity used for powering the hardware used for training the best model. They are grateful to the works of Strubell et al. [16], Schwartz et al. [22], Alexandre et al. [14], and Henderson et al. [32] for raising awareness on this important issue and hence the importance of research on this topic.

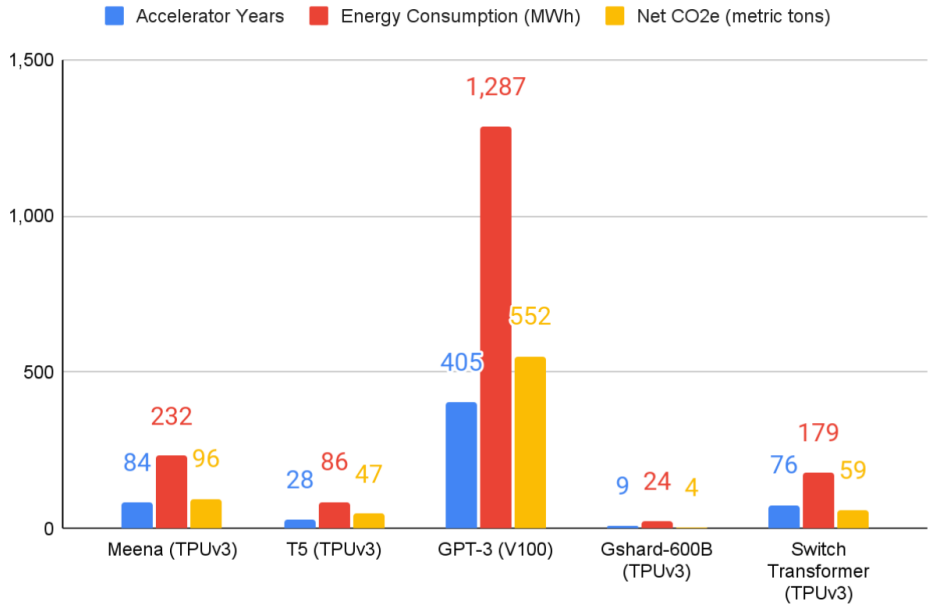


Figure 10: Accelerator years of computation, energy consumption, and CO_2e for five large LLMs (Patterson et al. [29]).

This paper gives orders of magnitude of the training costs of such models as displayed in Figure 10. Using Google Flights[11] they also compare those results to the emissions of a direct round trip of a whole passenger jet between San-Francisco and New York which is 180 tCO_2e. And show that T5 training emissions are 26%, Meena is 53%, Gshard-600B is 2%, Switch Transformer is 32%, and GPT-3 is 305% of such a round trip respectively.

We want to underline that the carbon footprints of the training of these biggest models are both enormous and relatively small. Enormous, because only by doing computation, those models reach an energy consumption of 1 GWh. On average, a French household consumes around 5,000 kWh over a year [12], so training the largest models have a consumption equivalent to 200 French households over a year. But also small, with the plane trip comparison, while 28 million planes have flown in 2022 [13]. And there aren't that many big AI models, because of the resources needed to train them.

But again, the values are reported only for training the best model and not for all the research & development. GHG emissions are reported only for the training phase and not for the other tasks that are necessary for building the AI service and for keeping it running. The GHG emissions are only reported for the use of some devices as a part of the AI service. No other categories of impacts are reported in this paper.

[11] Google Flights: https://support.google.com/travel/answer/9671620?hl=en&ref_topic=2475360

[12] Consumption of a French household: https://particuliers.engie.fr/electricite/conseils-electricite/conseils-tarifs-electricite/consommation-moyenne-electricite-personne.html

[13] Number of planes in 2022, Statista: https://www.statista.com/statistics/564769/airline-industry-number-of-flights/

In a further paper from the same group of researchers (Patterson et al. [51]), they acknowledge the existence of life-cycle emissions (which additionally includes the embedded carbon emissions resulting from the manufacturing of all components involved, from chips to data-center buildings.) alongside operational emissions (the energy cost of operating the ML hardware including data-center overheads) and they state that estimating life-cycle emissions is a larger, future study for them.

In the paper, they then highlight opportunities to improve energy efficiency and hence CO_2e emissions of such models such as model optimization, judicious choice of the data center location, and use of more energy-efficient infrastructures for training.

Meta's LLMs Meta (previously Facebook) disclosed emissions for some of their most used models. In Wu et al. [38], they assessed the emissions from their LM and RM class of models, used in various Meta products, for translation and recommendation tasks respectively. Their results are in Figures 11 and 12.

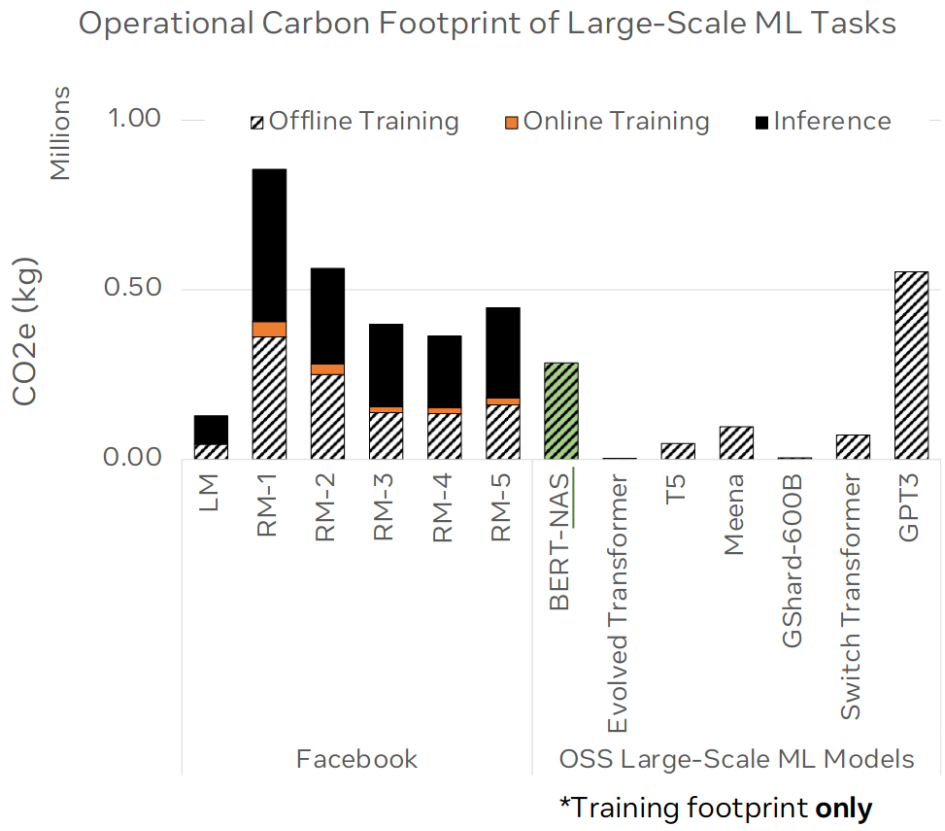


Figure 11: Carbon footprint of Facebook models versus Open source models. Comparison between inference and training cost

In Figure 11, training versus inference emissions are compared. The models from Meta are compared to other well-known AI models' training emissions. The inference footprint has been assessed from the deployment of those models to a date preceding the publication of the paper. Of course, the inference footprint keeps growing with the daily use of the model. At the end of this assessment, inference and training have almost the same share in the operational footprint. As a comparison, the carbon footprint of RM-1, of about 800 tCO_2eq emitted corresponds on average to the GHG emissions of 1 person traveling 3.5 million km by plane, according to ImpactCO2 calculator [14] (600 times the distance Paris - New-York). It is worth noting that those figures are only the operational footprint of the models, that is to say, the emissions of the *use-phase* only.

[14] ImpactCO2: https://impactco2.fr/

Wu et al. also disclosed the embodied footprint of these models. Embodied footprint attributes the GHG emissions of the manufacture of the hardware used to run the models over the lifetime of the AI service. This increases the models' footprint by around 60%.

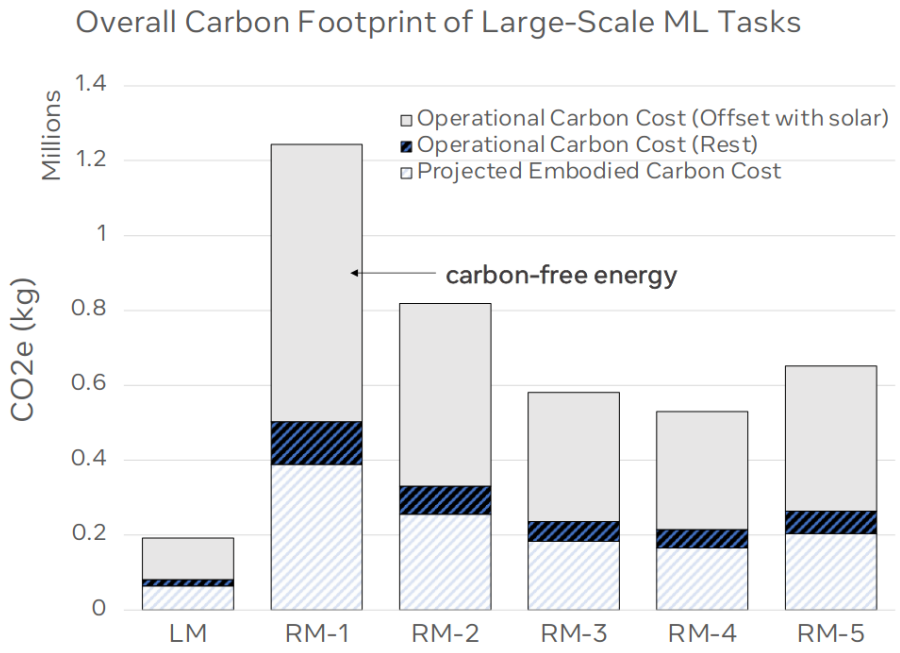


Figure 12: Carbon footprint of Facebook models over the manufacturing and use phase of the life cycle

The carbon footprint of training BLOOM (25 tCO_2eq) is noticeably smaller than the one for training LLM models of comparable size (502 tCO_2e for GPT-3). This can be explained by the structure of the models themselves, but also the relatively low carbon intensity used for BLOOM (57gCO_2eq/kWh) compared to the other carbon intensities that can be up to 8 times higher 429 gCO_2e/kWh for GPT-3).

From here comes the intuitive idea of training the models within data centers powered by electricity with low carbon intensity. But this may have side effects on the electricity grid: if the public renewable installations of a country are used to power the data centers instead of powering the rest of the electrical grid, then the country's carbon intensity increases. In such a case, there are no overall gains in terms of GHG, it's rather only a carbon accountability shift away from the data center owners.

3.2 AI for Green

AI for Green applications are AI-based solutions that aim at having positive environmental impacts. "Smart" AI-based heating can be an example of such a solution since it aims at reducing the electricity consumption of a building.

Ligozat et al. underline that to know the environmental impact of such a solution, one has to compare the LCA of the whole service - the building enhancement in the previous example - with the AI with the LCA of the same building without the AI service. An AI solution can only be said to have a positive effect on the environment if its benefits outweigh its cost.

Surprisingly, such a comparison is almost never done by *AI for Green* solution developers. Rolnick et al. [52] suggest a list of AI-based solutions to deal with climate change. They describe how ML could be a powerful tool in reducing greenhouse gas emissions and helping society adapt to a changing climate. Out of these, Ligozat et al. review the 57 applications labeled as *High-leverage*, which supposedly have the most positive impact on the environment. Ligozat et al. have categorized the levels of evaluation or reporting for these applications as follows:

(a) No mention of the environmental gain

(b) General mention of the environmental gain

(c) A few words about the environmental gain but no quantitative evaluation or only indirect estimation

(d) Evaluation of the energy gain without taking the AI service into account

(e) Evaluation of the energy gain taking the use phase of the AI service into account

(f) Comprehensive evaluation of the environmental gain (comparison of LCAs).

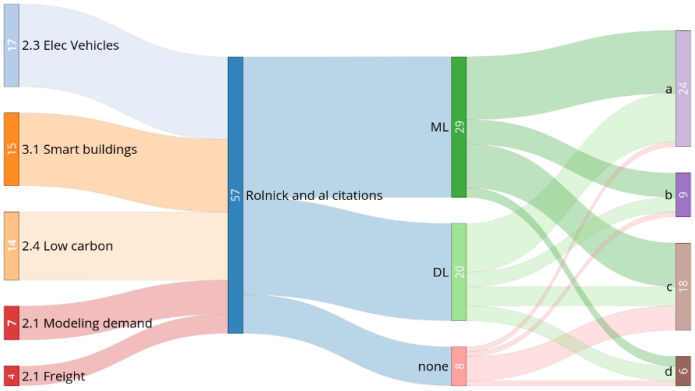


Figure 13: Sankey diagram of parts of Rolnick's paper references in terms of environmental evaluation (Ligozat et al. [34])

However, out of those 57 applications, not a single one was properly evaluated: none of them were categorized as (f) or (e). The evaluation of the energy gain was assessed for only 6 of them (category (d)), and for the others, the energetic gains were put into words, but never quantitatively evaluated (category (c) or below). The category-wise distribution of the 57 applications is shown in Figure 13.

Thus, Ligozat et al. conclude that the environmental benefits of these AIs built for *green* applications are *only potential*. They need to be proven properly, and must not be taken for granted.

Finally, the authors push the developers of such solutions to think carefully about the rebound effects and the societal changes that might be implied by the wide adoption of *AI for Green* solutions.

Conclusion

The respective environmental impacts of ICT and AI, the existing methodologies to assess the environmental impact of AI, especially its energy consumption, and some AI case studies have been discussed in this report. In addition to these topics, many orders of magnitude have been provided to help contextualize these impacts in terms of global ecological questions.

We want to underline the following :

1. There is no understanding of the environmental impacts of AI as a sector. However, these are included in those of the ICT sector since AI is a part of it.
2. Carbon emissions from the ICT sector are estimated to contribute to between 2 % and 4% of global emissions (Freitag et al. [26]).
3. The ICT sector, like any other sector, must reduce its impact if it is to comply with the Paris Agreement (Intergovernmental Panel on Climate Change (IPCC) [6]). At present, the ICT sector is facing the risk of exponential growth in its carbon emissions despite constant optimization. Massive and concerted efforts including large-scale actions by politicians and industry, will be needed to reduce these emissions (Freitag et al. [26]).
4. Optimisation and the search for energy efficiency cannot be the only solutions to GHG emissions because of the rebound effect (Freitag et al. [26]). This effect means that when technical improvements increase the efficiency of usage of a resource, it creates new uses, resulting in an increase in the overall consumption of the resource in question, rather than a decrease. This partly explains the growth in ICT emissions, despite several optimizations.
5. AI, combined with big data, data science, and IoT, is worsening ICT's carbon emissions as it creates a growing need for data, which must be processed and stored by powerful hardware (Freitag et al. [26]).
6. The carbon footprint is far from being the only impact on the environment. The depletion of natural abiotic resources (minerals and metals) accounts for around a quarter of digital pollution in the assessment made by (ADEME-ARCEP [30]) for France. This justifies the use of a multi-criteria approach [15] to assess the environmental impact of digital services.
7. It is important to assess the three orders of impacts - the direct, indirect, and systemic impacts - of AI services (Hilty et al. [2]). The direct impacts must be studied exhaustively using recognized, standardized, and multi-criteria environmental assessment methodologies such as the LCA (Ligozat et al. [34]).
8. In the literature, the relative shares of environmental impacts of the different tasks involved in an AI service (data collection, production, storage, training, and inference) are not well-known, and the same is true for the shares of environmental impacts for the different stages of its life cycle (manufacturing, transport, use and end of life). One rare example is the AI models used in various Meta products. For this example, operational emissions from training are as important to take into account as those from inference. And emissions from hardware manufacturing are as important as operational emissions (Wu et al. [38]). So, reducing the carbon footprint of a model to the emissions resulting from the energy consumed during its training is insufficient.
9. The environmental impacts of *AI for Green* services, aimed at using AI to solve environmental problems, are insufficiently evaluated, and their real benefits remain to be proven. (Ligozat et al. [34]).

In many respects, most papers in the literature review consider the impact of digital technology (ICT) in general rather than focusing on AI. This is because AI raises environmental problems of the same nature as the rest of ICT. Yet, a few elements set AI apart from the rest of digital technology in the three orders of impacts as it has been shown in this report.

[15] For example, depletion of abiotic resources (fossil fuels, minerals, and metals), acidification, ecotoxicity, carbon footprint, ionizing radiation, fine particle emissions, ozone creation, raw materials, waste production, primary energy consumption, final energy consumption

Most importantly, AI brings a lot of crucial ethical questions, not addressed in this report, since it is reshaping our relationships with a lot of human domains like information, knowledge, work, entertainment, intellectual property, or even art. These changes further legitimize studying AI's impacts as a standalone domain.

References

[1] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[2] Hilty and Lorenz. *2008 Hilty ICT and Sustainability Chapters 7*. Apr. 2008.

[3] Johan Rockström et al. "A safe operating space for humanity". In: *Nature* 461.7263 (Sept. 2009), pp. 472–475. ISSN: 1476-4687. DOI: 10.1038/461472a. URL: https://doi.org/10.1038/461472a.

[4] Peter Gao, Andrew Curtis, Bernard Wong, and Srinivasan Keshav. "It's Not Easy Being Green". In: *ACM SIGCOMM Computer Communication Review* 42 (Sept. 2012), pp. 211–222. DOI: 10.1145/2377677.2377719.

[5] Gemma Brady, Niraj Kapur, Jon Summers, and Harvey Thompson. "A case study and critical assessment in calculating power usage effectiveness for a data centre". In: *Energy Conversion and Management* 76 (Dec. 2013), pp. 155–161. DOI: 10.1016/j.enconman.2013.07.035.

[6] Intergovernmental Panel on Climate Change (IPCC). "Climate change 2014 - Synthesis Report". en. In: (2014).

[7] International Telecommunication Union. "Methodology for Environmental Life Cycle Assessments of Information and Communication Technology Goods, Networks and Services". In: (Dec. 2014), p. 202.

[8] Mohammad Islam, Kishwar Ahmed, Shaolei Ren, and Gang Quan. "Exploiting Temporal Diversity of Water Efficiency to Make Data Center Less "Thirsty"". In: June 2014.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Dec. 5, 2017. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762[cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 05/02/2023).

[10] Mohammad A. Islam, Kishwar Ahmed, Hong Xu, Nguyen H. Tran, Gang Quan, and Shaolei Ren. "Exploiting Spatio-Temporal Diversity for Water Saving in Geo-Distributed Data Centers". In: *IEEE Transactions on Cloud Computing* 6.3 (2018), pp. 734–746. DOI: 10.1109/TCC.2016.2535201.

[11] OpenAI. *AI and compute*. May 2018. URL: https://openai.com/research/ai-and-compute.

[12] Crefeda Rodrigues, Graham Riley, and Mikel Luján. *SyNERGY: An energy measurement and prediction framework for Convolutional Neural Networks on Jetson TX1*. Aug. 2018. DOI: 10.13140/RG.2.2.36489.54881.

[13] UN-Water & United Nations World Water Assessment Programme. *Nature-based Solutions for Water 2018: The United Nations World Water Development Report 2018*. Accessed: 2023-07-18. 2018. URL: https://wedocs.unep.org/20.500.11822/32857.

[14] Lacoste Alexandre, Luccioni Alexandra, Schmidt Victor, and Dandres Thomas. *Quantifying the Carbon Emissions of Machine Learning*. 2019. arXiv: 1910.09700 [cs.CY].

[15] Alberto Boretti and Lorenzo Rosa. "Reassessing the projections of the World Water Development Report". In: *npj Clean Water* 2 (Dec. 2019). DOI: 10.1038/s41545-019-0039-9.

[16] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *CoRR* abs/1906.02243 (2019). arXiv: 1906.02243. URL: http://arxiv.org/abs/1906.02243.

[17] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models". In: *CoRR* abs/2007.03051 (2020). arXiv: 2007.03051. URL: https://arxiv.org/abs/2007.03051.

[18] T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared Kaplan. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[19] International Telecommunication Union. "Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement". en. In: (2020).

[20] Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. "DeLighT: Very Deep and Light-weight Transformer". In: *CoRR* abs/2008.00623 (2020). arXiv: 2008.00623. URL: https://arxiv.org/abs/2008.00623.

[21] Luccioni Sasha, Mallet Tristan, Friedler Sorelle, Laskaris Nicolas, Schmidt Victor, and Goyal Kamal. *Code Carbon: Track and reduce CO2 emissions from your computing*. https://codecarbon.io/. Accessed: 2023-04-26. 2020.

[22] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI". In: *Communications of the ACM* 63.12 (Nov. 17, 2020), pp. 54–63. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3381831. URL: https://dl.acm.org/doi/10.1145/3381831.

[23] Tristan Trebaol, Mary-Anne Hartley, Martin Jaggi, and H.S. Ghadikolaei. "A tool to quantify and report the carbon footprint of machine learning computations and communication in academia and healthcare". In: *Infoscience EPFL: record 278189* (2020).

[24] Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. "Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools". In: (Nov. 2021), pp. 11–21. DOI: 10.18653/v1/2021.sustainlp-1.2. URL: https://aclanthology.org/2021.sustainlp-1.2 (visited on 04/13/2023).

[25] Sofia Benqassem, Frederic Bordage, Lorraine de Montenay, Julie Delmas-Orgelet, Firmin Domon, Etienne Lees Perasso, Damien Prunel, and Caroline Vateau. *Behind the figures: understanding the environmental impacts of ICT and taking action*. 2021. URL: https://www.apl-datacenter.com/wp-content/uploads/2021/12/Case-studies-Environnemental-impacts-of-ICT-and-Digital-technologies-7-dec-2021.pdf.

[26] Charlotte Freitag, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon Blair, and Adrian Friday. *The climate impact of ICT: A review of estimates, trends and regulations*. 2021. arXiv: 2102.02622 [physics.soc-ph].

[27] Loic Lannelongue, Jason Grealey, and Michael Inouye. "Green Algorithms: Quantifying the Carbon Footprint of Computation". In: *Advanced Science* 8.12 (2021), p. 2100707. DOI: https://doi.org/10.1002/advs.202100707. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202100707. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202100707.

[28] Anne-Laure Ligozat and Sasha Luccioni. "A Practical Guide to Quantifying Carbon Emissions for Machine Learning researchers and practitioners". In: (2021). URL: https://hal.science/hal-03376391/document.

[29] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, L.M. Munguia, Daniel Rothchild, and David So. *Carbon Emissions and Large Neural Network Training*. 2021. arXiv: 2104.10350 [cs.LG].

[30] ADEME-ARCEP. *Evaluation de l'impact environnemental du numérique en France et analyse prospective - Note de synthèse réalisée par l'ADEME et l'Arcep*. 2022.

[31] Semen Budennyy et al. *Eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI*. 2022. arXiv: 2208.00406 [cs.LG].

[32] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. *Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning*. 2022. arXiv: 2002.05651 [cs.CY].

[33] Urs Holzle. *Our commitment to climate-conscious data center cooling*. Accessed: 2023-04-18. 2022. URL: https://blog.google/outreach-initiatives/sustainability/%20our-commitment-to-climate-conscious-data-center-cooling/.

[34] Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. "Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions". In: *Sustainability* 14.9 (2022). ISSN: 2071-1050. DOI: 10.3390/su14095172. URL: https://www.mdpi.com/2071-1050/14/9/5172.

[35] A.S. Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model*. 2022. arXiv: 2211.02001 [cs.LG].

[36] Romal Thoppilan et al. *LaMDA: Language Models for Dialog Applications*. 2022. arXiv: 2201.08239 [cs.CL].

[37] Noelle Walsh. *How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability*. Accessed: 2023-10-18. 2022. URL: https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/.

[38] Carole-Jean Wu et al. *Sustainable AI: Environmental Implications, Challenges and Opportunities*. Number: arXiv:2111.00364. Jan. 9, 2022. arXiv: 2111.00364[cs]. URL: http://arxiv.org/abs/2111.00364.

[39] Lucia Bouza Heguerte, Aurélie Bugeau, and Loïc Lannelongue. "How to estimate carbon footprint when training deep learning models? A guide and review". In: arXiv:2306.08323 (June 14, 2023).

DOI: 10.48550/arXiv.2306.08323. arXiv: 2306.08323[cs]. URL: http://arxiv.org/abs/2306.08323 (visited on 06/27/2023).

[40] International Energy Agency. en-GB. 2023. URL: https://www.iea.org/reports/world-energy-balances-overview/world.

[41] Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel. "An experimental comparison of software-based power meters: focus on CPU and GPU". In: (2023).

[42] Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel. "An experimental comparison of software-based power meters: focus on CPU and GPU". In: *CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing*. Bangalore, India: IEEE, 2023, pp. 1–13. URL: https://inria.hal.science/hal-04030223.

[43] Loic Lannelongue and Michael Inouye. "Carbon footprint estimation for computational research". In: (2023). DOI: 10.17863/CAM.93287. URL: https://www.repository.cam.ac.uk/handle/1810/345865.

[44] Laurent Lefèvre et al. "Environmental assessment of projects involving AI methods". working paper or preprint. 2023. URL: https://hal.science/hal-03922093.

[45] Peng Li, Jianyi Yang, Mohammad Atiqul Islam, and Shaolei Ren. "Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models". In: *ArXiv* abs/2304.03271 (2023).

[46] Melanie Nakagawa. *The journey to water positive*. 2023. URL: https://blogs.microsoft.com/on-the-issues/2023/03/22/water-positive-climate-resilience-open-call/.

[47] Katherine Richardson et al. "Earth beyond six of nine planetary boundaries". In: *Science Advances* 9.37 (2023), eadh2458. DOI: 10.1126/sciadv.adh2458. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adh2458. URL: https://www.science.org/doi/abs/10.1126/sciadv.adh2458.

[48] Wikipedia. *Information and communications technology*. Page Version ID: 1172797080. Aug. 29, 2023. URL: https://en.wikipedia.org/w/index.php?title=Information_and_communications_technology&oldid=1172797080#cite_note-2.

[49] AT&T. *Water Management*. URL: https://about.att.com/csr/home/reporting/issue-brief/water-management.html.

[50] Meta. *Sustainability — water*. URL: https://sustainability.fb.com/water/.

[51] David Patterson et al. *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*.

[52] David Rolnick et al. "Tackling Climate Change with Machine Learning". In: *ACM Computing Surveys* 55.2 (), pp. 1–96. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3485128. URL: https://dl.acm.org/doi/10.1145/3485128 (visited on 09/14/2023).

[53] Statista. *IoT devices installed base worldwide 2015-2025*. en. URL: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ (visited on 07/19/2023).