Evaluation of the environmental impacts of Natural Language Processing methods

Clément Morand

Master of Science in Informatics at Grenoble
Master Informatique
Specialization Data Sciences and Artificial Intelligence (DSAI)

Evaluation of the environmental impacts of Natural Language Processing methods

Clément Morand

June the 26th, 2023

Research project conducted at LISN

Under the supervision of:
Anne-Laure Ligozat & Aurélie Névéol

Defended before a jury composed of:
Massih-Reza Amini
Emilie Devijver
Gaël Guennebaud
Lorraine Goeuriot
Anne-Laure Ligozat & Aurélie Névéol

June 2023

Abstract

Faced with environmental challenges, we should question the ever-increasing use of digital technology. Thus, tools such as Green Algorithms or Carbontracker, offering studies of the environmental impact of calculations such as the training of a machine learning model, where developed with the aim of creating a 'green AI' research. However, these tools are only focused on the dynamic consumption induced by a calculation and only evaluate its carbon footprint. Other types of impacts such as the resources consumption are not evaluated. In order to remedy this shortcoming, we propose a multi-criteria estimation tool taking into account the production impacts of the equipment used to perform these calculations. We also study the use of digital technology by healthcare and in particular automatic language processing tools used in the hospitals. This study takes the first steps towards impact measurements at the scale of this field. Indeed, the French healthcare system is undergoing an important digital transition including an important use of AI. It is also the motivation for a lot of current computer science research.

Acknowledgement

I would like to express my sincere gratitude to Anne-Laure and Aurélie for this internship and for their invaluable assistance and comments in reviewing this report. I would also like to thank Ludmila for her support throughout this internship and the very fruitful discussions.

Résumé

Face aux enjeux environnementaux, nous nous devons de questionner l'utilisation toujours croissante du numérique. Hors, les outils comme Green Algorithms ou Carbontracker, proposant des études d'impact environnementaux des calculs tels l'entraînement d'un modèle de machine learning, ne s'intéressent qu'à la consommation dynamique induite par un calcul et n'évaluent que son empreinte carbone. Ces outils ne prennent pas en compte d'autres types d'impacts tels que l'épuisement des ressources naturelles. Das le but de combler ce manquement, nous développons un outil d'estimation multicritère tenant compte des impacts de production du matériel utilisé pour réaliser ces calculs. Nous nous intéressons aussi à l'utilisation du numérique par la santé et en particulier des outils de traitement automatique des langues par la recherche en santé, dans l'optique de parvenir à des mesures d'impacts à l'échelle de ce domaine. En effet, la santé est à la fois un domaine qui connaît une rapide transition numérique, en incluant l'utilisation d'IA ; et un secteur applicatif aux cœur des motivations dans la recherche en informatique actuelle.

Contents

Acronyms

ADP Abiotic ressources Depletion Potential. 11, 12, 15, 32, 33

AI Artificial Intelligence. 1, 2, 3, 4, 5, 6, 7, 41, 45

APHP Assistance Publique des Hopitaux de Paris. 41, 44

CHU Centre Hospitalier Universitaire. 44

CI Carbon Intensity. 11, 14, 19, 30, 31, 32

EDS Entrepôt de Données de Santé. iv, 6, 40, 41, 42, 43, 45, 46, 47

GWP Global Warming Potential. 10, 12, 14, 15, 32, 33, 34, 38

HPC High Performance Computing. 36, 43

ICT Information and Communication Techonlogies. v, 1, 2, 4, 5, 6, 7, 11, 39, 40, 41, 43, 44, 45, 46, 49

LCA Life Cycle Assessment. 1, 2, 3, 4, 5, 9, 10, 12, 15, 17, 33, 34, 35, 36, 44, 49

NLP Natural Language Processing. 1, 2, 6, 9, 12, 39, 40, 41, 42, 44, 46, 47, 49

NVML NVIDIA Management Library. 3, 17

PE Primary Energy. 10, 12, 15

PUE Power Usage Efficiency. 11, 13, 14, 19, 27, 29, 30, 31, 32

RAPL Running Average Power Limit. 3, 17

TDP Thermal Design Power. 4, 10, 11, 13, 19, 20, 22, 25, 29, 30

— 1 —
Introduction

In the last 20 years, the usage and carbon footprint of ICT has grown continuously with a trajectory that could place them as a sector as energy consuming as civil aviation for instance ([Gupta et al., 2020, Freitag et al., 2021]). Machine learning is now taking a big part in ICT and it has been shown that training models can have an important carbon footprint [Strubell et al., 2019, Luccioni et al., 2022].

Facing the ever increasing computation demand of Artificial Intelligence (AI), researchers call for the development of a "Green AI" [Schwartz et al., 2020]. In order to understand and measure the impacts of AI, and with the aim to promote a more frugal AI research, a number of tools were developed.

However, most of the research effort is concentrated exclusively on the impacts of the training phase of models and only takes into account the energy consumption induced by training the model, not the impacts of producing the hardware needed to run these computations. These hardware manufacturing impacts are far from negligible: [Wu et al., 2022] shows that they account for 30% of the total impacts of "large scale ML tasks" at Facebook. Furthermore, with the shift towards less Carbon Intensive electricity, the share of embodied impacts is bound to grow.

A methodology accounting for the impacts over the whole life cycle (from production to use and end of life) of products is the Life Cycle Assessment (LCA). This methodology is widely recognised with ISO standards (ISO 14040 and 14044). [Ligozat et al., 2021] provides keys on how to adapt this methodology for measuring impacts of AI solutions.

LCA has its limitations, such as the complexity of applying it to ICT because of limited data availability [Ligozat et al., 2021, Clément et al., 2020] or the fact that it isn't able to capture structural or societal impacts such as rebound effect (or Jevons' Paradox) [Kaack et al., 2021, Rasoldier et al., 2022]. Other societal impacts include ethics questions [Bender et al., 2021].

Still, LCA offers a broad view of impacts by taking into account not only the direct impacts and offering to include multiple criterion for determining impacts, such as resource depletion, and not only the carbon footprint.

In this work, we tackle the questions of the evaluation of the environmental impacts of Natural Language Processing (NLP) methods. How can we accurately evaluate the environmental impacts of a series of NLP Experiments? How can we incorporate LCA considerations in a tool conducting such evaluations? What are the impacts of NLP methods used in Epidemiology and Public Health? To what extent does Healthcare policy making currently take environmental impact into consideration?

Indeed, the French Healthcare system is undergoing an important digital transition and the development of clinical data warehouses in numerous French hospitals [Jannot et al., 2017] encourages the development and use of digital tools and NLP in general because the textual content of the patients health records becomes more accessible. This textual content is of great interest since it has been shown that it contains quantity of information that is not recorded in the existing structured data [Escudié et al., 2017].

In this work, we present a tool named *MLCA*[1] (for Machine Learning life Cycle Assessment) aiming at providing researchers with LCA estimates before undertaking new AI projects to help them put in perspective the expected benefits of their project (may it be in terms of accuracy gains for instance) with the expected impacts and decide if it is worth following that path.

This tool is then evaluated in a series of case studies showing the quality and usability of our tool.

Finally, the first steps towards a domain wide evaluation of the impacts of NLP are taken through a series of interviews in order to obtain a wide-view of ICT use in Healthcare and the different questions it poses.

The main contributions of this work are as follows:

1. An estimation tool named *MLCA*
2. An evaluation of the usability and quality of this tool
3. An overview of ICT usage in the French Healthcare system, its probable evolution, the current state of reflection on the impact of these usages, and, how this reflection on the impacts of the ICT impacts the use and investment towards new tools.

First, Chapter 2 presents the state of the art and related work, then Chapter 3 details the methodology used for creating MLCA. Chapter 4 then evaluates MLCA on a series of case studies before Chapter 5 opens towards a domain wide evaluation. Finally Chapter 6 concludes and discusses.

[1] available at `https://github.com/blubrom/MLCA`

— 2 —
State of the Art

In this Chapter, we present the important concepts to understand this study and related work in the area. First, Section 2.1 introduces existing work on carbon footprint estimation for deep learning and in particular NLP programs and highlights the need for the integration of LCA considerations in the evaluation of the environmental impacts of carrying computation. Then, Section 2.2 details LCA and ways to adapt this framework to evaluate the impacts of computer programs. Subsequently, Section 2.3 highlights the need to think of impacts in a global context to foresee rebound effects and also to ensure adequacy with global sustainability targets. Finally, Section 2.4 discusses the use of ICT and NLP in particular in health, considering the ethics and environmental questions that may be raised.

2.1 Carbon footprint of Deep Learning

[Strubell et al., 2019] was the first paper to present the high level of carbon emissions associated with training Natural Language processing models. [Schwartz et al., 2020] followed, calling for the development of a systematic reporting of the costs associated with training models and a "green AI" research field. [Verdecchia et al., 2023] presents a review of the current state of "green AI" research.

In addition, [Parcollet and Ravanelli, 2021] and [Dinarelli et al., 2022] argue that models should not be evaluated only on raw performance but also on efficiency, proposes a method to evaluate the increase in carbon footprint per percentage of gain in precision. For instance [Parcollet and Ravanelli, 2021] shows that for one of their experiment, half of the costs of state-of-the-art speech system are used to gain .3% Word Error Rate.

In order to allow researchers to evaluate their models in terms of impacts, a number of tools have been developed such as CarbonTracker [Anthony et al., 2020], CodeCarbon [Schmidt et al., 2022], Experiment-Impact-Tracker [Henderson et al., 2020], Green Algorithms [Lannelongue et al., 2020] or ML CO_2 Impact [Lacoste et al., 2019].

[Jay et al., 2023, Bannour et al., 2021] survey these different existing tools and review the strengths and weaknesses of each tools. There exists two main categories of tools. On the one hand, there are measurement tools such as [Anthony et al., 2020, Schmidt et al., 2022] that use the Running Average Power Limit (RAPL) tool to obtain live values about the CPU and DRAM consumption [David et al., 2010] and the NVIDIA Management Library (NVML) [NVIDIA Corporation, 2021] to get live consumption values for the GPU. On the other hand, there are modeling-based tools such as [Lannelongue et al., 2020, Lacoste et al., 2019] that use

the Thermal Design Power (TDP) of processing units to obtain an estimate of the consumption. With its *Pragmating Scaling Factor*, Green algrithms encourages its users to reflect on the production process of the computer program being evaluated and the potential multiple experiments needed to tune hyper-parameters of a model or to find the Neural architecture of said model.

All of these tools focus on giving the carbon footprint of the energy consumed to run a computer program. However, this approach does not take into account the manufacturing impacts of the hardware used to run the program. Furthermore, producing results about the Carbon Footprint only risks incentivising the user to decrease the greenhouse gas emissions by shifting impacts. This could for instance happen when replacing working but older hardware with new and more energy efficient one. The energy efficiency gains could reduce the carbon footprint, but discarding functional hardware increases the production of e-waste and manufacturing new hardware increases the consumption of resources.

2.2 Life Cycle Assessment

A methodology accounting for the impacts over the whole life cycle (from production to use and end-of-life) of products is the Life Cycle Analysis (LCA). This methodology is widely recognised with ISO standards (ISO 14040 and 14044). The reference handbook on LCA is [European Commission et al., 2010]. LCA is a multi-criteria evaluation of the environmental impacts (common metrics or criteria include Global Warming Potential, Human Toxicity, Resource depletion, Water use, Land use, Marine eutrophication, ...). In order to avoid impacts shifting (reducing impacts from one category of impacts, typically carbon footprint, by increasing the impacts in one or multiple other categories), it is important to have a multi-criteria approach. Figure 3.1 presents different phases of the life cycle of an item and impacts each can have.

Two types of LCA can be performed: *Attributional* LCA and *Consequential* LCA. Attributional LCA is aimed at explaining the different potential impacts that can be attributed to a particular product system. It is set in a static environment. For instance it could try to answer the question: "What are the impacts of transporting 10,000 people a day by bus over 15 km?" Oppositely, Consequential LCA looks at the impacts of change in a dynamic environment, with possible macro-economic responses to a change. For instance, a Consequential LCA could try to answer the following question: "Given the current bus network, what would be the impacts of adding 1000 new passengers a day?"

All the tools presented in section 2.1 try to conduct an attributional analysis (i.e. They try to answer the question: What is the carbon footprint of training a ML model?) when they in fact conduct a short term consequential analysis (They give an answer to the question: Given the existing data-center infrastructure, what is the change in consumption caused by training a ML model).

[Ligozat et al., 2021] proposes a framework for adapting the attributional LCA approach to evaluate AI solutions. To evaluate an AI solution, one must not only consider the energy consumption induced by the training phase of the model but also the hardware manufacturing needed to produce the server on which this training phase takes place. Ideally, this analysis would also include information about the end-of-life of the hardware used but, because of the lack of available data on ICT end-of-life, it is a complex task. One should also not only evaluate the impacts of producing an AI solution (i.e. training the model) but also consider the other life

cycle phases of the model such as the data collection required, the architecture search and the inference phase.

As pointed out in [Luccioni et al., 2022], the inference phase, even if it has a unitary cost way lower than the training phase, can have a major importance when numerous inferences are run, which is the expected case for large models such as Large Language Models. For instance, [Wu et al., 2022] shows that, for a number of Language Models deployed at Facebook, the energy consumption of the inferences is at least as high as the energy consumption for training the models.

Also pointed out in [Luccioni et al., 2022] and [Gupta et al., 2020] is the importance of the manufacturing impacts of the hardware used. These manufacturing impacts can account for 40% of the total impacts of a server over its whole life cycle, as such, one could expect that they account for a similar share of the impacts attributable to training an AI model. In fact, [Wu et al., 2022] states that, for Facebook, embodied impacts represent 30% of the total impacts of "large scale ML tasks".

[Boavizta, 2021] proposes a tool for simplified LCA of servers[1] based on [Gröger et al., 2021], the follow up study to [Schödwell et al., 2018]. These studies aim at creating a methodology for evaluating the environmental performance of Cloud services based on LCA methodology. One important missing component of this tool and methodology is the fact that it does not account for GPU being present in servers, although for most computationally heavy tasks such as training machine learning models, GPU or TPU are used to perform these tasks in a manageable time-frame.

Indeed, one important difficulty when applying an LCA approach in the ICT is the lack of available quality data [Clément et al., 2020]. This is especially the case when looking at GPUs or TPUs where no manufacturing firm provides insights on the manufacturing impacts of these devices. [Loubet et al., 2023] conducted an LCA for comparing the use of desktop computers with raspbery PI devices with centralised servers in the context of a higher education class. This LCA is, to our knowledge, the only available LCA analysis taking into account a GPU.

2.3 Limits of applying Life Cycle Assessment for the Information and Communication Technologies

Conducting an attributional LCA does not allow to explore all of the possible impacts of the considered solution. Indeed, it cannot explore the social consequences and ethical challenges posed by a solution. Nor can it explore the changes induced by introducing the new solution such as rebound effect [2].

Ethical challenges of AI are discussed in [Bender et al., 2021]. [COMETS, Comité d'éthique du CNRS, 2022] discusses the role of research and the need for researchers to reflect on the possible impacts their work can have on society. [Gossart, 2015] shows a series of occurrences of rebound effect in the ICT. [Bol et al., 2021] shows the importance of the rebound effect in ICT and the global trend of growth of the impacts of ICT despite exponential efficiency gains. Rebound effects can result in a higher environmental impact, or more or less steady impacts, such as for Datacenters. [Patterson et al., 2022] also observes this slow growth of

[1] accessible at `https://github.com/Boavizta/boaviztapi`

[2] Rebound effect also known as Jevons' paradox is the fact that a gain in efficiency in a process can result in a smaller than expected or even increase in the total consumption because of a wider adoption of the new solution.

impacts despite exponential growth in demand and concludes that this trend will continue to hold, leading to a future decrease in impacts due to computation run on datacenters. However, The stabilisation of emissions is in itself not compatible with the global need of reduction of the emissions in all sectors. Furthermore, the proposed solution to maintain this stability of impacts due to the energy consumption is the frequent evolution of hardware to increase the energy efficiency of the datacenters. This frequent need for new hardware would also lead to an increase in the manufacturing impacts of datacenters. [Patterson et al., 2022] also states that the impacts of datacenters are already negligible since big companies buy and produce 'green' energy to power their datacenters and also offset their carbon emissions. Yet, this point of view only accounts for greenhouse gas emissions and therefore does not consider the pollution from e-waste production or the land use for instance. Also, one could discuss the preemption of renewable energy by digital companies in detriment to other activities. Finally, one could also discuss the true potential of carbon offsetting[3] and the pertinence of removing carbon offset from carbon emissions accountability.

In addition to the limits that we just mentioned, LCA only takes into account direct or first order impacts, i.e. those stemming from the life cycle of digital equipment. [Hilty and Aebischer, 2015] draws a framework for understanding the impacts of an ICT solution distinguishing between three levels of impacts, the first one being the "Life-cycle impacts", the second one enabling impacts, referring to actions that are enabled by the application of ICT, and the third one structural impacts, i.e. impacts leading to persistent changes observable at the macro level. [Kaack et al., 2021] insists on the need to evaluate second and third order effects to prevent rebound effects and understand them.

[Rasoldier et al., 2022, Hauschild, 2015] highlight the importance of putting environmental impact measurements in perspective with global sustainability objectives. Such global sustainability objectives can come in the form of the planetary boundaries framework [Sala et al., 2020] of international targets such as the Paris Agreements. Putting impacts in perspective in such a way permits to better judge the costs benefits balance of a solution. It also allows to judge if some optimisation effort is sufficient.

2.4 Information and Communication Technologies and Artificial Intelligence in Health

We are currently in a period of transition towards more and more ICT use in the French Healthcare system. With the development of the EDS projects in all big French Hospitals [Jannot et al., 2017, Haute Autorité de Santé,H.A.S., 2022], lots of NLP tools are developed and being used to allow the reuse of data produced by the hospitals. [Yu et al., 2022] and [Selvan et al., 2022] present some results about the carbon footprint of machine learning for health applications.

This development of AI use in healthcare comes with a lot of ethical challenges, in particular when AI is used in clinical aid / decision tools [CCNE and CNPEN, 2022]. The impacts of a massive deployment of ICT and in particular AI tools poses a number of questions. [Beede et al., 2020] shows the importance of considering the social context in which the solutions are deployed to design tools that can really help medical activities.

[3] https://www.theguardian.com/environment/2023/jan/18/revealed-forest-carbon-offsets-biggest-provider-worthless-verra-aoe

Also, as any other sector, healthcare will need to reduce its emissions, maybe to a lesser extend than other sectors as we may want to prioritise healthcare over other activities. When creating a new AI solution, researchers often point to the potential benefits their solution could have in healthcare. However, as pointed out by [COMETS, Comité d'éthique du CNRS, 2022], potential harm that could be caused by the use of a new solution should also be taken into consideration. Furthermore, it is often admitted that, if it can help save one life, the use of a massive quantity of resources is justified. This principle holds only if getting those resources does not cause more harm than it can mitigate. For instance, ICT requires rare metals to be manufactured. The extraction of those metals can often imply child labour and appropriation of common resources by private companies. There is a need to take into consideration the impacts not only where a solution is deployed but also the impacts induced by this solution in other countries when making the impacts/benefits balance of a healthcare solution.

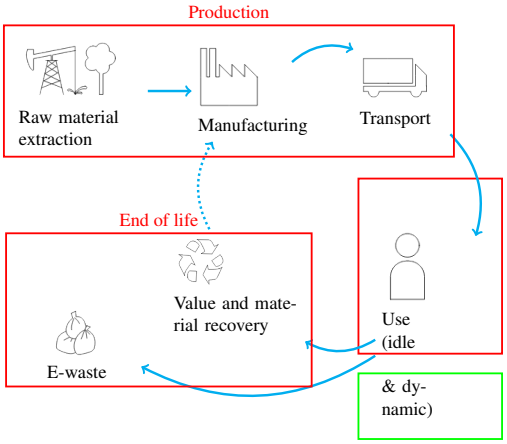
— 3 —

Defining and implementing an estimation tool for the environmental impacts of computation

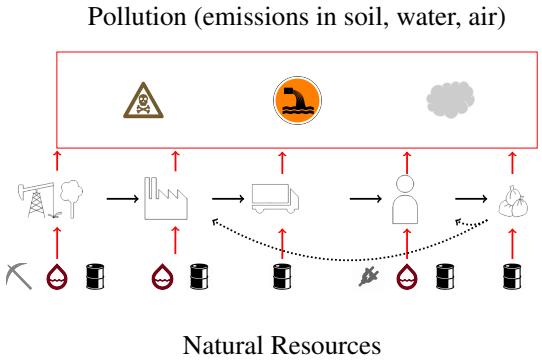
In this chapter, we describe the methodology and implementation used to create *MLCA*, a tool aimed at providing researchers with attributional LCA estimates for numerical computation. The objective is to allow researchers to estimate the expected environmental impacts/ benefits balance before running their experiments and decide if it is or not worth pursuing them.

The question MLCA tries to answer, (e.g. the functional unit) is as follows: "What are the impacts of running program X on the hardware Y during Z hours", where X,Y and Z are parameters provided by the user. Program X can for instance be training a NLP model.

The environmental impacts of running program X are those due to the hardware that it is run on. To estimate them, we thus need to analyse analyse different phases of the life cycle of hardware Y and evaluate the different impacts each considered part of the life cycle has. Figure 3.1 presents different phases in the life cycle of hardware equipment and different types of impacts each phase can have.



(a) different phases of the life cycle of hardware equipment. Use Dynamic is highlighted because most existing tools are focused on the dynamic use of the hardware induced by running a program.



(b) different types of impacts each phase of the life cycle can have. The barrel represents fossil fuels, the droplet represents water, the pickaxe represents metals and the plug represents electricity. The top row pictograms represent, from left to right emissions in soil, emissions in water and emissions in the air.

Figure 3.1: Presentation of the possible scope for a LCA

A summary of the desired features for MLCA (Life cycle phases considered (manufactur-

ing, distribution...) ; diversity of impacts (not focused only on carbon footprint) ; GPU support (since computing intensive programs such as training an AI uses servers with specialised hardware such as GPU or TPU)) and existing tools features can be found in Table 3.1. The column 'estimates consumption' is important because we want to have a tool that can be used before the code runs.

The choice for creating an estimation tool is driven by the fact that we want a tool that can be used before running an experiment but also by the fact that estimates of the consumption based on the TDP of the processing units used (such as in Green Algorithms) seems to provide better quality estimates of the real consumption than software measures (such as in CarbonTracker) [Jay et al., 2023].

As we can see, combining Boavizta's tool [Boavizta, 2021] with Green Algorithm's methodology [Lannelongue et al., 2020] is the best match since it allows to get both usage (Dynamic and Idle) and manufacturing impacts, multiple impact indicators, and GPU support. Since we wanted an estimation tool, We did not turn ourselves to measuring tools such as CarbonTracker [Anthony et al., 2020] or CodeCarbon [Schmidt et al., 2022]. We also did not turn ourselves to the ML CO_2 Impacts [Lacoste et al., 2019] tool because it only accounts for the GPUs and no the rest of the hardware.

Tool	Life cycle phase considered					multiple impacts evaluated	estimates consumption	GPU support
	Man.	Dis.	Use. Idle	Use. Dyn.	EoL.			
Green Algorithms	✗	✗	✓	✓	✗	✗	✓	✓
ML CO_2 Impact	✗	✗	✗	✓	✗	✗	✓	✓
CarbonTracker	✗	✗	✓	✓	✗	✗	✗	✓
CodeCarbon	✗	✗	✓	✓	✗	✗	✗	✓
Boavizta	✓	✗	✗	✗	✗	✓	-	✗

Table 3.1: Feature comparison of different existing tools to study environmental impacts of running computations

We will therefore base our work on Boavizta's code[1] and methodology for evaluating hardware manufacturing impacts of a server [Boavizta, 2021] and model dynamic consumption in a similar way to [Lannelongue et al., 2020].

Working principle of the Manufacturing phase impacts estimation [Boavizta, 2021] proposes a tool for simplified LCA of servers based on [Gröger et al., 2021], the follow up study to [Schödwell et al., 2018]. These studies aim at creating a methodology for evaluating the environmental performance of Cloud services based on LCA methodology. It is a bottom-up approach, estimating impacts for each component and aggregating them to obtain the total impacts for the server.

The impacts are computed with three different metrics. First, Global Warming Potential (GWP), measured in gCO2 eq for the emissions of greenhouse gas such as CO_2. Second, Cumulative Energy demand or Primary Energy (PE), measured in MJ [Frischknecht et al., 2015], for the total energy consumption. PE can be interesting to show that some tasks, even

[1] accessible at `https://github.com/Boavizta/boaviztapi`

though they have a low carbon footprint, can necessitate an important quantity of energy to be executed. Third, Abiotic ressources Depletion Potential (ADP) measured in kgSbeq [van Oers et al., 2020, Bruijn et al., 2002], represents the use of ressources. This third category of impacts is especially pertinent when considering ICT equipment since they use an important quantity of different rare metals to be manufactured.

Working principle of the Use phase impacts estimation In Green Algorithms, and in similar tools, the energy consumption induced by running a computer program is estimated as follows. The TDP of the GPU and CPU are used to model the consumption of the processing units. Also the consumption of the memory allocated to running the computer program is estimated by multiplying the quantity of memory by a consumption/GB factor [Lannelongue et al., 2020].

Once the consumption is estimated, it is generally multiplied by a factor such as the Power Usage Efficiency (PUE) to also account for the power consumption of the infrastructure. [Avelar et al., 2012] presents a complete overview of this metric and [Brady et al., 2013] shows how to compute this metric and the situations where it is adapted to use it and the ones (most of the time) where it is not. The total energy consumption obtained is then multiplied by a Carbon Intensity (CI), dependant on the country in which the computation is run, accounting for the Carbon footprint of producing one kWh. The CI of a country is influenced by the percentage of renewable energy sources and countries such as Iceland have a close to 0 g CO_2 e/kWh while the USA has a CI of $\sim$ 400 g CO_2 e/kWh and South Africa of $\sim$ 800 gCO_2 e/kWh [Lannelongue et al., 2020].

$$gwp = CI * PUE * \frac{(p_c + p_g + p_m) * t}{1000}$$

where p_c refers to the power of CPUs, p_g the power of GPUs and p_m the power of memory. This shift from dynamic energy consumption to carbon footprint of the 'total' energy consumption follows more of a top-down approach

In the end, the scope of this analysis is the manufacturing and usage of the hardware used during the execution of program X, not the distribution and end of life. If hardware Y is a cloud-based server, Network usage is also out of the scope of this analysis. Out of the scope is also the storage of the potential outputs of running program X. We also want to evaluate multiple categories of impact. Table 3.2 shows the different phases of the life cycle and the different impact categories considered in MLCA.

	ADP	GWP	PE	Human toxicity	Water Consumption	...
Manufacture	✓	✓	✓	✗	✗	✗
Distribution	✗	✗	✗	✗	✗	✗
Usage	✓	✓	✓	✗	✗	✗
End of Life	✗	✗	✗	✗	✗	✗

Table 3.2: Summary of the scope of MLCA

Adapting Boavizta's method for estimating the manufacturing impacts and Green Algorithms' method for estimating the energy consumption caused by running a computer program requires some work. First of all, Training a model is done using GPU computing. However,

Boavizta's methodology does not account for GPUs. We therefore need to model and incorporate GPUs in Boavizta's methodology for estimating manufacturing impacts (section 3.1). Then, we slightly modify the way CPU manufacturing impacts are estimated (section 3.2). These first two tasks allow us to estimate the manufacturing impacts of the hardware used. We then needed to attribute part of these impacts to a specific task (for instance, training an NLP model during X hours) (section 3.3). Once we know how to compute manufacturing impacts, we need to estimate the impacts induced by the use of the hardware to perform the task (section 3.4). To provide good quality estimates, and ease of use, we need to collect data about different models of GPU and CPU (section 3.5). Once we are able to estimate the total impacts, we need to put this estimation in perspective (section 3.6). Section 3.7 discusses and concludes on the design of MLCA.

3.1 Manufacture impacts of GPUs

To provide estimations of the manufacturing impacts of different GPUs, we decided to adapt the estimate for CPUs in [Gröger et al., 2021]. As for CPUs, a GPU's impact is modeled as the impact of the die (modeled in function of its surface, the processing unit) to which is added a constant impact that refers to components present on all GPUs such as gold for the connections or the number of Inductors, Resistors, Capacitors present on a board. Since GPUs tend to come with a relatively large amount of memory, we also need to take it into account. For this, we use the same procedure as for the rest of the memory of the server. In the end, GPU impacts are estimated as follows :

$$GPU_{impact} = die_{size} * die_{impact_{per-cm^2}} + base_{impact} + memory_{size} * memory_{impact_{perGB}}$$

where impact $\in$ {ADP, PE, GWP}.

To our knowledge, the only LCA that comprises GPUs is [Loubet et al., 2023]. Therefore, we will base our work on their results to obtain base impacts for GPUs. This would need to be completed with the analysis of some more powerful GPUs to get results that are more certain and that are coherent with the reality of GPUs used for training deep learning models. In [Loubet et al., 2023] Results for scenario 2 are given for 6 servers. Each server contains 2 GPUs, each with a die of .81cm². If we divide the total results for GPUs in scenario 2 by 12 (6 servers * 2 GPu), we obtain impacts per GPU. Then, by removing the estimated impacts for .81cm² of die, we get some base results. We also need to convert from Copper equivalent to Antimony equivalent. to do so, we base ourselves on [van Oers et al., 2020] where it is shown that one kg of Copper is approximately equivalent in terms of ADP to 0.02 kg of Antimony equivalent.

Since results presented in [Loubet et al., 2023] do not present Primary energy, we are not able to provide any figure and will therefore use the base value for CPU.

For the die impacts, we use the same factors as for CPUs. it is fine to use the same value since we wo not be able to get more precise values and we can hypothesise that the process for manufacturing GPU dies is mostly the same as for CPU dies.

3.2 Modification of the way CPU impacts are estimated

We also changed a little bit the way CPU impacts were computed. It follows the same modeling as presented for GPUs but in [Gröger et al., 2021], they proposed a modeling for estimating the die-size when the only known information is the number of cores (In the form $die_{size} = A * number_{cores} + B$). in Boavista's code, they reused this model even when the complete die-size is known (you can modify the value of A (die_size_per_core) but you always add B). We decided to use the known value of the die size when available and not a modeling based on the number of cores.

3.3 Allocation of manufacturing impacts

Once we are able to estimate the manufacturing impacts of the hardware used, we need to attribute part of these impacts to the task we are considering (i.e. the training of the model). The results of this attribution of the total manufacturing impacts of the hardware used for a specific task are called the *embodied impacts*. We attribute uniformly (meaning that each hour of use accounts for the same part of the impacts). For this, we need the total number of hours we can use the hardware before it needs replacement. We use as base value the data from [Luccioni et al., 2022] on the Jean Zay cluster of a replacement rate of 6 years and 85% average usage ($\texttt{total available hours} = 365 * 24 * \texttt{replacement rate} * \texttt{average usage}$) embodied impacts are thus obtained with

$$\texttt{embodied}_{impact} = \texttt{manufacturing}_{impact} \frac{\texttt{hours usage}}{\texttt{total available hours}}$$

3.4 Estimating energy usage impacts

As previously mentioned, to estimate the direct impacts, we use a modeling based on the TDP of the processing units. the dynamic energy consumption is therefore

$$E_{dynamic} = \texttt{hours usage} * \sum_{p \in \{\text{CPU,GPU}\}} (n_p * u_p * \texttt{TDP}_p) + \texttt{memory}_{size} * P_{perGB}$$

where P is the Power, n_p is the number of processing units and u_p is the average usage of the processing unit during the whole period of use. (One could argue that we need to take into account the memory from the GPU as well but it is not implemented yet)
In the same spirit as using a PUE to account for the server's energy efficiency (accounting for the infrastructure that allows the server to run our specific task) we multiply the dynamic energy by a *dynamic ratio* to obtain the total energy consumption.

$$E = E_{dynamic} * \texttt{dynamic ratio} * 1E-3$$

We multiply by $1E-3$ to convert from Wh to kWh. this dynamic ratio is set to a default value calculated from the data gathered on the Jean Zay supercomputer in [Luccioni et al., 2022]. When running a series of experiments, they observed that the energy consumption was distributed as follows : 27kWh in "Infrastructure" mode (computing node off but the rest of

the infrastructure running), 64 kWh in "Idle" mode (computing nodes and the rest on but no jobs running) and 109 kWh in "Production" mode (jobs running) for a total consumption of 200kWh. The dynamic ratio corresponds to the total consumption divided by the consumption in Production mode.

$$\texttt{dynamic ratio} = \frac{\text{TOTAL}}{Production} \simeq \frac{\text{TOTAL}}{\sum_{j \in \texttt{Jobs}} (E_{dynamic})_j} \simeq 1.834$$

This dynamic ratio corresponds to the average energy overheard for running the computing node. this is almost the same definition as the PUE but it does not describe exactly the same thing.

In the end, Energy related impacts are computed as follows :

$$\texttt{Energy}_{impact} = E * impact_{perkWh}$$

where, for instance, the $impact_{perkWh}$ corresponds to the CI if impact corresponds to GWP.

The case for not using the Power Usage Efficiency As explained in [Avelar et al., 2012, Brady et al., 2013], PUE should be used to monitor the evolution of a single datacenter over the years and is not intended for comparison across datacenters. The main issue with PUE is that it is defined as

$$PUE = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}}$$

and IT Equipment Energy does not obviously correspond to the dynamic consumption so one cannot simply multiply the energy by the PUE and assume that dynamic energy and IT Equipment Energy will cancel, therefore leading to converting from dynamic energy to total energy.

3.5 Database

Since our estimates are based on the harware configuration inputed by the user. In order to provide sensible estimates, one needs to be able to input the exact hardware configuration used by their experiments. To this end, we assembled a database of over 150 CPUs and of around 30 GPUs. We wanted to bring together the databases from Green Algorithms and from Boaviztapi. We found out that Green Algorithms' database was mainly based on TechPowerUP databases[2] and presented CPUs (and) by their names and TDP and that Boaviztapi's database was mainly based on the Wikichips website[3] and presented CPUs by their die size and architecture. We decided to use these two sources of data to create a new and unified database for CPUs. for GPUs we base ourselves solely on the TechPowerUP database for GPUs.

3.6 Putting impacts in perspective

In order to prevent the temptation of optimising without considering global objectives [Hauschild, 2015, Rasoldier et al., 2022] we have decided to put the estimated impacts in perspective with two scenarios for sustainability. The first one is the French *Stratégie Nationale Bas Carbone*

[2] `https://www.techpowerup.com/`

[3] `https://en.wikichip.org/wiki/WikiChip`

(SNBC) with an objective of reducing the average anual carbon footprint per capita from 10T CO2eq to 2T[4] and the second one is the framework of the Planetary Boundaries [Sala et al., 2020].

Results are presented in annual person consumption in these two scenarios.

3.7 Conclusion about the design of MLCA

In this chapter, we have presented MLCA, a tool aimed at researchers to provide them LCA estimates for the experiments they are considering to run. This should help researchers make the impacts/benefits balance for their experiments and decide if they are worth pursuing or not.

This tool takes into account the manufacturing and the use phase of the hardware used to run a program but not the Distribution nor the End of Life. Distribution is often not considered because it has only minor impacts relative to the other phases. For the End of Life, this phase generally does not emit a lot of greenhouse gases and is often discarded in consequence. However, the end of life can also have major impacts in terms of toxicity for instance. The major problem to considering the end of life is the lack of available information.

The impacts of an experiment are evaluated with three impact metrics that are GWP, PE and ADP for the greenhouse gas emissions, the primary energy demand and the resource depletion. We would have liked to compute other impact metrics such as the water consumption but were not able to do so because of the lack of data availability.

In the next chapter, we will evaluate the quality of the results produced by MLCA and its usability in different contexts.

[4] https://www.ecologie.gouv.fr/strategie-nationale-bas-carbone-snbc

— 4 —

Case studies on impacts measurement

In order to validate our tool MLCA, we first need to conduct some experiments to ensure that it produces results consistent with the state of the art. These experiments will also ensure that MLCA can be used in a variety of situations. MLCA estimates manufacturing impacts of the hardware used and energy consumption over the usage duration (typically during the training phase of a model). We therefore want to test those two parts.

Firstly, we will present experiments aimed at testing the estimates for the dynamic energy consumption (and the results we will present will therefore focus only on the energy consumption estimated and on the Carbon footprint induced by this energy consumption). These experiments will start in section 4.1 by reproducing the same exact results as the Green Algorithms tool, by first choosing a scenario were we know that MLCA and Green Algorithms use the same data. Then, in sections 4.2 and 4.3 we will focus on reproducing results presented in the two surveys of existing tools ([Bannour et al., 2021] and [Jay et al., 2023]). Finally, in sections 4.4, 4.5 and 4.6 we will try to reproduce results obtained with a different method than the one used by MLCA. This will be done by trying to reproduce results from [Dinarelli et al., 2022], [Cattan et al., 2022] and [Strubell et al., 2019]. We chose those articles for several reasons: they all use a different tool (even if measures presented in these three articles are based on the RAPL and NVML tools); we were able to contact the authors of [Dinarelli et al., 2022, Cattan et al., 2022] to obtain further details about the hardware configuration they use; [Cattan et al., 2022] presents results about the inference phase of models which is a phase that is rarely studied; and [Strubell et al., 2019] was the paper that encouraged NLP researchers to consider the impacts of the models they produced.

Secondly, in section 4.7 we will compare the results that MLCA produces with LCA results produced by Dell about the impacts of the servers they sell. This will allow us to validate the estimations of embodied impacts MLCA generates.

Finally, in section 4.8 we will try to reproduce the results from [Luccioni et al., 2022]. This last step is important because this paper conducts an analysis of the global warming potential induced by the Bloom model. This analysis takes into account embodied emissions and we use figures they present to define the default dynamic ratio MLCA uses.

4.1 Consistency with Green Algorithms

To do a first sanity check, we verify that we are able to reproduce the same results as Green Algorithms on the dynamic consumption part:

We choose a configuration that we know is available in both databases (GA version 2.2 at the time of this experiment):

- 1 CPU A8-7680 (4 cores)
- 1 GPU NVIDIA GTX 1080 Ti
- 64 GB Memory
- Use time of 12h 0min
- no PUE / dynamic ratio
- carbon intensity of France is used (51.28 g CO_2 e/kWh)

We are using Green Algorithms v2.2 for an expected result of 196.32g of CO_2 e and 3.83 kWh of dynamic consumption (this link[1] should in theory get you to the page with this exact setup and results).

If we now run the experiment with MLCA, we see that we indeed obtain the same results of 196gCO_2 e and 3.83 kWh of dynamic energy consumption.

4.2 Impacts of Named Entity recognition: Replicating results from [Bannour et al., 2021]

In order to replicate results, we first need to gather some information about the hardware configuration used to run the experiments. Then, we will face the challenge of inconsistencies in the data presented in the paper. Finally, we will be able to run experiments that give the same energy consumption estimates as those presented in the paper.

4.2.1 Detailing the Hardware configurations

The authors provided us with information about the hardware configurations used to run the experiments.

The facility setup is the Lab-IA [2] cluster. We can see that the only nodes using a 20 core CPU are: n[101-102]:

- 2 x Intel Xeon Gold 6148 20 cores / 40 threads @ 2.4 GHz (Skylake)
- 384 GiB of RAM
- 4 x NVIDIA Tesla V100 with 32 GiB of RAM (NVLink)

[1] `http://calculator.green-algorithms.org//?runTime_hour=12&runTime_min=0&appVersion=v2.2&locationContinent=Europe&locationCountry=France&locationRegion=FR&PUEradio=Yes&PUE=1&coreType=Both&numberCPUs=4&CPUmodel=A8-7680&numberGPUs=1&GPUmodel=NVIDIA%20GTX%201080%20Ti&memory=64&platformType=localServer`

[2] `https://doc.lab-ia.fr/`

using 32 GB of RAM and not the full 384 GB available.

The lab server on the other hand is the Segur machine, using one GTX 1080 Ti with 11GB of memory. It is a Dell PowerEdge R730 with 2 GTX 1080 Ti, 2 Intel Xeon E5-2620 v4 CPU and 125 GB memory (only 11 of which are requested).

While we do not have the Intel Xeon Gold 6148 in our CPU database, we can see on Intel's website[3] that it has a TDP of 150W, was released in 2017 with a process of 14nm with the Skylake architecture; this is sufficient information to add one entry to our database, knowing the information about the Skylake architecture from WikiChips [4].

4.2.2 Problems with the provided data

Some of the results presented in the paper do not seem coherent from one table to the other (tables 3 and 4 of the paper). If we try to convert from energy consumption to carbon emissions using the presented carbon intensity of 39 gCO_2 e/kWh we do not at all find the same results as the ones presented. For instance, for the first method (Yu2020) for the French Press benchmark, it is indicated 1.38kWh consumption and 350.15g CO_2 e.

We can see that if we are to use the presented carbon intensity, we get emissions of `53.8` gCO_2 e for a 1.38kWh energy consumption. This is really far from the 350 gCO_2 e presented in the paper.

Trying to understand the problem

Let us check if the factor to convert from table 4 to table 3 is constant. If it is, it would maybe explain the problems. The CI used to produce the results could be the one of another country than France.

We obtain results around 250 gCO_2 e/kWh with some non negligible variations (the smallest conversion factor is of 191.5 gCO_2 e/kWh while the highest is of 283.5 gCO_2 e/kWh)

According to GA's v2.2 database, this carbon intensity of around 250gCO_2 e/kWh would approximately correspond to Lithuania's one. According to the version 1.1 of the data (version seemingly used in the article), the closest one would be Hungary.

Still, we can observe quite important variations in carbon intensity to convert from the presented energy consumption to the presented carbon emissions, which would tend to infirm the hypothesis of just an error of selection in the carbon intensity used.

Even if there are obvious problems with the presented data, we still want to try and replicate the presented results. Indeed, if the data is flawed only in the table presenting the energy consumption or only in the table presented the carbon footprint, we might be able to reproduce the results of one of the tables (i.e. either the consumption or the carbon footprint).

4.2.3 Experiments

It is said that the default PUE used is 1.67. In order to replicate the results, we will use a dynamic ratio of 1.67, even though the dynamic ratio and the PUE do not have the same meaning, since they are both used in the same way here.

[3] `https://www.intel.fr/content/www/fr/fr/products/sku/120489/intel-xeon-gold-6148-processor-27-5m-cache-2-40-ghz/specifications.html`

[4] `https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)`

We can see in the latest version of Green Algorithms' GPU TDP database [5] that they have a TDP value of 300W for a Tesla V100 GPU whereas we have a TDP of 250W for the same card in our database. In order to see if we can replicate the same consumption and see the difference resulting from this data-point inconsistency we will try two versions. One with a V100 and one with a card with a TDP of 300W in our database: the NVIDIA A100 PCIe 80 GB. This will of course also impact the manufacturing impacts but we are here only focusing on reproducing the same direct impacts.

Table 4.1 presents the estimates that MLCA (MLCA) produces in comparison with the Expected values presented in the paper. We can see that we are able to obtain the same exact energy consumption estimates up to rounding (when we do the modifications to the inputed setup for the facility) except for Yu2020, French Press, Server where we have a slightly lower estimation than the one proposed in the paper. We can also see that, as expected, the estimates we do when considering the "real" setup are lower than the ones presented in the paper and this can be entirely explained by the difference in TDP in the database. We can also conclude that the problem in the presented data lies in the estimates of the carbon footprint and not in the estimates of energy consumption.

4.3 Replicating results from [Jay et al., 2023]

In order to replicate the results from this paper, we first need to gather some information from the paper and its supplementary material which is designed to allow for reproducible experiments.

- The hardware used is a Nvidia DGX-1 with two Intel Xeon E5-2698 v4, 512 GB of memory and 8 NVIDIA Tesla V100-SXM2-32GB.
- The Carbon Intensity for France used in Green Algorithms V2.2 is 51.28gCO_2 e/kWh (latest version of Green Algorithms' Carbon Intensity Database).
- To convert from kWh to kJ, one must multiply the result by 3.6E+3.

We can see in the latest version of Green Algorithms' GPU TDP database [6] that they have a TDP value of 300W for an NVIDIA V100 GPU whereas we have a TDP of 250W for the same card in our database. As a first version, just to see if we are able to obtain the same exact results as those presented in the paper, we will use as GPUs a card with a TDP of 300W in our database: the NVIDIA A100 PCIe 80 GB.

We can also see that the CPU model used is the Xeon E5-2698 v4 with a tdp 135. However, it is not available in Green Algorithm, the model used is the Xeon E5-2697 v4 with a TDP of 145W and 18 cores. In order to reproduce the results presented in the paper, we will use in our setup one CPU with 40 cores, a TDP of 324W (145/18*40) and a die size of 9.12cm² (2*the die size of a Xeon E5-2698 v4, not relevant for the computation of energy)

In the notebook accompanying the paper, we can see that the link explaining the configuration used for the CPU benchmarks are exact copies of the ones for GPU benchmarks. Since

[5] `https://github.com/GreenAlgorithms/green-algorithms-tool/blob/master/data/latest/TDP_gpu.csv` consulted in May 2023

[6] `https://github.com/GreenAlgorithms/green-algorithms-tool/blob/master/data/latest/TDP_gpu.csv`

Method	Task	Hardware	Expected Energy (kWh)	Estimated Energy (kWh)	Estimation trying to match Facility only (kWh)	Expected Carbon (gCO2e)	Estimated Carbon (gCO2e)	Estimation trying to match Facility only (gCO2e)
(Yu2020)	French Press	Server	1.38	1.16		350.15	45.1	
		Facility	1.03	0.861	1.03	260.26	33.6	40
	EMEA	Server	0.07	0.0673		16.67	2.62	
		Facility	0.06	0.0499	0.0595	14.31	1.95	2.32
	MEDLINE	Server	0.08	0.0843		20.68	3.29	
		Facility	0.08	0.0669	0.0797	20.03	2.61	3.11
(Ma2016)	French Press	Server	0.41	0.414		104.4	16.1	
		Facility	0.4	0.341	0.406	102.08	13.3	15.8
	EMEA	Server	0.02	0.0158		3.8	0.616	
		Facility	0.02	0.0179	0.0213	4.99	0.697	0.83
	MEDLINE	Server	0.02	0.0225		5.57	0.878	
		Facility	0.02	0.0216	0.0258	5.67	0.843	1

Table 4.1: Comparison of estimations realised with MLCA (Estimated) with the results computed by [Bannour et al., 2021] (Expected) for the different NER experiments, the columns 'trying to match' present a scenario where we voluntarily choose an NVIDIA A100 PCIe 80 GB card to match the TDP used in Green Algorithms

Benchmark	Value (kJ)	Difference (kJ)
EP	176.04	-0.134
LU	381.6	-0.043
MG	134.28	-0.049

(a) GPU Benchmark

Benchmark	Value (kJ)	Difference (kJ)
EP	25.74	0
LU	15.444	0
MG	64.44	0.09

(b) CPU Benchmark

Table 4.2: Comparison of estimations produced by MLCA (Value) with the results computed by [Jay et al., 2023] (expected results). Estimates are produced trying to match the results presented in the manuscript by inputting hardware that is not the real hardware but that presents the same TDP as the one used to produce the expected results. Difference presents our estimate (Value) minus the expected results.

Benchmark	Value (kJ)	Difference (kJ)
EP	149.04	-27.134
LU	324.36	-57.283
MG	117	-17.329

(a) GPU Benchmark

Benchmark	Value (kJ)	Difference (kJ)
EP	23.04	-2.7
LU	13.824	-1.62
MG	57.6	-6.75

(b) CPU Benchmark

Table 4.3: Comparison of estimations produced by MLCA (Value) with the results computed by [Jay et al., 2023] (expected results). Estimates were produced by inputting the exact hardware used in the experiments. Difference presents our estimate (Value) minus the expected results.

the results for the CPU benchmarks are different from the ones in the GPU benchmark, there must have been a mistake when copying the links for the CPU Benchmarks. We will therefore not be able to ensure we use the same settings as the ones used for producing the results in the paper. We will assume, as this is a CPU Benchmark, that the inputed CPU usage was 1 and inputed GPU usage was 0. This configuration leads to an energy consumption of 8.58Wh for one minute. Since this value is strangely similar to the value of 7.58Wh/min used in the paper, we will also assume that there was a mistake when copying results from the Green Algorithm website and therefore use the value of 8.58Wh/min instead of the value of 7.58Wh/min to compute the expected results.

Table 4.2 present the results obtained when trying to match the expected results (same hardware setup as used for obtaining values with Green Algorithms) and Table 4.3 present the results obtained when using the hardware setup really used.

We can see that we are able to obtain results that are exactly the same as the expected ones up to rounding errors (difference 3 orders of magnitude lesser than the value). We can also see that even though the input value to Green Algorithms does not exactly correspond to the hardware setup used, the difference to the expected results is not too high. The difference between our estimate using 'correct' data and the expected values is around 10% of the estimated value. These results demonstrate the importance of inputting the right hardware if one wants precise results.

4.4 Impacts of Spoken Language Understanding: Replicating results from [Dinarelli et al., 2022]

As for other experiments aiming at reproducing results, we first need to gather enough information to run our experiments. We will also check the consistency of the results presented in the paper. This will allow us to run our estimates. We will focus on two results that we will try to reproduce. First the fine tuning of the SSL model which is the most time consuming task presented and then we will focus on the training time for the spectro model, this should allow us to get a good overview of the results.

4.4.1 Trying to find information about the hardware setup

Thanks to the collaboration of the authors, who gave us some insight on the hardware used for running their experiments, we are able to produce some estimates.

Hardware for the fine-tuning

The author said that a node from the Jean Zay supercomputer with 4 GPUs with 32GB memory was used for the fine tuning of the wave2vec model. If we look at the Idris' website [7] we think that the nodes used were from the **v100-32g**, since it is the only node with matching requirements in terms of number of GPUs and memory per GPU.

These nodes have the following hardware configuration :

- 2 Intel Cascade Lake 6248 (20 cores at 2,5 GHz)
- 192 GB memory per node
- 4 GPU Nvidia Tesla V100 SXM2 32 GB

Because we do not have the Intel Cascade Lake 6248 in our database, we need to find some information about it. We can see on Intel's webpage [8] that it is a processor of the Cascade Lake architecture. On Wikichip [9], we can see that Cascade Lake Processors use dies largely similar to those of the Skylake cores [10]. Combining all of these pieces of information, we can get an estimation of the details of an Intel Cascade Lake 6248 :

- model: "Xeon Gold 6248"
- manufacturing date: "2019"
- process: 14nm
- number of cores: 20
- die size: 694 mm² (XCC configuration)

[7] http://www.idris.fr/jean-zay/cpu/jean-zay-cpu-hw.html#gpu_p13

[8] https://www.intel.fr/content/www/fr/fr/products/sku/192446/intel-xeon-gold-6248-processor-27-5m-cache-2-50-ghz/specifications.html

[9] https://en.wikichip.org/wiki/intel/microarchitectures/cascade_lake#LCC_SoC

[10] https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)#Core

Hardware for training the models

We are told that training uses only one GPU at a time and that it uses roughly half of the time a RTX 2080 Ti and the other half a GTX 1080 Ti; to represent this, we will put the two different models in the list of GPUs and use a 'gpu usage' of 0.5. We are also told that the training uses 80 GB memory with no additional information on the hardware used. Since we do not know any more precise information, we will use the default values of MLCA to complete the missing pieces of information.

4.4.2 Coherency of the results

One first good news is that information are coherent with themselves. Using the indicated (in the paper) carbon intensity of 51gCO_2 e/kWh used and indicated energy consumption, we are able to find back the carbon emissions indicated in the table. The only problem is that for table 1 of the paper, it seems that there was a translation error when filling the table. The figures are written in the French notation with "," separating units from decimals and not the usual ".". For instance, if we look at the first line of table 1 of the paper, we can read a consumption of 4,473 kWh, that we can translate to 4.473 kWh. We obtain $4.473 * 51 = 228.123 gCO_2e$, the same value as indicated in the paper.

We then only need to be able to find coherent energy consumption values to obtain comparable results.

4.4.3 Estimating energy consumption

Fine tuning of the SSL model

When running an estimate of the impacts of the fine tuning, we can see that we obtain an estimate of 5.46kg CO_2 e for the direct impacts and a dynamic consumption of 107 kWh, which is close to the 4.729kg CO_2 e and 97.720 kWh presented in the paper. The fact that results are not a perfect match and slightly higher than presented can be explained by the fact that the measures in the paper were carried out based on a measurement tool (CarbonTracker). (The evaluation of the impacts of the fine tuning of the SSL model presented in the manuscript ([Dinarelli et al., 2022]) are borrowed from [Evain et al., 2021]. [Evain et al., 2021] used the methodology from [Parcollet and Ravanelli, 2021])

Replicating results from Table 1 of the paper

We now turn our focus towards replicating the measure of impacts for the spectro experiments presented in the Table 1 of the paper.

Table 4.4 compares our estimates with the presented measures on the spectro experiments. We can see that we obtain carbon emission estimates around 3 times higher than those presented in the paper. This difference is important but results are still in the same order of magnitude. It is expected that we obtain higher estimates than the measurements as, as presented in [Jay et al., 2023] estimation tools tend to produce higher results than measurement tools. However, the difference could potentially be explained by the lack of information about the GPU usage (if GPUs were running at 30% capacity for instance) during training.

model	expected power (kWh)	estimated power (kWh)	expected carbon (gCO2e)	estimated carbon (gCO2e)
spectro 3 steps	4.473	10.1	228	517
spectro 2 steps	2.989	6.78	152	346
spectro 1 step	1.708	4.44	87	226

Table 4.4: Comparison of estimates produced by MLCA (estimated) with the measures computed in table 1 of [Dinarelli et al., 2022] (expected) on the spectro experiments

4.5 Impacts of Transformers: Replicating results from [Cattan et al., 2022]

This paper studies the gains and impacts of choosing to use one type of NLP model in a system. It evaluates the impacts of training the models but also of running inferences.

We try to replicate results that were obtained by scaling up the results obtained in [Cattan et al., 2022] for one inference to account for the weekly number of requests the search engine of Qwant receives. As always, we will need to first find the hardware configuration used, then we will check the coherency of the expected results and run our experiments.

4.5.1 Hardware configuration

We where told that the hardware used was an NVIDIA DGX equipped with 8 NVIDIA Tesla V100 SMX2 16GB. We were not able to find such a configuration on NVIDIA's website but since the Tesla V100 SMX2 32GB GPU present in an NVIDIA DGX-1 server have the same exact TDP, we will suppose that this is the hardware used.

4.5.2 Running experiments

Table 4.5 compares our estimates with the presented measures. We can see that we obtain results as low as 4 orders of magnitude lower than the expected results. This massive difference cannot be easily explained and is a really surprising result.

4.5.3 Explaining the massive differences between our estimates and the expected results

In our estimates, the consumption of one DGX-1 is estimated at 2460W (if we were to suppose that CPUs are running at full capacity) and 2190W if we suppose that CPUs do not run. This is significantly lower than the 3500W provided by NVIDIA and can be due at least in part to the fact that we do not account for storage in our estimation.

Our results are way lower than those presented. However, the presented results seem at least surprising. If we use the consumption value provided by NVIDIA of 3500W for one DGX-1 DGX-1 datasheet [11] and suppose it was used for 8 hours like for ATIS-FR with XLM-Rlarge,

[11] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-1/dgx-1-rhel-centos-datasheet-update-r2_Updates_NV_web_fr_FR.pdf

Task	Model	Time (Hours)	Expected Energy (MWh)	Estimated Energy (MWh)	Expected CO2 (Kg)	Estimated CO2 (Kg)
MEDIA	FlauBERTbase	20.19	204.24	0.0442	147.84	2.27
	CamemBERTlarge CCNet 135 Gb	50.63	512.67	0.111	371.14	5.69
	CamemBERTbase OSCAR 138 Gb	20.23	204.67	0.0443	148.15	2.27
	CamemBERTbase CCNet 135 Gb	15.57	157.39	0.0341	113.96	1.75
	CamemBERTbase OSCAR 4 Gb	15.89	160.7	0.0348	116.35	1.79
	CamemBERTbase CCNet 4 Gb	15.64	158.08	0.0343	114.42	1.76
	CamemBERTbase Wiki 4 Gb	15.38	155.46	0.0337	112.57	1.73
	FrALBERTbase Wiki 4 Gb	9.11	92.02	0.02	66.61	1.02
	XLM-Rbase	17.2	173.94	0.0377	125.9	1.93
	XLM-Rlarge	55.68	563.95	0.122	408.25	6.26
	mBERTbase	17.95	181.41	0.0393	131.36	2.02
	distill-mBERTbase	15.06	152.08	0.033	110.11	1.69
	small-mBERTbase-fr	16.45	166.24	0.036	120.35	1.85
ATIS-FR	FlauBERTbase	3.08	30.88	0.00675	22.33	0.346
	CamemBERTlarge CCNet 135 Gb	7.36	74.23	0.0161	53.75	0.827
	CamemBERTbase OSCAR 138 Gb	3.27	32.57	0.00716	23.56	0.367
	CamemBERTbase CCNet 135 Gb	2.55	24.79	0.00559	17.94	0.286
	CamemBERTbase OSCAR 4 Gb	2.52	25.18	0.00552	18.25	0.283
	CamemBERTbase CCNet 4 Gb	2.59	25.49	0.00567	18.48	0.291
	CamemBERTbase Wiki 4 Gb	2.5	24.95	0.00548	18.1	0.281
	FrALBERTbase Wiki 4 Gb	1.39	13.71	0.00305	9.93	0.156
	XLM-Rbase	2.4	25.72	0.00526	18.63	0.27
	XLM-Rlarge	8.02	76.08	0.0176	58.6	0.901
	mBERTbase	2.48	24.72	0.00543	17.94	0.279
	distill-mBERTbase	2.35	23.25	0.00515	16.79	0.264
	small-mBERTbase-fr	2.46	24.56	0.00539	17.79	0.276

Table 4.5: Comparison of estimates produced by MLCA (Estimated) with the measures realised in [Cattan et al., 2022] (Expected).

we would expect a consumption of 28kWh. This is extremely far from the 76MWh presented. There is therefore a problem in the data presented in the manuscript (like a problem with the training times) or in the hardware configuration used.

Furthermore we can see that conversion from energy consumption to carbon emissions make us remark that the carbon intensity seemingly used is approximately 1.38 gCO_2 e/kWh. This is extremely low as the Carbon Intensity for France is estimated between 50 and 200 gCO_2 e/kWh.

In order to get further insight on what could cause these inconsistencies we will try and reproduce results from [Cattan et al., 2021] which uses the same configuration. If results from this paper are consistent with our estimates, this would tend to confirm that there is a problem in the data presented in [Cattan et al., 2022] and not in our estimates.

4.5.4 Table from [Cattan et al., 2021]

It is said that only one V100 GPU is used for training the different models (we will suppose that it was done on one DGX-1 server).

By dividing the presented Carbon Footprint by the Energy consumption for each model, We can see that the carbon intensity used seems to be of 295 gCO_2 e / kWh. We can also see on Experiment-Impact-Tracker's repository [12] that they by default use a PUE of 1.58, in order to replicate their results. We will choose to use this value of 1.58 as dynamic ratio.

We can suppose that during training only the GPU is used at full capacity. We can also try a scenario where one core of one GPU is used during training. This would lead to including a CPU usage of 1/20 (since the CPU has 20 cores).

Table 4.6 presents the results of these experiments. We can see that for upper and lower estimates, we obtain results slightly higher that those presented in the paper but in the same order of magnitude. This is expected since estimation tools tend to provide higher (and closer to reality) estimates than measurement tools. However, we can also see that the estimation tool ([Jay et al., 2023]) does not capture some subtleties. For instance small-$mBERT_{base}$ training is quicker than $mBERT_{base}$ one. However this does not translate to smaller energy consumption. This can most probably be explained because the GPU usage is in average higher when training small-$mBERT_{base}$ than when training $mBERT_{base}$. Without fine knowledge of the processing units usage, we cannot provide very precise estimations and track small changes such as this one.

All of these results tend to confirm that there are problems with the data available in [Cattan et al., 2022] but that the data from [Cattan et al., 2021] confirms us the hardware configuration used.

4.5.5 New experiment

After pointing out the problems in the data to the authors, they ran a new experiment on the Segur machine. Table 4.7 presents the newly obtained results using Experiment-Impact-Tracker:

From these results, and knowing that the Segur machine is equipped with 2 20 core CPUs with 125 GB RAM and 2 GTX 1080 Ti, we can estimate that approximately 2 cores (1.04/.53) were used at full capacity during training, which equates to 1/20 usage. The two GPU also

[12] `https://github.com/Breakend/experiment-impact-tracker`

model	estimate	time (s)	expected energy (kWh)	estimated energy (kWh)	expected carbon (kgCO2e)	estimated carbon (kgCO2e)
$\text{CamemBERT}_{\text{base}}$	lower	7207	1.08	1.41	0.317	0.415
$\text{CamemBERT}_{\text{base}}$	upper	7207	1.08	1.43	0.317	0.421
$\text{CamemBERT}_{\text{large}}$	lower	19445	3.1	3.77	0.914	1.11
$\text{CamemBERT}_{\text{large}}$	upper	19445	3.1	3.83	0.914	1.13
$\text{FrALBERT}_{\text{base}}$	lower	3816	0.57	0.75	0.167	0.221
$\text{FrALBERT}_{\text{base}}$	upper	3816	0.57	0.761	0.167	0.225
$\text{XLM-R}_{\text{base}}$	lower	7676	1.14	1.5	0.337	0.441
$\text{XLM-R}_{\text{base}}$	upper	7676	1.14	1.52	0.337	0.448
$\text{XLM-R}_{\text{large}}$	lower	21137	3.3	4.1	0.973	1.21
$\text{XLM-R}_{\text{large}}$	upper	21137	3.3	4.16	0.973	1.23
$\text{mBERT}_{\text{base}}$	lower	7333	1.07	1.43	0.317	0.422
$\text{mBERT}_{\text{base}}$	upper	7333	1.07	1.45	0.317	0.428
$\text{samll-mBERT}_{\text{base}}$	lower	7190	1.09	1.4	0.321	0.414
$\text{samll-mBERT}_{\text{base}}$	upper	7190	1.09	1.42	0.321	0.42
$\text{distil-mBERT}_{\text{base}}$	lower	6466	1.06	1.26	0.314	0.372
$\text{distil-mBERT}_{\text{base}}$	upper	6466	1.06	1.28	0.314	0.378

Table 4.6: Comparison of estimates produced by MLCA (estimated) with the measures of impact realised in [Cattan et al., 2021] (expected) for training the different models. The lower estimates correspond to a scenario where we suppose CPUs were not used during training and upper estimates correspond to a scenario where we suppose that one CPU core was used during training

$\text{cpu}_{\text{hours}}$	1.04 h
$\text{gpu}_{\text{hours}}$	0.99 h
$\text{estimated}_{\text{carbon impact kg}}$	0.02 kgCO_2e
$\text{total}_{\text{power}}$	0.43 kWh
kw_{hrgpu}	0.25 kWh
kw_{hrcpu}	0.02 kWh
$\text{exp}_{\text{lenhours}}$	0.54 h

Table 4.7: Presentation of the new measures obtained by the authors when fine-tuning on 10 epochs on the MEDIA task with fr-ALBERT. The measured impacts can be divided by ten to obtain results produced in the same conditions as in [Cattan et al., 2022]. The authors obtained these new measures running this experiment on the Segur machine.

seem to have been used at full capacity. we can deduce the used Carbon Intensity by dividing the estimated carbon by the measured power.

This result of 56 gCO_2 e/kWh leads us to think that the Carbon Intensity of France was used (which would be logical since the experiment was run in France).

We also know that Experiment Impact Tracker uses a PUE of 1.58, in order to try and reproduce these results, we will use a dynamic ratio of 1.58. We will also try with the base dynamic ratio and see the difference

All of this allows us to run the following experiment to try and reproduce these results.

	Expected	Estimated	Match
energy (kWh)	0.43	0.507	0.436
Carbon (kgCO2e)	0.0241	0.0284	0.0244

Table 4.8: Comparison of estimates produced by MLCA (Estimated & Match) with the new measures. The Match scenario uses a dynamic ratio of 1.58 while the Estimated scenario uses the base dynamic ratio of 1.83

Table 4.8 compares the estimates produced in both scenarios with the new measures. We can see that we obtain very close results (a little bit higher just as expected) when trying to get an exact match by using a dynamic ratio of 1.58 and estimates are increased when using the base dynamic ratio which stands around 1.83.

In conclusion, we have encountered abnormally high values in [Cattan et al., 2022]. When trying to understand how these values might have been obtained, we looked at [Cattan et al., 2021] where we were able to reproduce close estimations to the measures of impacts presented for the training of the different models. After pointing out the unexpected values in [Cattan et al., 2022] to the authors, they ran new experiments whose results could be replicated.

4.6 Impacts of NLP: Replicating estimations from [Strubell et al., 2019]

4.6.1 Information about the hardware configuration

It is described in the paper that estimates are conducted by training all models for a maximum of 24h. They use RAPL and NVIDIA System Management Interface to measure the average consumption of the CPUs and GPUs. All models are trained on one NVIDIA TITAN X except for ELMo which is trained on 3 GTX 1080 Ti. They then transcribe these results to estimates by using the training time given in the paper and the description of the hardware given in the paper.

No figures are presented regarding the average consumption of the memory, CPU and GPU (separated). We only know about the model of GPU used for estimating the consumption and the total estimated consumption for training each model. We will therefore not give any value for the CPU and ram and run our estimates as is to see what results we obtain. We expect not to obtain exact results: it will not be possible given the information missing. Since they use measurement tools, we can think that using a modeling using the TDP will give an higher result but since we do not know the quantity of memory used and the CPU used, we are not sure that

the results will be higher (even if we can hypothesize that the CPU average consumption is negligible compared to the GPU consumption.)

One reassuring point is that GTX 1080 Ti, V100, P100 and Titan X GPUs have the same TDP so the consumption estimated should make sense.

They use a PUE of 1.58 and a Carbon Intensity of 0.954 pounds CO_2 e/kWh for American electricity production which is equivalent to 432.72 gCO_2 e/kWh.

4.6.2 Checking the Coherency of the presented results

Since there are no estimates given for models trained on TPUs, we will in the first time at least ignore these models.

Since table 3 of the paper presents the estimated consumption used, we can first check the coherency of the table by seeing if we can reproduce the same energy consumption by multiplying the power by the training time and the PUE

We can see that, up to rounding we obtain the same results. We can also check that we obtain the same carbon emissions and now serenely proceed with running our estimations.

4.6.3 Running our estimations

For a first check, we will compare the estimated power consumption of just the GPUs with the presented hardware consumption. The TDP of a P100 GPU is 250W, also the same as the one of a GTX 1080 ti.

model	estimated (W)	measured (W)
$\text{Transformer}_{\text{base}}$	2000	1415.78
$\text{Transformer}_{\text{big}}$	2000	1515.43
ELMo	750	517.66
$\text{BERT}_{\text{base}}$	16000	12041.51
NAS	2000	1515.43

Table 4.9: Comparison of the power consumption measured in [Strubell et al., 2019] (measured) and estimates of the consumption based on the TDP of the used GPU and training time.

Table 4.9 compares the estimated power consumption when only accounting for the GPUs with the power consumption measures presented in the paper. We can see that, as expected since the provided consumption result from using measurement tools, the estimated consumption is bigger (approximately one third bigger) than the measured consumption. Still, it remains in the same order of magnitude.

Table 4.10 presents the results of our estimates on two different scenarios. The first one (match) uses the same PUE and CI as presented in the paper while the second (base) uses the base values of MLCA for the dynamic ratio and CI of the USA. We can see that we obtain estimates that are, as expected, a little bit higher than those presented. We can explain the higher estimated energy when using the base values for MLCA because of the difference in Dynamic ratio. We use as base value a dynamic ratio of 1.83 when the match scenario uses a dynamic ratio of 1.58. We can also see that the estimated carbon footprint is slightly higher in the match scenario than in the base scenario; this can be explained by the difference in CI

model	expected energy (kWh)	estimated energy match (kWh)	estimated energy base (kWh)	expected CO2e (kg)	estimated CO2e match (kg)	estimated CO2e base (kg)	expected CO2e (lbs)	estimated CO2e match (lbs)	estimated CO2e base (lbs)
$\text{Transformer}_{\text{base}}$	27	38	44	11.79	16	16	26	36	36
$\text{Transformer}_{\text{big}}$	201	267	310	87.09	116	115	192	255	253
$\text{BERT}_{\text{base}}$	1507	2000	2320	652.17	865	859	1438	1907	1893
NAS	656347	871000	1.01e+06	284018	377000	374000	626155	831143	824529
ELMo	275	404	470	118.84	175	174	262	385	383

Table 4.10: Comparison of the measures presented in [Strubell et al., 2019] with estimates produced by MLCA (estimated). The base scenario uses the base values for dynamic ratio and CI for the USA in MLCA while the match scenario uses the PUE and CI presented in [Strubell et al., 2019]

used. Indeed, the CI for the USA in the base values is 370gCO_2 e/kWh instead of the 432gCO_2 e/kWh when trying to match.

4.6.4 Hyper-parameter search

To complement the case study on hyper-parameter search and costs not only on training one model but of the whole process, let us try and reproduce similar results, which we would be able to study also in terms of the other impacts estimated by MLCA.

Number of Models	Hours	Expected energy (kWh)	Estimated energy (kWh)	Expected electricity cost ($)	Estimated electricity cost ($)
1	120	41.7	55	5	7
24	2880	983	1320	118	158
4789	239942	82250	110000	9870	13200

Table 4.11: Comparison of the measured energy consumption and cost presented in [Strubell et al., 2019] (Expected) with estimates produced by MLCA (Estimated)

Table 4.11 compares the estimated energy consumption and electricity costs with the expected ones. We can see that we still obtain higher energy consumption values than the ones presented. This fact can be mostly explained by the difference between using a PUE of 1.58 and a dynamic ratio of 1.83.

4.6.5 Integrating Life cycle to previous analyses

If we now look at the full estimates produced by MLCA for the entire NLP pipeline and not only on the energy related impacts, we can see that the full impacts estimated for performing the whole model search, hyper-parameter tuning and training represents the annual GWP of 22 persons if we place ourselves in a scenario where we would respect the "Stratégie Nationale Bas Carbone" for France by 2050 (a limit of 2 tCO_2 e per year per person). If we place ourselves in the framework of the Planetary boundaries, where if we want to stay sustainable, societies must not overpass the planetary boundaries, The whole process accounts for the maximal annual impacts of 44 persons in terms of Green House Gas emissions (a limit of 985 kgCO_2 e per person per year) and the annual impacts of 24 persons in terms of resource depletion (a limit of 0.0317 kgSbeq per person per year).

Of course, if computations were to be run in a country with a less carbon intensive electricity mix, global warming potential would be lower. Still, the impacts on resources depletion are very important, and, in this estimation, we do not take into account any (1 GB) memory on the server that runs the experiments.

If we were to add memory, for instance 512 GB of memory, we would obtain the following estimation with expected impacts as high as the maximal annual ones of 86 persons in terms of GWP and 33 persons in terms of ADP when not exceeding the planetary boundaries.

As a comparison, if we were to make the same estimates but running in France, we would obtain the following (with a CI of 98gCO_2 e/kWh) : It would still represent the maximal annual emissions of 32 persons in terms of GWP or in terms of ADP.

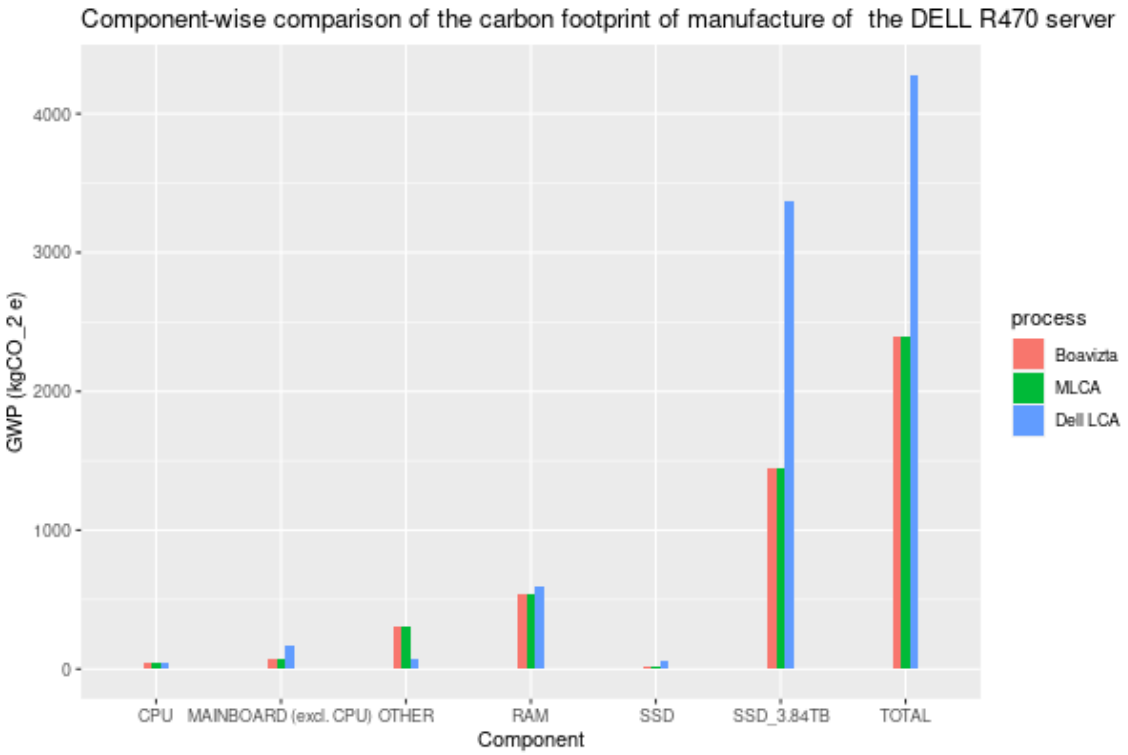


Figure 4.1: Component-wise comparison of the GWP of manufacturing of the Dell R470 server

This small change in ADP impact can be explained by the fact that most of the impacts on resource depletion are due to manufacturing the hardware used (a very small amount of the impacts are due to the energy consumption). This demonstrates the importance of both considering embodied impacts and of considering other impacts than just the carbon footprint.

4.7 Comparing manufacturing impacts to Dell LCAs

In order to validate the embodied impact estimations MLCA produces, we compare the results MLCA produces with the LCA results presented by Thinkstep for DELL on the R740 [Thinkstep, 2019] and By Sphera for Dell on the R6515, R7515, R7525 servers [Sphera, 2021] [13].

4.7.1 Dell R740

This first LCA of a server was already used by Boavizta to validate their tool. Since MLCA does not greatly differ from Boavizta's one on the manufacturing impacts estimate for servers with no GPU, we expect to obtain close results to them.

Figure 4.1 compares the GWP value obtained for the different components by the Boavizta tool, MLCA (MLCA) and the Dell LCA result. As we can see, the Boavizta tool and MLCA obtain very close results. This is expected since MLCA is based on Boavizta with some changes

[13] All Product Carbon Footprint and LCA produced on the different Dell Products can be found at `https://www.dell.com/fr-fr/dt/corporate/social-impact/advancing-sustainability/sustainable-products-and-services/product-carbon-footprints.htm#tab0=3`

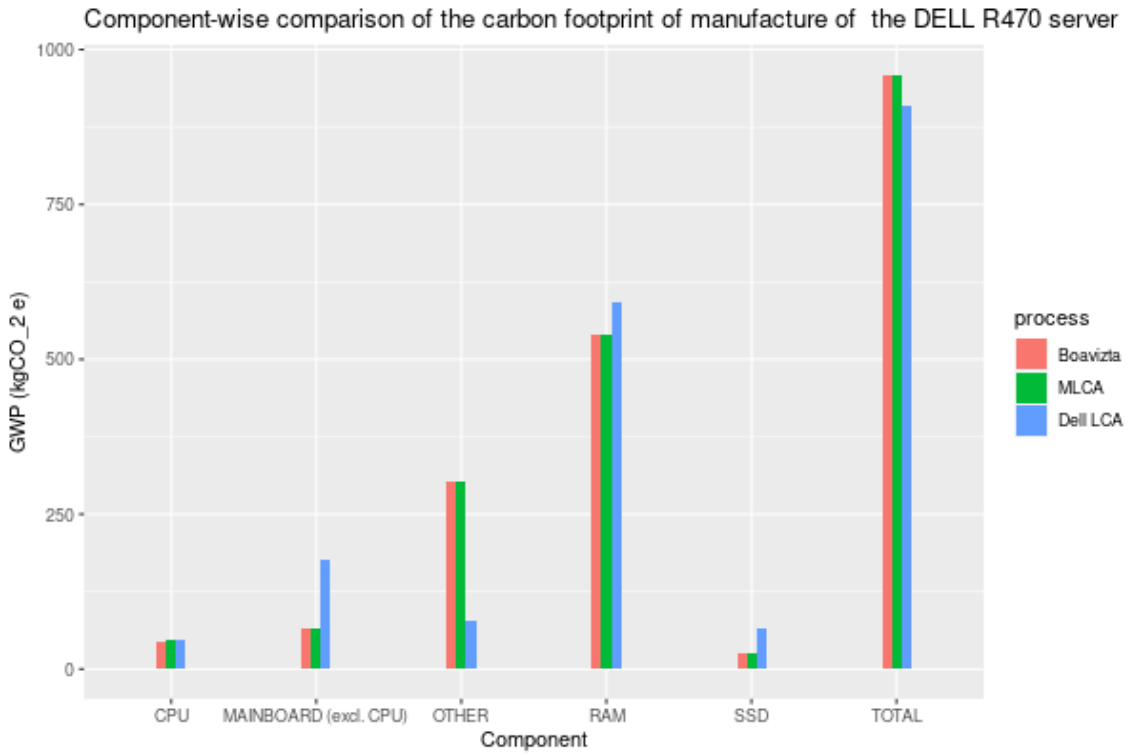


Figure 4.2: Component-wise comparison of the GWP of manufacturing of the Dell R470 server not considering the 3.84TB SSDs

in the way CPU impacts are computed and some bugfixes. We can see however that MLCA obtains a way lower estimate than the expected results (Dell LCA) but it seems to be mainly explained by the difference in the estimate for the 3.84TB SSDs. We can see that MLCA provides significantly lower estimates than the expected result for these components. If we look at the server without these big disks in figure 4.2, we can see that we obtain pretty close estimates for the CPU, RAM and Total impacts with a lower estimate of the impacts of the motherboard and an overestimate for the other components that compensate a little.

4.7.2 Dell R6515, R7515, R6525, R7525

In this section we compare the results MLCA produces with LCA results produced by Sphera for Dell on the R6515, R7515, R7525 servers [Sphera, 2021]. Since there are few configuration differences between the R6515 and R7515 and between the R6525 and R7525, we will only focus on the R6515 and R6525.

First, we compare the total manufacturing impact estimated by MLCA with the expected results?

For the manufacturing of the R6515, we obtain an estimate of 1200 $kgCO_2$ e when the expected results stand at 1343 $kgCO_2$ e. For the R6525, we obtain an estimate of 1600 $kgCO_2$ e when the expected result stands at 1709 $kgCO_2$ e.

We can see that we obtain close results. If we now take a deeper look at the repartition of impacts by components, Figure 4.3 compares MLCA (MLCA) with the expected results from the Dell LCA. As in section 4.7.1, we can see that MLCA underestimates the SSD impacts and produces a close estimate of the total manufacturing GWP impact. We also see that for the

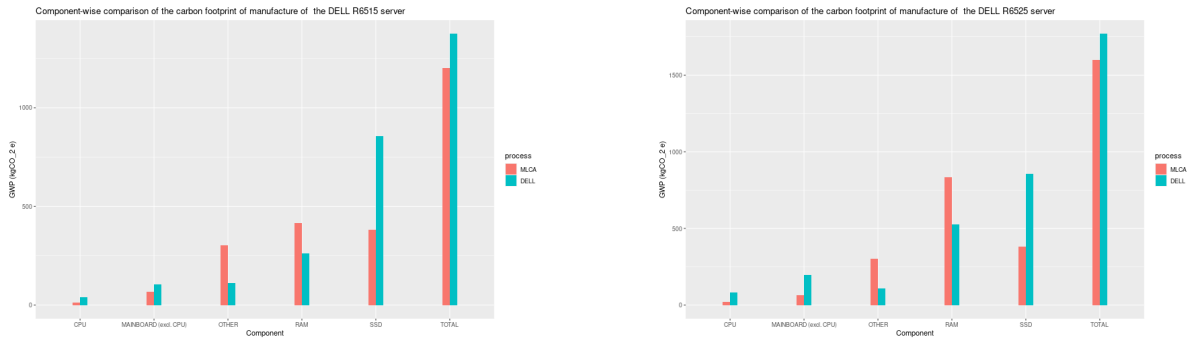


Figure 4.3: Component-wise comparison of the GWP of manufacturing for the Dell R6515 (left) and R6525 servers (right)

mainboard, we get a lower estimate that counterbalances the overestimate for the other components. This time, the estimates for the RAM and CPU impacts are farther from the expected result than in the comparison for the Dell R470. Still, the overall results tend to confirm the adequacy of the results MLCA produces with expected results about the manufacturing impacts of a server.

The comparison of MLCA with the Dell LCA results confirms that overall, MLCA produces adequate results component by component even if it tends to underestimate the impacts of storage.

4.8 Replicating the Bloom estimates from [Luccioni et al., 2022]

Now that we have tested MLCA on the results it produces about the energy consumption then on the manufacturing impacts, let us test our tool on all of this at the same time on the case of the Bloom carbon footprint study.

4.8.1 Gathering information about the setup

To replicate their experiments, we first need to gather some information on the time duration and hardware setup for the training phase.

We can see in the paper that the training phase lasted for 118 days, 5 hours and 41 minutes for a total of 1,082,990 GPU hours (table 1 of the paper).

In section 4.1 of the paper, we can read that training used on average 48 computing nodes with 8 GPUs each.

Combining the real time and these information about the setup, we obtain an estimate of the number of GPU hours of 1,089,670.4 hours this gives us a pretty close figure to the real GPU time.

It is written in the paper that training took place on the Jean Zay supercomputer, using HPE's Apollo 6500 Gen10 Plus [14]. We can read on their website that it uses AMD EPYC 7000 Series CPUs. Combining this information with information about the Jean Zay supercomputer

[14] `https://buy.hpe.com/fr/fr/compute/apollo-systems/apollo-6500-system/apollo-6500-system/hpe-apollo-6500-gen10-plus-system/p/1013092236`

on IDRIS's website [15], we can see that only the **gpu$_{p5}$** partition uses such CPUs. We can conclude that for each of the 48 used nodes, the server configuration is :

- 2 CPUs : AMD Milan EPYC 7543
- 512 Go of Memory
- 8 NVIDIA A100 SXM4 80Go

4.8.2 Comparing the server footprint with the PCF sheet

In section 4.1, it is stated that they use values provided in the HPE ProLiant DL345 Gen10 Plus PCF [16], the closest server with information provided. In this PCF sheet, we can read that servers are of type rack and that the estimated Carbon Footprint is of 2503.2 kg CO_2 e.

```
GWP: {'manufacturing': 2300.0, 'use': 1170.0, 'unit': 'kgCO2eq'}
PE: {'manufacturing': 29000.0, 'use': 39700.0, 'unit': 'MJ'}
ADP: {'manufacturing': 0.17, 'use': 0.000198, 'unit': 'kgSbeq'}
RAM impact GWP: {'value': 1800.0, 'unit': 'kgCO2eq'}
```

If we try MLCA with the server configuration used for training, we can see manufacturing impacts of 2300 kg CO_2 e. This impact is close to the 2500 kgCO_2 e provided on the PCF sheet and is mainly impacted by the quantity of memory used, as it accounts for 1800 kg CO_2 e.

4.8.3 Comparing the GPU footprint with the chosen value

In section 4.1 of the paper, it is stated that a value of 150 kg CO_2 e is chosen. Taking a look at the source, there is no real justification given for that value. Given that in [Loubet et al., 2023] a small GPUs manufacturing is estimated at emitting around 30 kg CO_2 e, we could hypothesize that GPU manufacturing impacts would be in the order of 50 to 150 kg CO_2 e.

For the specific model used, the "NVIDIA A100 SMX4 80GB", MLCA provides an estimate of 330 kgCO_2 e for the manufacturing of one GPU. This impact is mainly influenced by the quantity of memory on the GPU with a carbon footprint of 290 kgCO_2 e, leaving 40 kgCO_2 e for the rest of the GPU. This estimate of 40kgCO_2 e for the GPU without any memory is consistent with the values provided in [Loubet et al., 2023]. The importance of the memory present on the GPU in its manufacturing impacts show the need for an LCA of a modern GPU used for High Performance Computing (HPC) to obtain good quality estimates

4.8.4 Estimating the total impacts

With all of the previous information, we can run the estimation; we find embodied impacts of 7T CO_2 e for the servers and 8.1T for the GPUs to compare with the 7.6T for the servers and 3.6 T for the GPUs in the paper. Most of the difference is due to estimated impacts of 330 kgCO_2 e for one GPU while it was estimated to 150 kgCO_2 e in the paper.

For the dynamic consumption, we obtain an estimate of 23.7T CO_2 e, mainly due to the GPUs (accountable for 22.4T, the only difference with the figure obtained in the paper being

[15] http://www.idris.fr/jean-zay/cpu/jean-zay-cpu-hw.html#gpu_p13

[16] https://www.hpe.com/psnow/doc/a50005151enw

the slightly off conversion from real time to GPU hours) while the memory, not accounted for in the paper brings another 1.35T CO_2 e.

Process	Expected CO_2 (TCO_2 e)	Expected share of total (%)	Estimated CO_2 (TCO_2 e)	Estimated share of total (%)
Embodied emissions	11.2	22.2	15	25.4
Dynamic consumption	24.69	48.9	23.7	40.2
Indirect consumption	14.6	28.9	19.8	33.6
Total	50.5	100	59	100

Table 4.12: Comparison of estimations produced by MCLA (Estimated) with the measured/estimated impacts presented in table 3 of [Luccioni et al., 2022] (Expected) over the different sources of emissions

Table 4.12 compares the results MLCA produces on the different sources of CO_2 emissions with the expected results. As we can see, results for each category of emissions are pretty similar even if we obtain a higher estimate for embodied emissions due to a higher estimate of the manufacturing impacts of a GPU. More surprisingly a higher Indirect consumption estimate than the one presented in the paper estimate even if we are supposedly based on the same figures to convert from dynamic to Indirect consumption. This can be explained by the fact that the manuscript presents results only on the Idle consumption induced by the Dynamic consumption and does not include results about the Infrastructure consumption previously mentioned.

4.9 Conclusions

After these experiments trying to evaluate the validity of MLCA, we can draw some conclusions, firstly about the challenges of replicating results and then about the validity of MLCA.

4.9.1 About the replication of results

Overall, reproducing results from different papers proved way harder than expected. Indeed, unless a real effort is made to allow replication of results present in a manuscript, it is most of the time really difficult to find enough information to run estimates and reproduce these results. This is also particularly true for results produced using a measurement tool: If the hardware on which those results were produced is not detailed, it is impossible to reproduce the experiments and check the quality of the results presented. Most of the time, we were able to conduct experiments thanks to the authors, who gave us some insight about the hardware configuration of their experiments that we could not find in the manuscripts.

Even when we had enough information to run our estimates precisely enough to hopefully match the expected results, we faced multiple times important errors and inconsistencies in the data presented in different tables. This was for example the case with some results presented in [Bannour et al., 2021] and in [Cattan et al., 2022]. This was also the case to a lesser extent in [Jay et al., 2023] where reproducibility was greatly facilitated by the supplementary material provided. Still, some problems occurred and some assumptions needed to be made in order to reproduce exact results. After pointing out the problems with the data presented in [Cattan

et al., 2022], the authors conducted new experiments to resolve the problems with their data and we were able to reproduce these new results.

The question of the reproducibility of results is not new and has already been explored. For instance [Cohen et al., 2018] defined three different levels of reproducibility. The first one being the ability to reproduce the same exact *value*, this level is rarely feasible as processes are not always deterministic. The second one is the ability to obtain results close to the ones presented, it is called the *finding* level. Our experiments tried to reproduce results at this level since we often tried to find results obtained with a different method. Sometimes, when facing abnormal values in a manuscript, we would only be able to reproduce results to the level of the *conclusion* i.e. finding that an experiment that lasts longer has higher impacts than a shorter experiment on the same hardware.

[Digan et al., 2021] Proposed a list of recommendations to ensure a high level of reproducibility of the results presented in an article. Some rules that were sometimes not respected in the manuscripts we worked on where (R03) 'System metadata (e.g. RAM, CPU, OS, etc.)' ; (R04) 'Record parameters of tools' ; (R28) 'Absence of manual steps' which could explain incoherences between tables or abnormal values in a table.

In the end, we can note that the problems of the reproducibility of results also apply to evaluations of the environmental impacts of experiments.

4.9.2 About the validity of the tool

Running new experiments often required us to gather some information about a CPU not present in our database. This was not needed for GPUs. It seems like there is much more diversity in CPUs used than in GPU used. The difference in the results obtained when inputting different hardware than the one used have been explored to a certain extent in sections 4.2 and 4.3.

However, it was relatively easy to find all the information we needed when encountering a new CPU and when running estimations about GPU intensive tasks such as training NLP models, the CPU usage is often set close to 0. Moreover, CPU manufacturing does not play a huge part in the manufacturing impacts of a server in terms of GWP.

Even if the manufacturing impacts of a CPU or a GPU are not so important in terms of GWP, they are responsible for an important part of the impacts in terms of mineral resource usage (ADP).

We were unfortunately not able to find experiments to demonstrate the validity of other indicators than the Global Warming Potential.

Still, we can see that overall, we were able to reproduce results for the dynamic consumption and for the embodied impacts. These experiments also demonstrate the usability of MLCA in diverse scenarios.

— 5 —

Towards a domain wide impact evaluation: the case of French Public Health

In this chapter, we take the first step towards an evaluation of the impacts of ICT the French healthcare system. We focus on the evaluation of NLP tools and methods used for health. In order to better understand what technologies and tools are used by clinicians, healthcare researchers and administration, we conducted a series of interviews. Section 5.1 presents the methodology used for conducting this exploratory study and section 5.2 presents a synthesis of the interviews and analysis of the respondents reported experience.

5.1 Methodology of this study

The objectives of this series of interviews are to obtain an overview of ICT in the French health system with a focus on NLP tools. We want to understand the current state, ongoing transformations and plausible trajectories of use of these tools. We also want to understand the current state of reflection on the impact of these usages, and, how this reflection on the impacts of the ICT influence the use and investment towards new tools.

In order to gather all of this information, we chose a qualitative data strategy to conduct a series of semi-structured interviews ([DiCicco-Bloom and Crabtree, 2006]). We established the following protocol for conducting the interviews:
An interview lasted approximately one hour with one or two interviewers and the interviewee. We would first introduce ourselves, our objectives and the type of information we were seeking. then we would ask the following questions:

1. What is your job and background?
2. What are the digital tools you use in your work or know are being used in health?
3. What infrastructure exists or is needed to support this/these usages?
4. What is the reflection on the environmental impacts induced by this/these usages?
5. What is the reflection on the ethics of using the ICT in health?

This protocol was not strictly followed to accommodate for the specific knowledge and interests of each interviewee and to allow them to detail some points.

To identify people to contact we first looked at the existing links between LISN and French hospitals. This allowed us to identify researchers in NLP that work in different hospitals, mainly in Paris and we also looked at people working in governmental agencies with a focus on the impacts of the ICT in health. We also contacted some participants that have been recommended to us by the interviewees. We contacted 9 different persons and obtained 7 positive responses, one decline and one pending. All participants consented to having their contribution mentioned by name in this report. Table 5.1 gives an overview of the profiles of the interviewees.

Name	Background	hospital staff	NLP Researcher	Governmental agency staff	management of an EDS	Located in Paris
Christel Gerardin	CS & MD	✓	✓	✗	✗	✓
Antoine Neuraz	CS	✓	✓	✗	✗	✓
Bastien Rance	CS	✓	✓	✗	✗	✓
Romain Bey	CS	✓	✓	✗	✓	✓
Stéfan Darmoni	MD	✓	✓	✗	✓	✗
Nathalie Baudinière	MBA	✗	✗	✓	✗	✓
Brigite Seroussi	MD	✓	✓	✓	✗	✓

Table 5.1: Interview participant details. Although participants have a multidisciplinary background, the table shows their most saillant competence or expertise with respect to their current position. CS stands for Computer Science training, MD for Medical Training and MBA for business administration training

The interviewees recommended follow-up reading related to their work environment. We subsequently based ourselves on the following other resources: the reports [Haute Autorité de Santé,H.A.S., 2022, Jannot et al., 2017, CCNE and CNPEN, 2022, Délégation Ministérielle au numérique en Santé, Cellule éthique : GT6 - Numérique Responsable, 2021] and a presentation by Perseval Wajsbürt and Romain Bey about the EDS-NLP project [1].

5.2 Summary of the interviews

Overall, interviewees affiliated to a governmental agency discussed ethics, environmental impact and software used in healthcare. Interviewees involved in clinical data wharehouse projects were well aware of infrastructure cost and usage. Finally, physicians involved in NLP research described the use of ICT in healthcare research and practice. In general, Interviewees showed an interest in the measure and understing of ICT impact in healthcare. However, they had little practical knowledge of existing tools and initiatives toward sustainability.

We will present the results obtained, question by question: First, section 5.2.1 usage of ICT in the French healthcare system. Then, section 5.2.2 details the Infrastructure needed and used to support these usages. Section 5.2.3 discusses the sustainability of ICT use in Healthcare. Finally, section 5.2.4 discusses ethics questions raised by ICT use in healthcare.

[1] `http://almanach.inria.fr/seminars-en.html`, February the 17th 2023

5.2.1 What are the digital tools currently used in healthcare ?

In hostpitals, ICTs are used towards three goals: Healthcare, Research and Administration.

From the healthcare perspective, patient follow up is documented within an electronic health record ("Dossier patient Informatisé"). Currently, more than 800 proprietary computer programs with limited interoperability are used at the Assistance Publique des Hopitaux de Paris (APHP) to create and update documents stored in electronic health records.

There is a trajectory towards "zero paper" hospitals with entirely digitised administration and patient follow-up with the example of the Saint Joseph hospital already operating in this way. Digital tools bring the healthcare system a lot of hope of automating repetitive and tedious tasks such as administrative form filling or easier and faster information search.

AI technologies and rapidly evolving NLP tools such as Large Language Models also bring hope of care improvement with clinical decision support tools representing each patient profile and assisting with the discovery of similar patients to understand their possible trajectories and choose adapted care for instance.

Development of the EDS An important technology that is being developed and deployed are the EDS: "Entrepôts de Données de Santé" which could be translated to *Healthcare Data Warehouse*. An EDS consists in the pooling of data from one or multiple medical IT systems under an homogeneous format with the aim at reusing this data for direction, research or care purposes [Haute Autorité de Santé,H.A.S., 2022]. In France, the first EDS was deployed in 2008 at the Georges Pompidou European Hospital [Jannot et al., 2017]. Other hospitals throughout the country have now deployed their own EDS (e.g. Montpellier, Rennes, Brest, Necker in Paris) or are working towards deployment (e.g. Strasbourg). Regulation started with the first CNIL authorisation in 2016/17 for APHP. Rouen's EDS was started in 2018 and was authorised in 2020. The legal framework surrounding the EDS was created in 2021. This is a recent technology developing at a rapid pace. Currently in France, the teams working on the EDS projects represent in median 7.5 full time jobs per EDS.

The EDS of the APHP currently contains (Perseval Wajsbürt and Romain Bey's presentation):

- in terms of administrative documents: 11 Millions patient identities for 40 Millions stays
- in terms of medico-economic data (PMSI, in particular used for billing purposes): 40 Millions medical acts and 40 Millions diagnosis
- in terms of unstructured data: 100 Million clinical reports and 30 millions of imagery results
- in terms of structures data: over 2 Billion biology results.

The EDS are aiming at building bridges between research, care and direction. Incorporating new data in an EDS is a three step process. First, there is data collection and copy. Then, there is data transformation (with harmonisation of the different program output, de-identification of the data and structuration of the data). Finally, there is the provision of the data in specialised "datamarts" for each type of data usage (studies, monitoring, applicatives). For all of these steps

there is an important need of NLP tools. Such tools are developed in open-source projects such as the MEDKIT[2] library or the EDS-NLP[3]/PDF/Pseudo projects.

Usages of the EDS Currently the EDS are mainly used for research purposes. Figure 5.1 shows the distribution of the types of queries in the Rouen EDS. and Figure 5.2 shows the finality of the different research projects using an EDS.

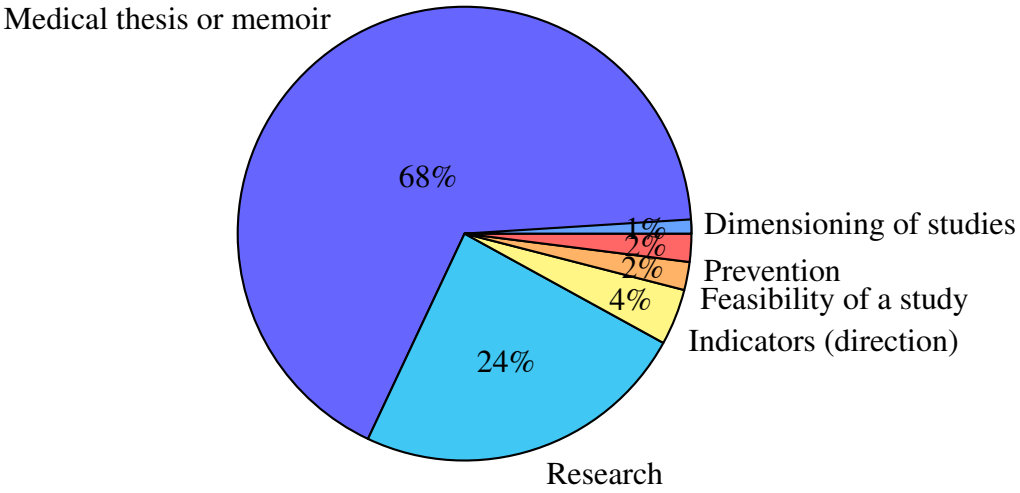


Figure 5.1: Distribution of purpose of the queries in the EDS of Rouen

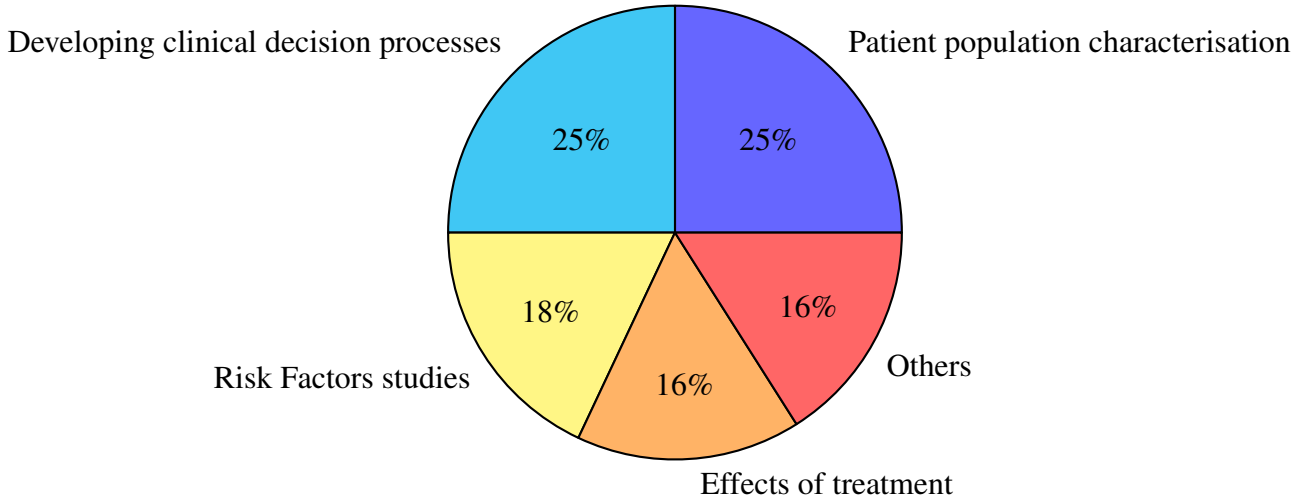


Figure 5.2: Distribution of finality of research projects using an EDS in France [Haute Autorité de Santé,H.A.S., 2022].

A lot of researchers are working on clinical decision making tools with for instance the development of medical language models specialised for patient representation. These models need to be frequently re-trained (once a week for instance) in order to update them with the new data coming in the EDS.

[2] https://heka.gitlabpages.inria.fr/medkit/
[3] https://github.com/aphp/edsnlp

Objectives of the tools The objectives of these tools are mostly to save time for clinicians with the automatising of data copying and similar possibly tedious tasks such as similar patient search. These tools can be used in multidisciplinary consultation meetings. Clinicians can also sometimes use tools and interfaces for the EDS to more rapidly query information and consult documents even if this usage is not endorsed and if dedicated software exists (but is slower and less convenient).

Another hope coming from the EDS is the development of efficacy metrics for the healthcare system in order to monitor hostpitals not on their consumption of drugs and medical acts but on their quality of care. There is the objective of facilitating monitoring in general. For research, EDS also come with the hope of enabling multi-centric studies (studies that span across multiple hospitals).

In 2019, the French Health ministry published a road map about the deployment of the ICT in healthcare. It confirms a trend towards increasing ICT deployment in the healthcare system. It is to be noted that the EDS are not so much encouraging ICT deployment as building on it.

5.2.2 What infrastructure exists or is needed to support this/these usages?

System duplication In the big hostpitals, there exists two parallel IT systems, The Care system and the EDS. These two systems use mostly separate infrastructure and have very different constraints.

On the one hand, The Care system needs to enforce high Reliability and Multiple data duplication for data storage. There are also dedicated servers for telephony and network access in the hospital. There are also some Care services that need computation power such as imagery services of resuscitation monitoring.

On the other hand, the EDS infrastructure needs more computing power for data processing and reuse and have less strict needs on data storage quality. The EDS are fed with a partial copy of the Care database with mostly text documents. Some other data sources such as imagery can be incorporated in a project-wise manner.

Computing power There exists some computing facilities internal to the hospitals. These computing facilities are used for Care-related computing power needs such as imagery. Teams that work on the EDS tend to also acquire smaller clusters at a smaller scale such as to the scale of a single hospital and not the entire hospital consortium. Researchers can use these clusters for non-critical and non-real time computation such as Model training in periods of lower computing power demand.

A turning point in infrastructure development Because of the increasing deployment of the ICT in the hospitals, there is an increasing demand for HPC. As it is not the objective of the hospitals to host computing facilities, there are discussions and contracts that are signed with Cloud service providers such as OVH. These contracts and use of private computing facilities pose questions about the security of the data transferring through these servers. They also pose questions about the sovereignty of the hostpitals and possible dependencies to private actors. hostpitals also do not use public computing facilities such as the Jean Zay supercomputer for privacy and security reasons.

Because we are in a period of growth in the presence and usage of the ICT in healthcare, the infrastructure is evolving to accompany this growth. Therefore, if we want to understand the possible future impacts of the ICT in the healthcare system, we need to anticipate this change.

5.2.3 What is the reflection on the environmental impacts induced by these usages?

Policies known by the researchers NLP researchers were not aware of important policies deployed at the scale of the APHP for instance. They are currently at a stage where they mostly try to always measure the impacts of the models they train. Some known policies concern extending the service life of the servers they use and prioritising the reuse of existing hardware in order to reduce new hardware purchase. Even if the environmental impact of the IT services of the APHP is not the main concern, the topic is gaining interest and a company should be mandated to perform an audit of the whole IT systems of the APHP.

Carbon footprint of the hospitals In the hospitals there exist instances dedicated to the development of the ICT that are in contact with the health ministry. Following the road map about the deployment of the ICT in healthcare, a report was published by governmental agencies in 2021: [Délégation Ministérielle au numérique en Santé, Cellule éthique : GT6 - Numérique Responsable, 2021]. It presents some figures about the impacts of the ICT in the healthcare system. For instance the IT system of an average Centre Hospitalier Universitaire (CHU) represented more than 5% of the carbon footprint of the hospital. Also, the Carbon footprint of the IT systems of all French hospitals in 2018 was evaluated to 190,000 tCO_2e.

One important point of the governmental strategy about the environmental impacts of the ICT in healthcare is that the reduction of the environmental impact should not come from a reduction in the number of applications but from optimising the usages with two main axes: The eco-conception of care programs and sobriety of the usages (use only what is necessary).

One point that was frequently raised is the need to evaluate the costs (or impacts) to benefits (in terms of care quality) balance of a new digital solution before implementing it. There is a need to be vigilant about the ICT in healthcare for it to maintain a positive balance and ICT use needs to be moderated with regards to ethics implications such as eco-responsibility and non-malfeasance.

Existing policies Governmental agencies developed two eco-scores, with the support of evolving regulation that enforced notions about the sustainability of health applications. The first one [4] quantifies the impacts of using an health application such as *Docotolib* (for scheduling doctor consultations) or applications that do follow-up of patients with diabetes. This score not only considers the carbon footprint but also the land and water use. It was created with the aim of educating software editors to eco-construct their applications and therefore tries to identify the parts of the application responsible for a high share of the impacts to give editors insight on what could be improved. The other tool that is developed is aimed at computing a simplified LCA of the IT system of an hospital based on an inventory of the devices and computer programs present in the system. This tool faces challenges with the absence of data on the manufacturing impacts of specialised devices such as MRIs.

[4] `https://ecoscore-appli.esante.gouv.fr/`

There can be some sort of rebound effect were digitisation can overlap with existing storage. Example of anatomopathology researchers that are concerned about the impacts of digitising the glass slides they are already physically storing.

Environmental challenges are a subset of the Ethical challenges posed by the usages of the ICT in healthcare.

5.2.4 Ethical challenges posed by the usage of the ICT in healthcare.

As previously mentioned, a turning point in the deployment of the ICT in the healthcare system was the 2019 roadmap published by the French Health ministry. This roadmap included an important place for the reflection about the Ethics of the ICT in healthcare. Subsequently, in the Inter-ministerial delegation for digital health, multiple task forces were created. Firstly, to create different sectoral referential to defining what it means to talk about the ethics of the ICT in healthcare. This meant for instance a referential for city healthcare. There were also some transversal questions such as the ethics of AI or the ethics of telemedicine. The Inter-ministerial delegation for digital health and the digital health agency (Agence du Numérique en Santé) are working on the ethics of software edition. Other agencies such as the Haute Autorité de Santé, are more focused on the ethics of the usages once software is deployed. In 2023, there was a legislation change in the "Loi de Financement de la Sécurité Sociale". Article 58 of this law rendered not only the security and interoperability but also the ethics enforceable to medical software editors.

Important points in the ethics of software edition include the accessibility and inclusivity of the computer program, the eco-conception, and the consent and transparency about the data processing.

About the ethics of AI, some important points are autonomy in decision-making, consent of the people whose data are included in training datasets or explainability of the models.

Researchers insisted on the importance of data security and privacy concerns. Models are considered as personal data and their usage and access are regulated as such. Another question concerns the accesses to data stored in the EDS. We were told of the example of psychiatrists that do not enter the most sensible data about their patients because they do not want any researcher to access them.

About the reliability of AI systems, regulations have evolved towards considering decision helping tools in the same way as other medicines. They therefore need to pass a clinical trial to ensure clinical efficacy and not only theoretical quality on a test set. This point is also slowing the deployment of number of tools created by researchers but that are never put in production because research budgets do not account for the costs of running clinical trials. One other important question about the efficacy of AI systems is the one of biases. Should a tool that in average improves quality of care but is biased be used?

This also poses a question of responsibility: the users of decision aid system could tend to rely on the system's proposal even if they do not think it is the correct decision. Indeed, if they were correct in thinking the system's proposal was erroneous, they can say that it is the system's fault. In the opposite case, if they are mistaken when not following the system's proposal they is a sort of double guilt.

On the explainability of AI, one could argue that it does not raise new questions and draw a parallel with other drugs. If care is provably improved when using the drug/AI, we will

tend to use it even if we are not fully able to understand the mechanisms that permit this care improvement. However, the lack of explainability can have a negative impact on the autonomy in decision making.

There is also a challenge when replacing humans with automatic tools. For instance, in the case of a mammography, the procedure requires two reads, one of which must be done by a highly specialised clinician. If we use an automatic tool for the first read, how can we ensure that the specialised clinicians required for the second reads are still trained. Also, how can we ensure that the tool used for the first read will still be reliable if there are slight modifications in the input such as change in dimension of the image.

IT systems of the hospitals and the use of the ICT in the hospitals can be really complex. There is therefore a challenge of communication with the general public and with the patients.

A really important ethics challenge about the usage of the ICT by the clinicians, is the breakdown in the clinician-patient relationship induced by the presence of digital tools. Numerous clinicians complain about the ever increasing number of data they are expected to input in the digital tools and that give them less time with the patients. There was for instance a coding strike where clinicians refused to enter the information expected because they felt like they were expected to fill bills all day long in addition to their workload.

This also highlights the risk of some sort of rebound effect when deploying new digital solution that do not substitute existing solutions but simply add-up, with clinicians needing to copy the same information in multiple places.

About the risks of digital dependency, there are a lot of solutions to ensure different levels of degraded operating of the hospitals when available power decreases. However, a really complex problem is the one of cyber-security and the vulnerability to attacks such as ransomwares that can fully block one hospital.

5.3 Conclusions about the interviews

In this Chapter, we have taken the first steps towards a domain wide evaluation of the impacts of NLP methods used in the French Healthcare system. To do so, we have conducted a series of semi-structured interviews with the aims of getting a complete panorama of ICT use in the Healthcare system, the infrastructure these usages require, the place for ethics and environmental issues in decision making and the possible development trajectories.

This series of interviews lead us to meet three types of profiles, with multidisciplinary backgrounds, some with a more important medical training and some with mostly an engineering / computer science background. Computer science researchers working in hospitals were able to provide many details about the usage of ICT for research purpose, the development process of tools using NLP methods, their purpose and expected use; People implied in the management of EDS projects could give detailed information about the infrastructure needed and being deployed but also the different types of usages of the EDS and their distribution; Finally, people working in governmental agencies provided us with a lot of information about policy making and the place for ethics and, in particular, environmental issues in the decision process.

After this series of interviews, a certain level of saturation (a decreasing amount of new information obtained after a new interview) was reached. This suggests that we have obtained a relatively good quality overview of ICT usage in French hospitals as well as the associated stakes.

Still, in order to get a broader view, maybe not so focused on NLP use, this series of interviews could be complemented by interviewing people with different profiles such as clinicians not working on the development of the tools and using them or even patients confronted with these new methods. This would also allow us to have a better understanding of the social transformations induced by digitization in hostpitals.

Some take-away messages that emerged from this series of interviews are:

1. ICT are ubiquitous within French healthcare (healthcare organisation, clinical practice, public health research).

2. The new availability of clinical data wharehouses (EDS) places the system at a turning point towards new deployment/uses of ICTs in healthcare.

3. Healthcare professionals seem open to sustainable solutions but there is a need for resources (e.g. expertise and time) to adequately implement them.

— 6 —
Conclusion

In this work, we interested ourselves in the evaluation of the environmental impacts of NLP applications. We first created a tool named MLCA aimed at NLP researchers to allow them to estimate, before undertaking a new project, the potential impacts of the computations they would run. This would allow them to then put these potential impacts in balance with the expected benefits of their work in order to decide if the potential benefits are worth the potential impacts. This tool was built upon existing methodologies and tools with the objective of integrating LCA considerations in the process. Adding these considerations allows to account for the manufacturing impacts of the hardware used and also to not only evaluate the carbon footprint but also other environmental impacts such as resource depletion.

This tool is then evaluated on a series of different case studies demonstrating its usability in different cases and the quality of the results it produces. This series of experiments also highlight challenges for the reproduction of evaluations of the environmental impacts. Such challenges in the reproduction of scientific results are not new but they also apply to the evaluations of the environmental impacts of experiments.

A series of semi-structured interviews was also conducted, giving us a wide view of the current state of ICT use in the French healthcare system, its probable evolution and the place for ethical and environmental considerations in the creation and deployment of new ICT solutions. These first steps open towards a domain-wide evaluation of the impacts of NLP applications on the case of the French Healthcare system.

Bibliography

[Anthony et al., 2020] Anthony, L. F. W., Kanding, B., and Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models.

[Avelar et al., 2012] Avelar, V., Azevedo, D., French, A., and Power, E. N. (2012). Pue: a comprehensive examination of the metric. *White paper*, 49.

[Bannour et al., 2021] Bannour, N., Ghannay, S., Névéol, A., and Ligozat, A.-L. (2021). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *EMNLP, Workshop SustaiNLP*, Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing, Punta Cana, Dominican Republic.

[Beede et al., 2020] Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

[Bender et al., 2021] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

[Boavizta, 2021] Boavizta (2021). Numérique et environnement : Comment évaluer l'empreinte de la fabrication d'un serveur, au-delà des émissions de gaz à effet de serre?̃

[Bol et al., 2021] Bol, D., Pirson, T., and Dekimpe, R. (2021). Moore's law and ict innovation in the anthropocene. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 19–24.

[Brady et al., 2013] Brady, G. A., Kapur, N., Summers, J. L., and Thompson, H. M. (2013). A case study and critical assessment in calculating power usage effectiveness for a data centre. *Energy Conversion and Management*, 76:155–161.

[Bruijn et al., 2002] Bruijn, H., Duin, R., Huijbregts, M. A. J., Guinee, J. B., Gorree, M., Heijungs, R., Huppes, G., Kleijn, R., Koning, A., van Oers, L., Sleeswijk, A. W., Suh, S., and de Haes, H. A. U. (2002). *Handbook on Life Cycle Assessment - Operational Guide to the ISO Standards*. Springer Dordrecht.

[Cattan et al., 2022] Cattan, O., Ghannay, S., Servan, C., and Rosset, S. (2022). Benchmarking transformers-based models on french spoken language understanding tasks. In *INTERSPEECH 2022*, Incheon, South Korea.

[Cattan et al., 2021] Cattan, O., Servan, C., and Rosset, S. (2021). On the usability of transformers-based models for a french question-answering task. In Angelova, G., Kunilovskaya, M., Mitkov, R., and Nikolova-Koleva, I., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, pages 244–255. INCOMA Ltd.

[CCNE and CNPEN, 2022] CCNE and CNPEN (2022). Diagnostic Médical et Intelligence Artificielle : Enjeux Éthiques. Avis commun du CCNE et du CNPEN, Avis 141 du CCNE, Avis 4 du CNPEN.

[Clément et al., 2020] Clément, L.-P. P.-V., Jacquemotte, Q. E., and Hilty, L. M. (2020). Sources of variation in life cycle assessments of smartphones and tablet computers. *Environmental Impact Assessment Review*, 84:106416.

[Cohen et al., 2018] Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., Ide, N., Névéol, A., Grouin, C., and Hunter, L. E. (2018). Three dimensions of reproducibility in natural language processing. In *Lrec... international conference on language resources & evaluation:[proceedings]. international conference on language resources and evaluation*, volume 2018, page 156. NIH Public Access.

[COMETS, Comité d'éthique du CNRS, 2022] COMETS, Comité d'éthique du CNRS (2022). AVIS n°2022-43, Intégrer les enjeux environnementaux à la conduite de la recherche – une responsabilité éthique.

[David et al., 2010] David, H., Gorbatov, E., Hanebutte, U. R., Khanna, R., and Le, C. (2010). Rapl: Memory power estimation and capping. In *Proceedings of the 16th ACM/IEEE International Symposium on Low Power Electronics and Design*, ISLPED '10, page 189–194, New York, NY, USA. Association for Computing Machinery.

[DiCicco-Bloom and Crabtree, 2006] DiCicco-Bloom, B. and Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, 40(4):314–321.

[Digan et al., 2021] Digan, W., Névéol, A., Neuraz, A., Wack, M., Baudoin, D., Burgun, A., and Rance, B. (2021). Can reproducibility be improved in clinical natural language processing? a study of 7 clinical nlp suites. *Journal of the American Medical Informatics Association*, 28(3):504–515.

[Dinarelli et al., 2022] Dinarelli, M., Naguib, M., and Portet, F. (2022). Toward low-cost end-to-end spoken language understanding. *arXiv preprint arXiv:2207.00352*.

[Délégation Ministérielle au numérique en Santé, Cellule éthique : GT6 - Numérique Responsable, 2021] Délégation Ministérielle au numérique en Santé, Cellule éthique : GT6 - Numérique Responsable (2021). L'impact environnemental du numÉrique en santÉ. accessible at `https://esante.gouv.fr/sites/default/files/media_entity/documents/RAPPORT_GT6_VF.pdf`.

[Escudié et al., 2017] Escudié, J.-B., Rance, B., Malamut, G., Khater, S., Burgun, A., Cellier, C., and Jannot, A.-S. (2017). A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. volume 17.

[European Commission et al., 2010] European Commission, Joint Research Centre, and Institute for Environment and Sustainability (2010). *International Reference Life Cycle Data System (ILCD) Handbook - General guide for Life Cycle Assessment - Detailed guidance.* EUR 24708 EN. Luxembourg. Publications Office of the European Union. First edition March 2010. EUR 24708 EN. Luxembourg. Publications Office of the European Union.

[Evain et al., 2021] Evain, S., Nguyen, M. H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021). Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, NeurIPS 2021 Datasets and Benchmarks Track, on-line, United States.

[Freitag et al., 2021] Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., and Friday, A. (2021). The real climate and transformative impact of ict: A critique of estimates, trends, and regulations. *Patterns*, 2(9):100340.

[Frischknecht et al., 2015] Frischknecht, R., Wyss, F., Knöpfel, S. B., Lützkendorf, T., and Balouktsi, M. (2015). Cumulative energy demand in lca: the energy harvested approach. *International Journal of Life Cycle Assessment 20*, page 957–969.

[Gossart, 2015] Gossart, C. (2015). Rebound effects and ict: A review of the literature. In Hilty, L. M. and Aebischer, B., editors, *ICT Innovations for Sustainability*, pages 435–448, Cham. Springer International Publishing.

[Gröger et al., 2021] Gröger, J., Liu, R., Stobbe, L., Druschke, J., and Richter, N. (2021). *Green Cloud Computing: lebenszyklusbasierte Datenerhebung zu Umweltwirkungen des Cloud Computing: Abschlussbericht*. Umweltbundesamt.

[Gupta et al., 2020] Gupta, U., Kim, Y. G., Lee, S., Tse, J., Lee, H.-H. S., Wei, G.-Y., Brooks, D., and Wu, C.-J. (2020). Chasing carbon: The elusive environmental footprint of computing.

[Hauschild, 2015] Hauschild, M. Z. (2015). Better – but is it good enough? on the need to consider both eco-efficiency and eco-effectiveness to gauge industrial sustainability. *Procedia CIRP*, 29:1–7. The 22nd CIRP Conference on Life Cycle Engineering.

[Haute Autorité de Santé,H.A.S., 2022] Haute Autorité de Santé,H.A.S. (2022). Entrepôts de données de santé hospitaliers en france.

[Henderson et al., 2020] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, 21(1).

[Hilty and Aebischer, 2015] Hilty, L. M. and Aebischer, B. (2015). Ict for sustainability: An emerging research field. In Hilty, L. M. and Aebischer, B., editors, *ICT Innovations for Sustainability*, pages 3–36, Cham. Springer International Publishing.

[Jannot et al., 2017] Jannot, A.-S., Zapletal, E., Avillach, P., Mamzer, M.-F., Burgun, A., and Degoulet, P. (2017). The georges pompidou university hospital clinical data warehouse: a 8-years follow-up experience. *International journal of medical informatics*, 102:21–28.

[Jay et al., 2023] Jay, M., Ostapenco, V., Lefèvre, L., Trystram, D., Orgerie, A.-C., and Fichel, B. (2023). 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*, pages 1–13, Bangalore, India. IEEE.

[Kaack et al., 2021] Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., and Rolnick, D. (2021). Aligning artificial intelligence with climate change mitigation. An updated version is now published with Nature Climate Change and can be found as:Kaack, L.H., Donti, P.L., Strubell, E. et al. Aligning artificial intelligence with climate change mitigation. Nat. Clim. Chang. 12, 518–527 (2022). https://doi.org/10.1038/s41558-022-01377-7.

[Lacoste et al., 2019] Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning.

[Lannelongue et al., 2020] Lannelongue, L., Grealey, J., and Inouye, M. (2020). Green algorithms: Quantifying the carbon footprint of computation.

[Ligozat et al., 2021] Ligozat, A.-L., Lefèvre, J., Bugeau, A., and Combaz, J. (2021). Unraveling the hidden environmental impacts of ai solutions for environment.

[Loubet et al., 2023] Loubet, P., Vincent, A., Collin, A., Dejous, C., Ghiotto, A., and Jego, C. (2023). Life cycle assessment of ict in higher education: a comparison between desktop and single-board computers. *The International Journal of Life Cycle Assessment*, pages 1–19.

[Luccioni et al., 2022] Luccioni, A. S., Viguier, S., and Ligozat, A.-L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model.

[NVIDIA Corporation, 2021] NVIDIA Corporation (2021). Nvidia Management Library (NVML). accessed may 2023 at `https://developer.nvidia.com/nvidia-management-library-nvml`.

[Parcollet and Ravanelli, 2021] Parcollet, T. and Ravanelli, M. (2021). The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. working paper or preprint.

[Patterson et al., 2022] Patterson, D., Gonzalez, J., Hölzle, U., Le, Q. H., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *techrxiv*.

[Rasoldier et al., 2022] Rasoldier, A., Combaz, J., Girault, A., Marquet, K., and Quinton, S. (2022). How realistic are claims about the benefits of using digital technologies for ghg emissions mitigation? *LIMITS'22: Workshop on Computing within Limits*.

[Sala et al., 2020] Sala, S., Crenna, E., Secchi, M., and Sanyé-Mengual, E. (2020). Environmental sustainability of european production and consumption assessed against planetary boundaries. *Journal of Environmental Management*, 269:110686.

[Schmidt et al., 2022] Schmidt, V., Goyal-Kamal, Courty, B., Feld, B., SabAmine, kngoyal, Zhao, F., Joshi, A., Luccioni, S., Léval, M., Bogroff, A., de Lavoreille, H., Laskaris, N., LiamConnell, Wang, Z., Saboni, A., Catovic, A., Blank, D., Stęchły, M., alencon, JPW, MinervaBooks, SangamSwadik, M., H., MarionCoutarel, Pollard, M., McCarthy, C., Husom, E. J., Vicente, F., and Tae, J. (2022). mlco2/codecarbon: v2.1.4.

[Schödwell et al., 2018] Schödwell, B., Zarnekow, R., Liu, R., Gröger, J., and Wilkens, M. (2018). Kennzahlen und indikatoren für die beurteilung der ressourceneffizienz von rechenzentren und prüfung der praktischen anwendbarkeit. *Umweltforschungsplan des Bundesministeriums für Umwelt, Naturschutz, Bau und Reaktorsicherheit; Umweltbundesamt: Dessau-Roßlau, Germany*.

[Schwartz et al., 2020] Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Commun. ACM*, 63(12):54–63.

[Selvan et al., 2022] Selvan, R., Bhagwat, N., Wolff Anthony, L. F., Kanding, B., and Dam, E. B. (2022). Carbon footprint of selecting and training deep learning models for medical image analysis. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 506–516, Cham. Springer Nature Switzerland.

[Sphera, 2021] Sphera (2021). Life Cycle Assessment Dell Servers R6515, R7515, R6525, R7525.

[Strubell et al., 2019] Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp.

[Thinkstep, 2019] Thinkstep (2019). Life Cycle Assessment of Dell R740.

[van Oers et al., 2020] van Oers, L., Guinée, J. B., and Heijungs, R. (2020). 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*, 25:294–308.

[Verdecchia et al., 2023] Verdecchia, R., Sallou, J., and Cruz, L. (2023). A systematic review of green ai.

[Wu et al., 2022] Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. (2022). Sustainable ai: Environmental implications, challenges and opportunities.

[Yu et al., 2022] Yu, J.-R., Chen, C.-H., Huang, T.-W., Lu, J.-J., Chung, C.-R., Lin, T.-W., Wu, M.-H., Tseng, Y.-J., and Wang, H.-Y. (2022). Energy efficiency of inference algorithms for clinical laboratory data sets: Green artificial intelligence study. *J Med Internet Res*, 24(1):e28036.