Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

MASTER IN INNOVATION AND RESEARCH FOR SUSTAINABILITY

MASTER´S FINAL WORK

DISSERTATION

CARBON FOOTPRINT OF ARTIFICIAL INTELLIGENCE (AI) MODELS: ESTIMATION AND REDUCTION APPROACHES

MINA VILDAN ŞARK

FEBRUARY - 2025

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

MASTER IN INNOVATION AND RESEARCH FOR SUSTAINABILITY

MASTER´S FINAL WORK

DISSERTATION

CARBON FOOTPRINT OF ARTIFICIAL INTELLIGENCE (AI) MODELS: ESTIMATION AND REDUCTION APPROACHES

MINA VILDAN ŞARK

SUPERVISION:
TIAGO CAPELA LOURENÇO

FEBRUARY - 2025

To the souls who take life's
sourest lemons and resemble
lemonade through dedication
and hard work, inspiring
others along the way.

GLOSSARY

AI - Artificial Intelligence

CO_2eq - Carbon dioxide equivalent

CPU - Central Processing Unit

DL - Deep Learning

DPU - Data Processing Unit

GHG - Greenhouse gas

GPT - Generative Pre-trained Transformer

GPU - Graphics Processing Unit

ICTs - Information and Communication Technologies

kWh - Kilowatt-hour

LLM - Large Language Models

ML - Machine Learning

NLP - Natural Language Processing

PUE - Power Usage Effectiveness

RL - Reinforcement Learning

t CO2eq - Tonnes of carbon dioxide equivalent

TPU - Tensor Processing Unit

ABSTRACT

Artificial intelligence (AI) has become an integral part of our lives, with its models increasingly used across various sectors. However, there is limited research on its environmental and sustainability costs. With this study, we intend to advance the understanding of how carbon emissions associated with AI models can be measured and reduced. This is done through literature review, the analysis of real-world case studies, and the elicitation of expert stakeholder's perspectives. This combination of methodologies enables a comprehensive evaluation of the practices currently used to calculate the carbon emissions associated with the training and inference of AI models, as well as of the strategies applied to mitigate their carbon footprint. Our results show that: (i) current estimations of the carbon emissions of training and deploying AI models are flawed due to the limited understanding of their complexity, unavailability of coherent estimation frameworks and incomplete data availability; (ii) the adoption of standardized emission's reporting among tech companies is a necessary step towards more accurate calculations; (iii) the implementation of carbon reduction techniques such as algorithm, hardware and data centre optimization can serve as possible solutions to minimize the carbon emissions of these models; (iv) all stakeholders involved in the AI model's lifecycles need to be publicly informed about the emission impact and actively engaged in mitigation efforts. This thesis acknowledges the growing data and computational resources that accompany the current advancement of AI models and discusses how this trend may affect their long-term environmental sustainability. Future research and research policies should focus on addressing major gaps in AI model carbon emission calculation and on developing effective mitigation strategies and technological innovation, which can support companies' efforts towards a more sustainable AI.

KEYWORDS: AI Models; Carbon Emissions; Sustainable AI; Emission Calculators; Carbon Reduction Frameworks; Stakeholder Engagement

JEL CODES: C80; M15; O33; Q01; Q55; Q56

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank everyone who contributed to making this Master's Final Work possible. A special thanks to my supervisor, Tiago Capela Lourenço, for his invaluable guidance, insights, and unwavering support throughout this work. I am also grateful to the professors of the Master's in Innovation and Research for Sustainability for sharing their knowledge and expertise. Lastly, I extend my deepest gratitude to my loved ones for their constant encouragement and support throughout this journey.

CARBON FOOTPRINT OF ARTIFICIAL INTELLIGENCE (AI) MODELS: ESTIMATION AND REDUCTION APPROACHES

By Mina Vildan Şark

Artificial intelligence (AI) models are widely used in several sectors, but research on their environmental and sustainability costs is limited. Using a combination of literature review, real-world case studies, and expert insights, this study explores how the carbon footprint of AI models can be measured and reduced. Findings indicate current emission estimates face challenges due to model complexities, inconsistent frameworks, and insufficient data. The importance of implementing standardized reporting, carbon reduction techniques, and stakeholder engagement is highlighted. Further research and policy developments should aim to enhance the calculation of AI carbon emissions and the promotion of sustainable AI innovation.

1. INTRODUCTION

1.1. Background

Climate change is one of the most pressing challenges of our time, and assessing and reducing carbon emissions is crucial to mitigate its impact. As digital services proliferate across various sectors, the environmental impact of Information and Communication Technologies (ICTs) has become a matter of concern (Berthelot et al., 2024). Recent estimates claim that ICT contributes between 1.5% and 4% of global carbon emissions (Ayers et al., 2024). Such a wide range is due, in part, to the difficulty in accurately estimating emissions, given the distributed nature of the global computing infrastructure. Artificial Intelligence (AI) is playing a significant role in the growing carbon footprint of the ICT sector, particularly because of large-scale generative models (Bolón-Canedo et al., 2024). A recent surge in large-scale generative models, such as ChatGPT and DeepSeek, has attracted particular attention due to the significant computational resources required for their training. Additionally, their deployment increases the use of end-user devices, networks, and data centres, as these models are offered as web services. This expanded usage contributes to global warming, heightens the demand for metals, and increases energy consumption (Berthelot et al., 2024). Considering a real-life example on ChatGPT which is widely being used, and it is found that a single query on the model produces 0.382g CO_2eq (10,000,000 queries per day is taken as a base) (Tomlinson et al., 2024). The growing adoption of AI models like ChatGPT raises considerable concern about their impact on increasing carbon emissions.

1.2. Problem statement

As the use of AI models become more widespread, the growing concern about their environmental impacts highlight the urgent need for sustainable AI, defined as the extent to which AI technology is developed to meet present needs without compromising future generations (Bjørlo et al., 2021). To address these concerns, van Wynsberghe (2021) outlines two critical branches of sustainable AI: AI for sustainability (e.g., when applied to decarbonization systems) and the sustainability of AI (i.e., which focuses on reducing carbon emissions and computing power). This indicates that AI can be a double-edged sword regarding the environment, since it can be very helpful in lowering carbon emissions and responding to climate change effects, while being a vast carbon emitter. Therefore, it is crucial that sustainability frameworks are used so that the potential benefits are balanced against the environmental impact of AI (Gaur et al., 2023).

The complexity of AI models is growing and so is their demand for energy-intensive computational power, the amounts of data used in their training and deploying, and the water consumption needed to refrigerate data centres that hold training data, which results in significant carbon and other greenhouse gas (GHG) emissions, thus posing serious sustainability challenges (Bolón-Canedo et al., 2024). On the other hand, there is an increasing number of studies that use AI models (especially machine learning) to support the reduction of carbon emission across multiple societal sectors and applications, creating a positive environmental impact. However, many of these studies overlook the carbon emissions generated by the training and use of the models themselves, which may result in an overall negative environmental impact, as indicated by Delanoë et al. (2023).

1.3. Research objectives

Given the above stated challenges, this thesis focuses on the sustainability of AI, specifically examining the carbon emissions associated with the training and deployment of AI models. The primary objective is to address the research gap in understanding the carbon emissions generated by AI models throughout their lifecycles, particularly during the training and inference phases. This study also aims to raise awareness among technology users, researchers, policymakers, and practitioners about the environmental challenges posed by AI models and considering sustainable practices. By shedding light on the challenges and opportunities related to the carbon emissions of AI models, this

thesis contributes to the literature on the half of sustainable approach to AI models. It also informs the research policy on the need for standardized reporting, transparent methods, and regulatory framing for greener AI use. It contributes to more informed policies by analysing measurement techniques, systemic reporting challenges, and potential reduction strategies.

1.4. Research questions

As a relatively new field, research about carbon emissions associated with the training and deploying of AI models remains limited. To address this gap, this thesis explores how the carbon footprint of AI models can be measured and reduced, highlights systemic reporting and calculation challenges, and examines the level of involvement and awareness among stakeholders (e.g., AI researchers, data centre managers, and cloud service providers).

Two main research questions (and sub-questions) are addressed in this thesis:

1. How can the carbon footprint of AI models be measured?

a. What are the systemic reporting and calculation challenges associated with the carbon emissions of AI models?

To measure the carbon emissions associated with AI model lifecycles, primarily on training and inference phases, this question examines the methods and tools currently available and commonly used. It also examines the measurement factors such as model size, amount of data, data centre location, and type of hardware use, among others. With the sub-question, this thesis explores the limitations in the current implications, and lack of standardized reporting guidelines.

2. How can the carbon footprint of AI models be reduced?

a. How involved and aware are stakeholders in the effort to mitigate carbon emissions of AI models?

Potential solutions and recommendations to reduce the carbon emissions of AI models are addressed in this question. Such techniques include the optimization of algorithms, energy-efficient models, the use of renewable energy sources for data centres, and sustainable training practices. With the sub-question, this thesis explores the current level of stakeholder involvement and awareness in the carbon emission reduction of AI models.

2. LITERATURE REVIEW

2.1. Artificial Intelligence (AI)

We name ourselves Homo sapiens, the wise species, highlighting the central role of intelligence in defining humanity. For centuries, humans have sought to understand how we think, exploring how a small amount of matter can perceive, comprehend, predict, and interact with a world far more complex than itself (Russell & Norvig, 2020). This curiosity gave rise to the desire to replicate human intelligence, leading to the introduction of "Artificial Intelligence" technology. The term AI was first introduced by John McCarthy in 1955, who defined it as "the science and engineering of making intelligent machines, especially intelligent computer programs." (McCarthy, 2007). It goes beyond simply by understanding intelligence; it builds intelligent entities capable of performing tasks autonomously (Russell & Norvig, 2020). It is the ability of a machine to display human-like capabilities such as reasoning, learning, planning and creativity. This allows technical systems to sense their surroundings, interpret what they observe, solve problems, and take action to accomplish a specific goal (European Parliament, 2020). Today, AI is one of the most prevalent and rapidly advancing technologies. Therefore, it is important to learn about its sustainability, including its environmental impact (e.g., climate change) and implications to national and European research and innovation policies.

2.2. The landscape of Artificial Intelligence

The landscape of artificial intelligence is vast, incorporating machine learning, with deep learning and reinforcement learning representing small but increasingly important segments of this broader domain, driven by advancements in complex problem-solving and real-world applications (Sarajcev et al., 2022) (Figure 1). Machine Learning (ML) is a subfield of AI that includes training algorithms for making predictions based on data. It is applied in an extensive range of tasks, incorporating classification (categorizing data into predefined labels), regression (predicting continuous values), and clustering (segmenting similar data points together without prior labels) (Jason Bell, 2022). Deep Learning (DL), a subdivision of ML, incorporates training artificial neural networks to execute typical tasks such as image classification, speech recognition and object detection (Shahinfar et al., 2020). Natural Language Processing (NLP) allows computers to process human language, such as text or voice, and synthesize a relevant response in the form of

speech and natural language (Khan et al., 2023). It relies mainly on DL techniques and the best known and used applications are Chat GPT-4 and Google's Gemini. Reinforcement learning (RL), a specialized subset of DL, trains models to respond to specific scenarios which is effective for decision-making tasks (Sutton & Barto, 2018).

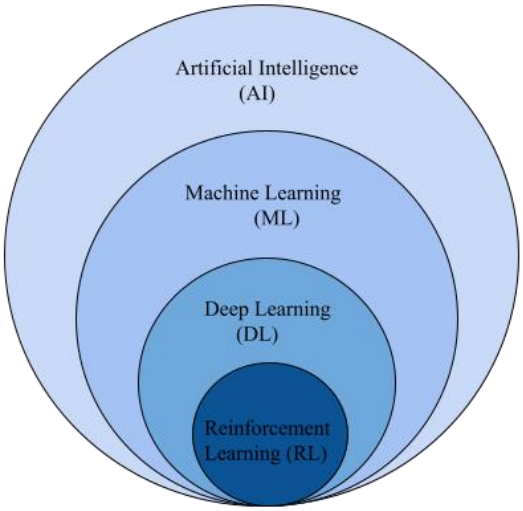

FIGURE 1 – Artificial Intelligence landscape

Natural Language Processing (NLP), a subfield of AI, is not explicitly included since it focuses specifically on language processing tasks and is not a broader ML or DL methodology.

Source: Sarajcev et al. (2022)

2.3. Artificial Intelligence model's lifecycle

AI models are programs or algorithms that enable machines to analyse data, identify patterns, make predictions, adapt to new situations, and perform tasks like humans by learning from experience (Sarajcev et al., 2022). AI models usually have five common phases in their lifecycles: data preparation, model building, model training, model deployment, and model management. Figure 2 graphically depicts a typical AI model lifecycle (Sarajcev et al., 2022).

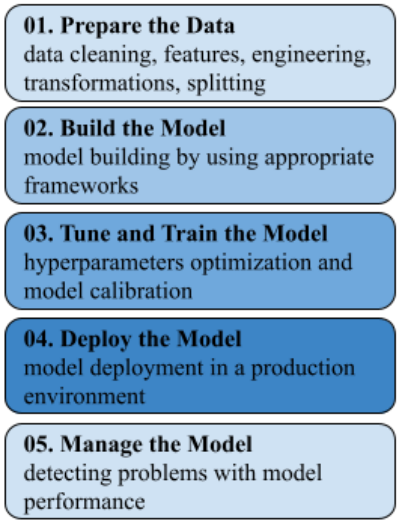

FIGURE 2 – AI model lifecycle

Source: Sarajcev et al. (2022)

Within this lifecycle, Berthelot et al. (2024) state that several phases focus primarily on data, namely: data acquisition, data production, learning, inference, and data storage, all of which are interconnected, especially across phases two to four. Furthermore, the authors distinguish the importance of data use in the training and inference phases of the models (steps 3 and 4). The authors explain that firstly, the most appropriate model is being identified, and it learns a suitable algorithm tailored to the task, then integrates the algorithm with the model and dataset, called the training phase. After that, when the model achieves the desired quality level, it becomes ready to be used with new data, named as inference phase (Berthelot et al., 2024).

2.4. Need for data, computational and energy resources in the AI model training and inference phases

In the lifecycle of AI models, data is an essential input since it serves as the foundation for model building and training, and significantly impacts their performance, accuracy, and complexity. As Röger et al. (2024) highlights, the quality and volume of data influences the accuracy of the AI models' predictions and the complexity of the models themselves, and insufficient data can lead to inaccurate or incorrect predictions.

This dependency on data extends to the parameters of AI models, which are designed to capture the patterns and relationships within the provided data. For instance, GPT-4 (1.7+ trillion parameters), Gemini Pro (540+ billion parameters), and Llama (70 billion

parameters) rely on billions to trillions of parameters to process and learn from data (Zvornicanin, 2024). Model parameters are the internal variables that the model uses to make predictions and decisions, and including weights, biases, and activation functions. The effectiveness of these parameters, however, hinges on the quality and comprehension of the input data, enabling the models to make more accurate predictions and decisions (TED AI, 2024). Thompson et al. (2020) state that models with more parameters tend to deliver higher accurate results and better performance. However, these authors indicate that achieving such performance levels requires more data and computational resources. On the other hand, Leuthe et al. (2024) argue that a higher volume of data does not necessarily lead to better model performance. In fact, they reveal that smaller, simpler models can sometimes perform better, emphasizing the importance of data quality over the data volume. Moreover, Al-Jarrah et al. (2015) claim that to handle large volumes of data, current AI models lack efficiency and scalability so that deploying more accurate systems is significant as it increases the use of preference.

AI models also require extensive calculations to properly align their parameters for optimal performance. These computations are carried out using various computational resources, which include both hardware and software components. The most used hardware components are CPUs (Central Processing Units), GPUs (Graphics Processing Units), and TPUs (Tensor Processing Units) (DataScientest, 2024). According to Strubell et al. (2019), advances in hardware technology for training models have significantly improved the result accuracy and high-demand computational models achieve better results due to these improvements. Achieving the desired level of accuracy, however, requires large-scale computational resources, which often involve substantial energy consumption (Strubell et al., 2019). Therefore, Cowls et al. (2023) highlights the differences between training and inference phases according to their computational needs. They explain that while training an ML model entails supplying the algorithm with a set of labelled data, which enables it to adjust its internal parameters and minimize errors, after training - in the inference phase - the model is finalized to generate predictions using new, unknown data, thus requiring different levels of computation in each phase. Luccioni et al. (2024) state that the inference phase requires significantly less computational power compared to the training phase, but that the former occurs much more frequently, often billions of times daily for popular services like Google Translate. Cho (2023) discusses

that in the inference phase, the energy consumption might be higher than the training phase and presents Google estimations that indicate that 60% of the energy consumed comes from the inference, while 40% belongs to the training. Moreover, Xu et al. (2021) state that the burden of inference is higher due to the increment in larger models and exemplify the finding that 90% of the infrastructure costs for ML production are attributed to inference.

2.5. Carbon footprint of AI models

A carbon footprint, expressed in terms of carbon dioxide equivalent (CO_2eq), measures the greenhouse gas emissions (GHG) of a device or activity (Cowls et al., 2023). These emissions accumulate in the atmosphere and oceans, contributing to climate change, a major threat to our planet. The Paris Agreement was made to combat climate change, to keep global temperature rise to below 2°C, ideally 1.5°C, above pre-industrial levels (UNFCC, 2024). According to data from the EU's Copernicus (2024) global warming has already breached the 1.5°C target in 2024. As climate change continues to negatively impact society, economically and socially, it is imperative to reduce GHG emissions to meet the targets set by the Paris Agreement.

The use of AI models contributes to GHG emissions by training using large amounts of data, inference, and being overall computationally intensive. As AI models become more complex, their carbon footprints also start to be considered in the fight against climate change and thus calculated. For instance, researchers at the University of Massachusetts Amherst discovered that a single AI model produces over 284 t of CO_2 around 6 months, equivalent to the emissions of five cars over their lifetime (Cho, 2023). Similarly, training GPT-3 with 175 billion parameters consumed 1287 MWh of electricity, and emitted 502 t of CO_2, which is comparable to the annual emissions of 112 gasoline-powered cars (Patterson et al., 2021). In addition, specific tasks performed by AI models also contribute to energy consumption. Researchers at Hugging Face and Carnegie Mellon University discovered that generating an image with an AI model consumes as much energy as fully charging a smartphone and generating text is less energy intensive; Producing text 1000 times consumes only 16% of the energy required to charge a smartphone (Heikkilä, 2023).

Measuring AI model carbon emissions is a crucial step in raising awareness and controlling its potential environmental impact. Several factors involved in the calculation highlighted by recent literature: (a) hardware; (b) data centre location and the energy grid; and (c) complexity, length of training, and running frequency of inference tasks.

2.5.1. Hardware

Hardware components such as GPUs, CPUs, and memory are very power-intensive and consume large amounts of electricity to make the complex computations needed for AI models training and inference. GPUs account for approximately 70% of power consumption, while CPUs contribute 15%, and RAM make up 10% (Bouza et al., 2023). This power consumption is influenced by various factors, including the size and complexity of the model, dataset size, and hardware infrastructure, and can range from hundreds to thousands of kilowatt-hours (kWh) (Schwartz et al., 2019). According to Li et al. (2023), as demand for high-performance computing (HPC) systems increases, GPUs were found to produce notably higher carbon emissions compared to CPUs, based on a comparison across 500 supercomputers. Additionally, the capacity of memory and storage devices also influences carbon emissions, similar to compute units (e.g. CPUs, GPUs). Lacoste et al. (2019) argue that the computing hardware choice is crucial and directly impacts carbon emissions, and states that CPUs are found to be up to 10 times less efficient than GPUs. Moreover, these authors add that while powerful GPUs are readily accessible to practitioners, each new type of GPU must be trained to address challenges that require large datasets and extended processing times.

2.5.2. Location of the data centre and the energy grid

Another major factor in the carbon emissions of AI models is data centres. Their emissions are driven by two key factors: (i) the power consumption of servers; and (ii) the carbon intensity of the energy grids that supply them (Cho, 2023). Most data centres still rely on fossil fuels and operate 24/7, contributing approximately 1% of global GHG emissions, equivalent to around 330 Mt CO_2eq in 2020 (Rozite et al., 2023).

Servers in data centres consume electricity and generate heat, making cooling systems critical to prevent overheating. Goldman Sachs (2024) states that cooling accounts for 40% of a data centre's electricity usage. The authors added that in 2021, data centres represented 0.9% to 1.3% of global electricity consumption, with projections indicating

a rise to 1.9% by 2030. Further estimates done by Goldman Sachs (2024) show that, as AI-related energy consumption increases, Europe's energy demand could grow by 50%, given that 15% of global data centres are located there and, in the U.S., data centres' national power consumption is expected to climb from 3% in 2022 to 8% by 2030.

Lacoste et al. (2019) discuss that the carbon emissions from data centres are heavily influenced by their location and the energy grid they rely on. These authors explain that data centres connected to electricity grids with a higher share of renewable energy will produce significantly fewer carbon emissions compared to those reliant on fossil fuels. Additionally, servers powered by renewable-heavy grids, like those in Quebec, Canada, emit as little as 20g CO_2eq/kWh, while servers in Iowa, USA, where fossil fuels dominate, emit up to 736.6g CO_2eq/kWh (Lacoste et al., 2019).

To measure the carbon footprint of a data centre, energy consumption across power supply, cooling, and maintenance must be evaluated (Bouza et al., 2023). This is done by using Power Usage Effectiveness (PUE), representing how the data centre uses energy efficiently, i.e., the lower the PUE, the more efficient the data centre is. For instance, Google Cloud Services with a PUE of 1.1 indicates that for every 1 unit of energy used by the servers themselves, 1.1 units are consumed by the entire data centre (including infrastructure like cooling and lighting) (Lacoste et al., 2019).

2.5.3. Complexity, length of training and running frequency of inference tasks

Other significant factors are related to the model complexity, the length of training, and the running frequency of inference tasks. The complexity is correlated with how large the model is, while the length of training refers to the time it takes to train the model. Running frequency is about how frequently the model is being used to run the inference phase.

During training, for instance, Large Language Models (LLMs) such as GPT-4 constantly require new data generation, resulting in higher demand for computational resources and significant energy use (Castro, 2024). AI Researchers at the University of Massachusetts Amherst found that BERT, Google's LLM, produced 0.6524 t of CO_2 during 79 hours of training. In addition to the training phase, Amazon Web Services indicates that 90% of the cost of AI models comes from the inference phase, while Schneider Electric estimates this at 80%. Wu et al. (2021) also found that the inference

phase in LLMs accounts for 65% of the emissions. Furthermore, Luccioni et al. (2024) state that it is challenging to find an exact equilibrium between the energy costs of the training and inference phases due to the energy requirements of each phase in the AI model's lifecycle. However, the authors exemplifies that the energy costs of deploying ChatGPT would surpass its training costs in a few weeks or months of the model use.

2.6. Carbon footprint and emission calculators

AI emission calculators are tools that facilitate the measurement of energy use and the carbon emissions of AI models. Patterson et al. (2021) simplify the calculation of the carbon footprint of an ML algorithm using the formula below in Equation (1), which can be generalized to calculate the CO_2 emissions for AI models as a base for the tools.

(1) $$(CarbonFootprint = \left(\text{electrical energy}_{\text{train}} + \text{queries} \times \text{electrical energy}_{\text{inference}}\right) \times \frac{\text{CO}_{2\text{e}_{\text{datacenter}}}}{\text{kWh}}$$

Equation (covers the amount of electrical energy consumed during the training phase of the model (electrical energy $_{\text{train}}$), the total energy consumed during the inference phase by multiplying the queries (how many times the model is used) by the electrical energy consumed (electrical energy $_{\text{inference}}$), and the carbon emissions associated with the electricity used by the data centre hosting the model ($CO_{2\text{datacenter}}$) which depends on the energy mix of the data centre (e.g. renewable, gas, coal) (Patterson et al., 2021).

Bannour et al. (2021) compare the use of six publicly available AI emission calculators: Carbon Tracker (CT); Green Algorithms (GA); Experiment Impact Tracker (EIT); ML CO_2 Impact (MLCI); Energy Use (EU) and Cumulator (Cu) (see Table 1). These tools are widely used to calculate CO_2 emissions from NLP experiments, assessing emissions based solely on energy consumption during the dynamic use, with the production and end of life phases not being considered, or being only partly considered. Furthermore, they observe that these tools are good at providing large emission results rather than pointing out the smaller contributions.

Despite the current usage, the authors state that more research is required to understand the differences between these tools, the way they reach results (and how

accurate these are), and comprehend the sources of AI model emissions. The metrics used in each of the six AI model emission calculator tools are described in Table 1.

TABLE 1 - EMISSION CALCULATOR TOOLS

Tool Name [a]	Metrics	Power usage effectiveness (PUE)
Carbon Tracker (Anthony et al., 2020)	Hardware used, energy consumption, and the carbon intensity of the electricity grid	Default PUE = 1.67 (2019)
Green Algorithms (Lannelongue et al., 2021)	Runtime, number of cores, memory requested, type of platform used (PC, local server, cloud computing), type of cores, location	Default PUE = 1.67 (2019)
Experiment Impact Tracker (Henderson et al., 2020)	The number of processors, energy consumed during computation, and hardware efficiency	Default PUE (1.58) (adjustable)
ML CO_2 Impact (Lacoste et al., 2019)	Hardware, runtime, cloud provider and location of the computing facilities operated	Partly PUE used for few cloud providers
Energy Usage (Lottick et al., 2019)	Energy mix data (coal, oil, natural gas, and low carbon fuels) estimates CO_2 emissions based on grid composition	No PUE used
Cumulator (Trébaol, 2020)	Runtime, GPU load and carbon intensity, with a fixed value for consumption of a typical GPU	No PUE used

[a] These tools estimate emissions by analysing key factors such as hardware usage, energy consumption, runtime, and carbon intensity of electricity sources. The "Power Usage Effectiveness (PUE) Value", a metric, measures how efficiently energy is used in data centres, with lower values indicating better energy efficiency. Some tools assume default PUE values, while others allow customization or not account for it. Source: Bannour et al. (2021)

To exemplify the differences between calculators, Bannour et al. (2021) look at NER (Named Entity Recognition) models, which are used to identify and classify named entities in text (such as names of people, locations, and organizations) and categorised them into predefined categories. The carbon emissions associated with three NER datasets, calculated by each of the six AI model emission calculator tools mentioned above are presented in Table 2.

TABLE 2 - NER (NAMED ENTITY RECOGNITION) MODELS CARBON EMISSION'S CALCULATION BY DIFFERENT TOOLS

CO_2eq (g.) [a]							Runtime (minutes)
NER	**Carbon Tracker**	**Experiment Impact Tracker**	**Energy Usage**	**Cumulator**	**ML CO_2 Impact**	**Green Algorithms**	
French Press							
Server	237.96	78	0.49	302	290	350.15	163:39
Facility	161.16	48	0.98	222	250	260.26	118:04
EMEA							
Server	9.7	30	0.00131	19	20	16.67	9:31
Facility	8.07	1	0.002	13.7	10	14.31	6:51
MEDLINE							
Server	13.44	30	0.00128	26.1	20	20.68	11:55
Facility	10.5	1	0.0026	19.4	20	20.03	9:11

[a] Emissions are reported in CO_2 eq (g); French Press, EMEA, and MEDLINE datasets represent different linguistic or subject-matter challenges for NER models. Server covers the direct energy and emissions from computation. Facility includes the emissions from supporting infrastructure that enables those computations. Source: Bannour et al. (2021)

Bannour et al. (2021) demonstrate that while ML CO_2 Impact and Green Algorithms present higher emissions compared to other tools, Carbon Tracker and Experiment Impact Tracker provide similar outcomes and Energy Usage tools give out the lowest emission calculation because they do not include hardware consumption. Results by Bannour et al. (2021) illustrate, in AI model footprint, local servers are responsible for higher emissions and energy use than the computing facilities.

2.7. Systemic calculation and reporting challenges

Currently available literature indicates that the systemic calculation and reporting of AI models' carbon emissions is still challenging, mostly because of lack of data on lifecycle emission, energy consumption, energy sources, and the complexity of the AI systems. Delanoë et al. (2023) state that hardware power, cloud location, and training time are not always relevant information for measurement because many research papers fail to provide detailed specifications about the AI models used, making precise energy consumption evaluation difficult and leading assumptions that impact the accuracy of

carbon emission estimation. Thus, the authors drew attention on the lack of standardized reporting in carbon emission calculation, while Luccioni et al. (2024) claim that documentation is based on the dynamic use of power consumption, primarily during training, due to the ease of quantifying energy use in that phase. However, Andrews (2020) mentions that each training session is complex, as it draws power from different functions and must be untangled from other phases to be properly calculated. Despite the training phase being the most tractable part of the lifecycle, recent advancements have led AI/ML researchers to give increasing importance to the inference phase (Cho, 2023).

Bannour et al. (2021) indicate that existing emission calculator tools provide different outcomes for the same AI model, and Lacoste et al. (2019) support this by stating that these tools only represent approximations of the true emissions due to the lack of precise energy consumption and carbon production data reported by organizations. Thompson et al. (2020) assessed 1,058 research papers on DL and concluded that most of them do not mention the computational requirements of AI models. These reported challenges contribute to the lack of standardization in systematically comparing and quantifying the carbon footprints of different AI models.

2.8. Carbon footprint reduction in AI models and Sustainable AI

Electrical energy consumption and resulting carbon footprint of AI models has led the AI community to adhere to a more sustainability mindset. Sustainable AI is defined by van Wynsberghe (2021) as the development and use of AI models that are environmentally accountable for our present and future societies. In addition to this movement, Green AI has also been gaining attention and is often used interchangeably in literature. According to Schwartz et al. (2019), Green AI seeks to achieve results by keeping the computational cost lower, while Alzoubi and Mishra (2024) define it as the reduction of carbon emissions and power consumption within the application of environmentally friendly AI systems. Leuthe et al. (2024) adds to this definition, including the promotion of sustainable energy sources uses, the focus on the sustainable design and use of AI models. To implement sustainable AI, Gaur et al. (2023) proposes adding a set of carbon emission reduction steps to the current phases, as presented below (Figure 3).

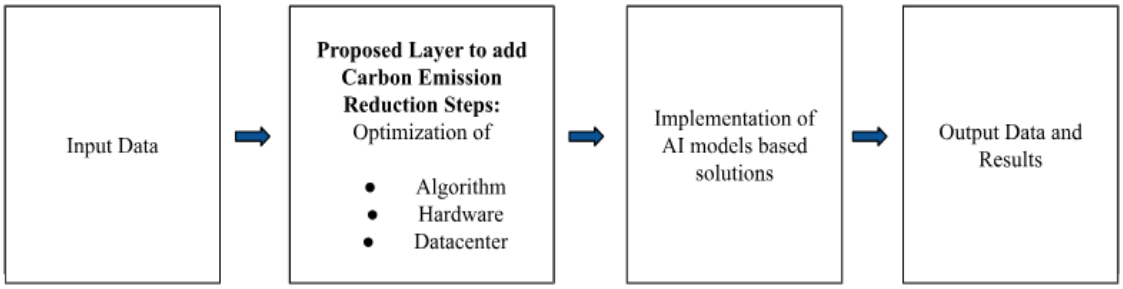

FIGURE 3 – Sustainable AI implementation

Source: Gaur et al. (2023)

2.8.1. Algorithm optimization

To address the high computational demands of AI models, Thompson et al. (2020) recommend optimizing energy usage and environmental impact by prioritizing computationally efficient algorithms from the outset. They further argue that developing green algorithms reduces computational requirements through optimization techniques, ultimately lowering energy consumption. For instance, Wu et al. (2021) highlight the example of NVIDIA DeepStream SDK, which is designed to build efficient, scalable AI-based video analytics applications, supporting multiple algorithms, and Hugging Face Transformers' Efficient Inference Mode, which reduces computational costs and latency during the inference phase of NLP models. Bolón-Canedo et al. (2024) also suggest that restricting the number of times a computationally expensive algorithm runs would be the easiest way to minimize energy use.

Alzoubi and Mishra (2024) emphasize that improving model efficiency lowers energy consumption and aids in selecting more suitable models. For instance, DeepSeek's[1] optimized distillation techniques, which transfer reasoning capabilities from larger models to smaller ones, resulting in reduced energy consumption without compromising performance (Hanbury et al., 2025). Similarly, Patterson et al. (2021) note that energy-efficient model selection enhances output quality while significantly reducing computational costs and provide the findings that using sparse (models with many zero weights, requiring fewer computations and less memory usage) instead of dense ones (models where most or all parameters are non-zero, requiring more computations and memory usage) can decrease computational requirements by 5 to 10 times. Gaur et al.

[1] DeepSeek: https://www.deepseek.com/

(2023) further argue that models with fewer parameters enhance computational efficiency, making them a practical choice for sustainable AI development.

2.8.2. Hardware optimization

Liu and Yin (2024) suggest that using faster GPUs can lower emissions and help organizations become more environmentally conscious, regardless of the model being used. Some GPUs are more efficient, and selecting efficient hardware contributes to energy savings (Bolón-Canedo et al., 2024). Thompson et al. (2020) further discuss that more efficient hardware not only scales up computational capabilities but also enhances computational efficiency. These authors also state that hardware specialization, particularly with GPUs, TPUs, and other specialized chips has led to significant computational gains. Patterson (2022) also advocates that specialized processors are much more energy-efficient than general-purpose processors, achieving 2 to 5 times better performance in terms of energy consumption. For instance, TPUs improved by 1.5 times in compute per dollar and 4.9 times in compute per watt between 2017 and 2020 (Thompson et al., 2020).

2.8.3. Data centre optimization

Lacoste et al. (2019) indicate that the selection of a lower carbon-based data centre location, a data centre in a region where electricity generation relies more on renewable energy sources such as wind, solar, hydro, or nuclear power, significantly impacts its carbon footprint. In support of this, Henderson et al. (2020) argue that a rapid reduction in carbon emissions could be achieved by implementing each training phase with carbon-efficient energy grids connected to data centres. The authors demonstrate that deploying tasks in Quebec, known for its clean energy sources, leads to a 30-fold reduction in emissions compared to Estonia. However, Lacoste et al. (2019) also note that information about the CO_2 emissions from servers connected to energy grids is often unavailable. This lack of data results in assumptions that servers in the same physical area leads to ignoring possible variations in CO_2eq emissions across different grid locations.

Moreover, Wu et al. (2021) state that selecting carbon-neutral data centres is challenging due to the long-term financial investments required and restrictions on geography and materials (e.g. rare metals). Adding to this, Bouza et al. (2023) highlight that running AI models in carbon-friendly regions can be influenced by factors like the

time of execution during the day or the allocation of energy sources at specific moments. To manage such constraints, the authors suggest analysing electricity maps of the countries before running tasks there. Additionally, the authors draw attention to the fact that transferring large datasets to carbon-free locations can sometimes have a higher environmental cost. Therefore, decisions on data centre location should be balanced with benchmarks comparing staying in the same server.

Nevertheless, most studies suggest that efficient data centres lead to 1.4 to 2 times better energy use and, when renewable energy is chosen, CO_2 emissions can drop by a factor of 5 to 10 (Patterson, 2022). Lacoste et al. (2019) supports these results by noting that data centres powered entirely by renewable energy contribute only with 20g CO_2eq/kWh, while fossil fuel-based centres can emit as much as 820g CO_2eq/kWh. The authors further emphasize that a single decision regarding location can drastically reduce emissions, with potential reductions of up to 40 times. As an example, Evans and Gao (2016) demonstrate that Google's investment in greener data centres through the selection of renewable energy sources and innovations in cooling systems, resulted in 3.5 times increase in computing power with the same energy consumption.

2.9. Stakeholder engagement for Sustainable AI

Wu et al. (2021) emphasize that achieving sustainability of AI models requires accountability from multiple stakeholders to support the development process. Similarly, van Wynsberghe (2021) stresses that all stakeholders must collaborate throughout the entire AI model lifecycle to ensure sustainable practices. Gaur et al. (2023) propose an approach that underscores the importance of stakeholder collaboration and coordination to enhance model efficiency and sustainability.

Verdecchia et al. (2023) examine studies that focused on the carbon emissions associated with AI models, emphasizing the responsibility of researchers and practitioners to design and use models sustainably. They also highlight the role of policymakers in ensuring government accountability for integrating sustainability into AI use. Delanoë et al. (2023) further underscore the critical role of policymakers in optimizing efforts to reduce CO_2 emissions. They advocate engaging a broad audience, including scholars, data scientists, developers, managers, and institutions across various sectors relying on AI models.

Tornede et al. (2023) states that Green AI has become a community-driven initiative, facilitating collaboration among researchers and developers. Similarly, Rolnick et al. (2019) advocate for engaging domain experts to simplify complex tasks and develop effective strategies. They point out that making ML models accessible through a common language or platform, along with ensuring interpretability, allows solutions to reach the right audience and enables stakeholders outside the ML community to understand and apply these models in real-world scenarios effectively.

While efforts to promote the sustainability of AI have increased in both literature and practice, challenges remain, and stakeholder engagement is critical. For instance, Verdecchia et al. (2023) highlight a lack of actionable insights and holistic design methodologies for sustainable AI. Similarly, Leuthe et al. (2024) examine how model developers can enhance design processes sustainably, advocating for making sustainable design information more accessible and applicable. Patterson (2022) recommends that practitioners select effective hardware in data centres powered by renewable resources for AI model training and deployment. Researchers are also encouraged to develop more efficient models, such as by integrating smaller models or reusing existing ones. Transparent reporting of energy usage and carbon footprints should also become standard practice, rather than focusing solely on model quality and accuracy (Patterson, 2022).

3. METHODOLOGY

3.1. Research design

This thesis focused on the carbon emission estimation and reduction emitted by AI models and a qualitative approach has been adopted based on literature, case studies and expert consultation. Given the complexity and recent emergence of the problem, this approach enabled the exploration of existing literature, recent studies and the analysis of relevant issues, focusing on the associated challenges and innovative solutions.

3.1.1. Literature review

As a first step, a systematic literature review was conducted on AI, ML and associated topics, and their connection to sustainability. Peer-review publications were searched and downloaded from Web of Science, ScienceDirect, and Google Scholar. This initial research revealed that most studies currently focus on how AI can reduce carbon emissions in various sectors, rather than addressing AI models' own carbon emissions. It also revealed the limited availability of peer-reviewed papers on this topic. Out of 97 papers found, only 23 of them were academically peer reviewed. Therefore, the research also relied on grey literature. Table 3 outlines the main categories and associated keywords used during the literature search.

TABLE 3 - KEYWORDS SEARCHED FOR DURING LITERATURE REVIEW

Category	Keywords [a]
General Themes	Artificial Intelligence, Generative AI, Machine Learning, Natural Language Processing, Deep Learning, Large Language Models, Sustainable AI, Green AI
Environmental Impact	Carbon Footprint, Carbon Emissions, Energy Consumption, Environmental Sustainability, Greenhouse Gas Emissions, Carbon Neutrality
Tools, Techniques, and Measurement	Carbon Emissions Tracking, Life Cycle Assessment, CodeCarbon, Experiment Impact Tracker, CarbonTracker, Cumulator, Green Algorithms, Climate Change AI
AI Applications and Sectors	Data Centres, Cloud Computing, High-Performance Computing
Innovations and Solutions	Green Algorithms, Energy-Efficient AI, Renewable Energy Integration, Carbon-Aware Computing, AI Model Optimization
Challenges and Opportunities	Energy Demand of AI, Trade-offs in Model Efficiency and Accuracy, Environmental Costs of AI Training, Policy Implications for AI Sustainability

[a] Each category's keywords were combined to increase the finding of relevant papers.

The results of this literature review are presented in section 2 above, and were organised into key areas of relevance, including: introduction to AI and AI model lifecycle; needed resources and its implications to AI carbon footprint; emission calculators; reporting challenges; sustainable AI; and stakeholder engagement.

3.1.2. Case studies

As a second step, a case study approach was applied to examine real-world applications of the carbon footprint calculation and reduction in AI models. This approach enabled the evaluation of current challenges in calculating and reporting AI model carbon emissions from AI models, as well as the exploration of solutions and frameworks being developed to reduce their emissions.

Case study selection criteria was based on the availability of peer-review and grey literature about specific models and organizations, and the authors own knowledge and working experience with the development, training and use of AI models in the real-world.

Case Study 1: BLOOM and its carbon emission calculation: This case study demonstrates how the emissions of an AI model are calculated, highlighting the factors considered, challenges in accurate energy consumption estimation, the importance of model lifecycle, and the tools and methods used for emission estimation, while addressing relevant reporting challenges.

Case Study 2: Microsoft's data centres contribution to carbon emissions: This case study highlights the importance of data centres in the carbon footprint, emphasizing the role of location and the energy required to operate the data centre, as well as the precautions taken to mitigate its negative impacts.

Case Study 3: Zeus: energy-efficient models and optimization: This case study discusses the optimal balance between energy consumption and training speed. It leads to understanding the significant role of optimization techniques in reducing energy consumption and carbon emissions of the models in achieving sustainable AI practices.

Case Study 4: NVIDIA hardware acceleration: This case study examines the role of hardware acceleration in reducing the carbon footprint of AI models by exploring the advancements in hardware technologies to make the computation of the models efficient and the challenges associated with its cost and adoption in the field.

3.1.3. Expert consultation

Expert consultations were conducted with AI/ML researchers who are directly involved in the organisations or AI models mentioned in the case studies. Following an open-ended interview protocol (see Annex A), the experts were asked about the research questions of this thesis and encouraged to share their relevant experiences. All interviews were done online, by sending the questionnaire and gathering written statements.

3.2. Research tools and procedures

In this thesis, MAXQDA[2] was utilized to analyse and organize scientific papers and grey literature. The software allowed for systematic coding of the papers, organizing relevant information under unified themes for easier analysis and citation. To find further relevant papers and speed up literature review, Elicit[3] and Scispace[4] tools were also used. Additionally, ChatGPT-4[5] and DeepL[6] were employed to enhance the thesis-writing process by improving flow, checking spelling, and providing paraphrasing support.

3.3. Study limitations

Several limitations to this study must be acknowledged. The lack of peer-reviewed academic papers on the AI models' contribution to carbon emissions lead to an extended reliance on grey literature, which might lack formal academic evaluation. Time constraints and lack of responses also restricted the in-depth analysis of case studies, limiting the number and range of stakeholders involved.

3.4. Ethical considerations

This study ensures transparency and proper attribution of all data sources, adhering to ethical research practices. All literature and case studies were publicly available, and no sensitive or personal data was used. Consulted experts gave written consent for the material gathered during the interviews. All personal information was anonymized, keeping the connections between experts and case studies undisclosed to prevent inadvertent identification.

---

[2] MAXQDA: https://www.maxqda.com/
[3] Elicit: https://elicit.com/welcome
[4] Scispace: https://typeset.io/
[5] ChatGPT: https://chatgpt.com/
[6] DeepL: https://www.deepl.com/en/write

4. RESULTS

Results from the literature review analysis are presented in section 2. That analysis provided the basis for the development of four in-depth case studies and related expert consultations, where the research questions were further explored. This section presents the findings from the case studies and expert consultations, which experts are chosen from the organizations; BLOOM, Zeus, and Microsoft mentioned in the case studies.

4.1. Case study 1: BLOOM and its carbon emission calculation

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model), is a 176-billion-parameter language model, created as part of the *BigScience Workshop (2022)* and trained on 1.6 terabytes of data in 46 natural and 13 programming languages. It was a collaborative initiative started in July 2022, involving over 1,000 researchers from more than 60 countries. BLOOM's carbon emissions were assessed using Life Cycle Assessment (LCA) approach, however, due to the limited data for each phase, carbon emissions are considered in equipment manufacturing for model training, and model deployment phases (Luccioni et al., 2022). Luccioni et al. (2022) state that the carbon emissions of training BLOOM are mostly connected to three main resources: embodied emissions, dynamic power consumption, and idle power consumption. The authors observe that the training phase required 1.08 million GPU hours, and that embodied emissions, i.e., linked to the production of computing equipment like servers and GPUs, contributed 11.2 t CO_2eq. Dynamic power consumption, the energy used during active model training, accounted for 24.69 t CO_2eq. Idle power consumption, the energy used by infrastructure when not actively training, added 14.6 t CO2eq. In total, training emitted 50.5 t CO_2eq, with dynamic consumption contributing the largest share (48.9%), followed by idle (28.9%) and embodied emissions (22.2%) (Luccioni et al., 2022). In the deployment phase's CO_2 calculation, CodeCarbon tool monitored a Google Cloud Platform (GCP) instance with 16 NVIDIA A100 GPUs over 18 days, processing 230,768 real-time requests. The instance consumed 914 kWh of electricity, with GPUs accounting for 75.3%, RAM for 22.7%, and CPUs for 2%. Even during minimal activity, 0.28 kWh was consumed every 10 minutes, highlighting the energy required to maintain BLOOM in memory. Operating in the US-central region, with a carbon intensity of 394 g CO_2eq/kWh, the model deployment emitted 340 kg CO_2eq over 18 days, averaging 19

kg CO_2eq daily. Luccioni et al. (2022) state that the model inference phase remains understudied compared to training. The authors reveal critical challenges on the study requiring further attention; Insufficient lifecycle information, including embodied emissions from GPU manufacturing, results in imprecise estimates, necessitating the greater transparency; Additional research is needed to address model inference complexities, particularly in real-time scaling and maintenance, focusing on bridging gaps between chip designers and users while optimizing energy consumption and emissions; Transparent reporting on carbon emissions, detailing energy use, carbon intensity, and research and development contributions to enable meaningful comparisons and understanding BLOOM's environmental impact.

4.2. Case study 2: Microsoft's data centres contribution to carbon emissions

Microsoft, a large US based software company, has committed to utilizing the Three Mile Island nuclear power plant in Pennsylvania to meet its AI energy needs (Luscombe, 2024). As AI is projected to drive a 160% increase in data centre energy demand by 2030 (Goldman Sachs, 2024), Microsoft aims to leverage nuclear power, to ensure that its data centres operate with zero emissions. According to Microsoft (n.d.), 2024 Environmental Sustainability Report, indirect emissions associated with the company's value chain come mostly from the construction of data centres. Associated carbon emissions, embodied in building materials, as well as in hardware components such as semiconductors and servers, have risen to 30.9% since 2020. The report shows that the company is being challenged due to the growing demand for its cloud supply, in turn leading to the expansion of its data centres. To address these challenges, Microsoft underscores the critical role of carbon-free electricity and the potential of advanced nuclear and fusion energy in achieving a decarbonized energy future. To lower emissions, Microsoft takes the initiative of monitoring the energy consumption of its data centres. By striving for a Power Usage Effectiveness (PUE) ratio close to 1.0 – it currently achieves 1.12 – Microsoft aims to enhance data centre efficiency by designing and building sustainable data centres. Additionally, transitioning servers to a low-power state has enabled Microsoft to reduce energy usage by up to 25% on unallocated servers, saving thousands of megawatt-hours monthly across its global data centres. The company also prioritizes optimizing cooling efficiency, utilizing predictive models to anticipate water consumption based on real-time weather data in water-cooled facilities. These models are designed to eliminate water

usage for cooling, reducing reliance on freshwater resources as the AI computation demand continues to grow.

4.3. Case study 3: Zeus: energy-efficient models and optimization

Researchers at the University of Michigan developed an online optimization framework called Zeus for reducing carbon emissions in DL model training. They state that existing practices primarily target optimizing DL training for faster completion over energy and carbon efficiency, resulting in inefficient energy usage. Zeus seeks to provide a solution to the current circumstances by automatically finding the optimal balance between energy consumption and training speed during the process (You et al., 2022). Zeus configures the GPU power limit and the model's batch size parameter (i.e. various workloads) to minimize power consumption and carbon emissions. The Zeus process starts with users submitting a request for a feasible batch size and power limit for their DNN training. Then, Zeus finds an optimal batch size and power limit configuration and launches the training. During and after training, Zeus collects statistics on DNN training and GPU power consumption. This feedback is used to update Zeus' internal states. The training task is ended when the target metric is reached, or the stopping threshold is exceeded. In this automated feedback loop system, Zeus continuously learns and adjusts its settings to optimize energy-time costs. By using this framework, and setting a GPU power limit, Zeus reduces its usage and slows down model training until adjustments are made. After analysing a wide range of GPUs, the researchers revealed that drawing maximum power does not always yield the best performance, and doing so leads to diminishing returns in terms of efficiency. The optimal energy consumption can be achieved at a lower power limit, reducing energy consumption by 3.0%–31.5%. (You et al., 2022) On the other hand, You et al. (2022) observed that large batch sizes reduce training time by improving the data processing speed, leading to more energy consumption for the same target accuracy. Batch size controls how many training samples the model processes before updating its understanding of the data. Analysis of several valid batch sizes (ranging from 8 to the maximum batch size supported by GPU memory) for six DL tasks, such as NLP, speech recognition shows that the energy-optimal batch size can lower energy consumption by 3.4%–65.0% compared to the default batch size for the same target accuracy. As a result, optimizing the GPU power limit and the right batch size configurations can achieve energy savings of 23.8%–74.7% for diverse

workloads. Zeus reduces energy consumption by 15.3%–75.8% and training time by 60.6% by simply selecting the maximum batch size and maximum GPU power limit. (You et al., 2022) However, some challenges remain. Identifying and navigating the trade-off between energy consumption and training time is complex, especially considering the non-linear relationships between these factors. Additionally, different models and GPUs have unique energy characteristics, which makes generalizing offline profiling results difficult. Lastly, the vast number of possible configurations, each demanding hours to days of evaluation, adds significant complexity to optimization.

4.4. Case study 4: NVIDIA hardware acceleration

NVIDIA is a US technological corporation that designs and supplies GPUs, considered key to drive advancements in the fields of AI and computing. NVIDIA develops hardware, software, and networking technology to enhance performance, energy efficiency and emissions reduction (Nvidia, 2024). To handle the AI models' demanding workloads effectively, modern data centres increasingly depend on accelerated computing, which significantly enhances computation by using parallel processing to handle frequently occurring tasks. Unlike traditional processors that execute tasks serially, accelerated computing offloads demanding workloads, offering lower costs, higher performance, and greater energy efficiency. By completing larger workloads more quickly, this approach reduces energy consumption and enables systems to return to low-power idle states faster than traditional computing. NVIDIA offers special GPUs for accelerating workloads to operate in parallel, which significantly enhances throughput while reducing the overall energy required to complete tasks. This results in considerable energy savings and a better total cost of ownership. Accelerated computing has revolutionized workloads that previously demanded tens of thousands of general-purpose servers, which consumed 10 to 20 times more energy, into highly efficient processes. While each GPU server may entail higher costs and greater power consumption, the need for fewer servers leads to substantial savings in both energy and expenses. For example, NVIDIA's CUDA GPUs accelerate Apache Spark, reducing its carbon footprint of data processing by up to 80% while achieving five times faster speeds and cutting computing costs by four times. Similarly, NVIDIA's Grace Blackwell Superchips provides 25 times better energy efficiency for LLMs. Additionally, NVIDIA's Blackwell GPUs demonstrate 20 times more energy efficiency than traditional CPUs for specific AI workloads.

Furthermore, NVIDIA's Data Processing Units (DPUs) can achieve a 25% reduction in power consumption by offloading critical data centre tasks from less efficient CPUs. By shifting from a CPU-centric infrastructure to GPU and DPU acceleration could save 30 trillion watt-hours of energy each year, comparable to the electricity consumption of nearly 4 million U.S. households. NVIDIA's accelerated computing reduces the cost and energy required to training AI models but additional investments in innovation are necessary to discover better scientific, engineering, and operational solutions that conserve energy, time, and costs.

4.5. Expert consultation

Ninety-two AI/ML experts were contacted through LinkedIn or email and invited to provide their views via interviews. Eleven experts indicated their interest in participating in the study. However, only three actively shared their insights and provided actionable responses: one computational linguist, one researcher and one computer scientist.

4.5.1. Insights on carbon emission calculation

The experts point out that the calculation process is challenging due to AI model complexities. The computer scientist states that accurate measuring requires complete access to the end-to-end process, which AI model developers do not have, and this necessitates many stakeholders in the supply chain (power generators, power distributors, data centre operators, cloud service providers, and AI companies) to collaborate. Also, the researcher and the scientist claim a lack of knowledge on the accuracy of carbon emission results due to limitations coming from the CO_2eq emissions used for each power source in a region's power mix, estimated power or energy consumption of computations, and the embodied carbon footprint of computing hardware obtained by extrapolating scarce existing data, resulting in as an estimate. They also indicate that any estimation must be accompanied by a discussion of error bounds or uncertainty for better accuracy results. The researcher claims that these error bounds are hardly discussed in the context of carbon emission, making the estimates unscientific and not practical. Further, he states that few people scientifically discuss this issue, and many endorse back-of-the-envelope guesses of carbon footprint without error quantification, which is the fundamental problem. Therefore, the researcher recommends that a deliberate quantification of the

potential errors introduced due to the lack of data availability must be done simultaneously, and shedding light on this problem is the most important challenge.

On the other hand, the computer scientist states that the lack of verifiable data propagation throughout the supply chain and necessary policy changes to incentivize stakeholders remain significant issues. Experts agree on the need for more publicly available efficient infrastructure and public education on energy consumption (e.g., facilitating the average person understand energy measures and comparisons). The computational linguist adds that carbon consumption metrics could include variables such as hardware replacement (e.g., how often GPUs need to be replaced) and infrastructure consumption metrics (e.g., energy used for cooling servers). Also, she emphasizes that adopting a standard for reporting energy usage in academic training experiments for technical fields may mitigate calculation challenges.

4.5.2. Insights on carbon emission reduction

Two experts state that minimizing energy consumption directly leads to reductions in operational carbon emissions and is the most reliable way of avoiding carbon accounting discrepancies. Further, they indicate that the other ways of minimizing embodied carbon emissions (e.g., using fewer computing devices) and operational carbon emissions (e.g., using renewable energy only) must be considered.

The computer scientist who builds large-scale distributed software systems to reduce AI's energy consumption underlines the importance of optimization, which requires careful measurement and understanding of the end metric. The researcher who created energy optimization methods for ML workloads states that the most challenging part is incentivization. Further, he claims that the time-based pricing model used by major cloud vendors such as Google, and Amazon do not encourage users to optimize their energy use in computing. Instead, users are incentivized to maximize performance during the time they are paying for, even if this leads to excessive or inefficient energy consumption, resulting in only users who pay for their electricity are motivated to optimize energy usage.

The experts suggest focusing on power, energy, and other essential metrics instead of composite metrics like carbon emissions, which might remove many workarounds stakeholders use today to avoid solving the problem. Also, they stress the need for policy solutions requiring stakeholders to use the proper tools and disclose raw metrics in detail.

5. DISCUSSION

5.1. Transparency and standardized reporting

Our results reveal critical challenges and opportunities in accurately measuring and mitigating the carbon emissions of AI models, primarily in the training and inference phases. The first issue highlighted is the lack of transparency regarding the entire AI model lifecycle, particularly concerning embodied emissions from hardware manufacturing and operational energy use. This identified gap aligns with Strubell et al. (2019), who emphasized the importance of comprehensive lifecycle assessments to provide more accurate AI model carbon emission calculation figures. Transparent reporting and standardized metrics, as advocated by the stakeholders interviewed, are consistent with calls from the broader academic discourse for greater accountability and consistency in assessing the environmental impact of AI systems. As Patterson (2022) argues, transparent reporting of energy usage and carbon footprints should become standard practice, moving beyond the current focus on model quality and accuracy. This shift would enable more meaningful comparisons across models and foster a deeper understanding of their environmental implications, thus systematically supporting efforts to reduce AI's carbon footprint.

5.2. Use of verifiable data and metrics for CO_2 calculation estimations

Our results underline the complexities of accurately calculating carbon emissions due to the absence of verifiable data across the supply chain. Similarly, the existing literature advocates the challenges for tracing energy consumption and carbon emissions through the supply chain. For example, Bannour et al. (2021) showcase the several outcomes from different emission calculators using variable metrics while Delanoë et al. (2023) state that the metrics used are not always relevant to the calculation. As a result, AI model carbon footprints are merely estimations. Consequently, experts consulted emphasized that quantifying error bounds and uncertainties is critical for improving the credibility of carbon footprint estimations.

5.3. Mitigation approaches to carbon emission reduction

Energy efficiency and optimization strategies surfaced as key themes, mainly through the lens of accelerated computing and improved hardware designs, such as NVIDIA's advancements in GPU technology. These findings support the conclusions of Wu et al.

(2021), who noted that advancements in hardware efficiency could significantly reduce the environmental footprint of AI models. Furthermore, Microsoft's efforts to lower Power Usage Effectiveness (PUE) and optimize cooling systems resonate with studies advocating greener data centre designs (Bouza et al., 2023).

5.4. Stakeholder involvement and engagement

The experts consulted provided insights into the adverse effects of time-based pricing models on energy efficiency provide a novel contribution, emphasizing the necessity for policy-level interventions and pricing reforms to promote sustainable computing practices. This aligns with Verdecchia et al. (2023), who advocate for holding multiple stakeholders accountable for the sustainable design and use of AI while emphasizing policymakers' role in integrating sustainability into AI governance. Furthermore, fostering public knowledge about the carbon emissions of AI models is essential to drive informed decision-making and collective action across the AI ecosystem.

6. CONCLUSION

This study aimed to support a better understanding of the challenges and opportunities related to the calculation and reduction of AI models' carbon emissions, through the analysis of scientific and grey literature, case studies, and expert surveys.

The first research question "How can the carbon footprint of AI models be measured?" and associated sub-question of "What are the systemic reporting and calculation challenges associated with the carbon emission impact of AI models?" sought to evaluate how these calculations are currently done, what metrics are used (e.g., hardware, data centre location, energy grid and energy types), how the complexity and length of training and inference phases is being factored in, and what is the availability of standardised emission calculators. It is observed that current studies focus mainly on using AI models to assess carbon emissions and other sustainability challenges, rather than on calculating (or advancing calculation protocols of) the carbon emissions emitted by the models themselves. Therefore, the sustainability of AI requires further dedicated scientific inquiry. For example, our study shows that it is not yet possible to accurately measure the carbon emissions of AI models due to the poor understanding of their lifecycle complexities, and limitations in data availability (e.g., the hardware that is used and the data centre infrastructures they rely on). Hence, the values currently available can be considered as initial, broadly informed estimates. The case study findings from BLOOM regarding the estimation of an open-source model and Microsoft's carbon emission increment due to the data centre relying on non-renewable resources drew attention to the need for standardized AI model carbon emission reporting among tech companies and the challenges of developing coherent metrics for its calculation.

Having elaborated on understanding the measurement of AI model carbon emissions, the thesis aimed to identify potential ways to mitigate them. The second research question: "How can the carbon footprint of AI models be reduced?" and the associated sub-question, "How involved and aware are stakeholders in the effort to mitigate carbon emissions of AI models?" aimed at analysing current proposed ways forward in sustainable AI, and asses the currently level of stakeholder engagement and awareness in carbon footprint mitigation. Our findings show that activities surrounding carbon footprint reduction are currently centred in lowering the computational cost, using environmentally friendly AI

models, decreasing energy consumption, and increasing the usage of renewable energy resources. Additionally, optimization techniques focused on algorithms, hardware, and data centres emerge as valid approaches to facilitate the reduction in power consumption associated with AI. The case study findings from Zeus on optimizing GPU power limits and batch sizes for energy-efficient training and NVIDIA's accelerated computing achieving greater energy efficiency compared to traditional CPUs emphasized the potential energy savings in AI model training and the need for standardized practices to balance energy consumption, cost, and performance. Expert consultations highlighted that optimization is the currently possible mitigation approach, and that there is a need for publicly available information to consciously involve stakeholders in the carbon reduction process towards sustainable AI.

The thesis acknowledges that AI models contribute to carbon emissions and that proper calculation, estimation and reduction techniques are crucial for mitigating the rapidly expanding carbon footprint of this sector. Further research on how to calculate AI model's carbon emissions, applied optimization techniques for mitigation, and transparency in the methods used for collecting data across the AI lifecycles, is paramount in facilitating the development of future reporting guidelines and in increasing stakeholders, developers and consumers' knowledge about the topic.

6.1. Recommendations

Based on the findings of this study, it is recommended that companies begin using publicly available emission estimation tools to provide data on their environmental impact from AI models, even if the results are only estimates and lack consistency. To enhance transparency, these calculations could include error bounds and be reflected in public reporting. Simple summaries of these outcomes could be analyzed and shared to raise awareness among all stakeholders. Additionally, efforts and investments should focus on optimizing existing models, with appropriate actions implemented to guide the development of more sustainable models in the future.

7. REFERENCES

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research,* 2(3), 87–93. https://doi.org/10.1016/j.bdr.2015.04.001

Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. *Journal of Cleaner Production*, 468, 143090. https://doi.org/10.1016/j.jclepro.2024.143090

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). *Carbontracker: Tracking and predicting the carbon footprint of training deep learning models*. arXiv preprint. http://arxiv.org/abs/2007.03051

Andrews, E. L. (2020). *AI's carbon footprint problem.* Stanford Doerr School of Sustainability. https://sustainability.stanford.edu/news/ais-carbon-footprint-problem

Ayers, S., Ballan, S., Gray, V., McDonald, R., World Benchmarking Alliance, Digital Development Partnership, & Korea Green Growth Trust Fund. (2024). Measuring the emissions & energy footprint of the ICT sector. In A. Pederson, L. C. De Freitas, J.-M. Canet, Orange Group, & C. Torgusson (Eds.), *Measuring the Emissions & Energy Footprint of the ICT Sector: Implications for Climate Action.* https://documents1.worldbank.org/curated/en/099121223165540890/pdf/P1785971 2a98880541a4b71d57876048abb.pdf

Bell, J. (2022). What Is Machine Learning? *In Machine Learning and the City: Applications in Architecture and Urban Design (pp. 207–216).* https://doi.org/10.1002/9781119815075.ch18

Bannour, N., Ghannay, S., & Névéol, A. (2021). Evaluating the carbon footprint of NLP methods: A survey and analysis of existing tools. *Proceedings of the 1st Workshop on Ethics in NLP (pp. 16–24).* Association for Computational Linguistics. https://aclanthology.org/2021.sustainlp-1.2/

Berthelot, A., Caron, E., Jay, M., & Lefèvre, L. (2024). Estimating the environmental impact of generative-AI services using an LCA-based methodology. *Procedia CIRP*, 122, 707–712. https://doi.org/10.1016/j.procir.2024.01.098

BigScience Research Workshop. (2022). https://bigscience.huggingface.co/

Bjørlo, L., Moen, Ø., & Pasquine, M. (2021). The role of consumer autonomy in developing sustainable AI: A conceptual framework. *Sustainability*, 13(4), 1–18. https://doi.org/10.3390/su13042332

Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096

Bouza, L., Bugeau, A., & Lannelongue, L. (2023). How to estimate carbon footprint when training deep learning models? A guide and review. Environmental Research Communications, 5(11), 115014. https://doi.org/10.1088/2515-7620/acf81b

Castro, D. (2024). *Rethinking concerns about AI's energy use.* Data Innovation. https://www2.datainnovation.org/2024-ai-energy-use.pdf

Cho, R. (2023). *AI's growing carbon footprint.* State of the Planet. https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/

Copernicus. (2024). https://climate.copernicus.eu/

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2023). The AI gambit: Leveraging artificial intelligence to combat climate change—Opportunities, challenges, and recommendations. *AI and Society*, 38(1), 283–307. https://doi.org/10.1007/s00146-021-01294-x

DataScientest. (2024). *Computational resources: Definition, operation, and role.* DataScientest. https://datascientest.com/en/computational-resources-definition-operation-and-role

Delanoë, P., Tchuente, D., & Colin, G. (2023). Method and evaluations of the effective gain of artificial intelligence models for reducing CO2 emissions. *Journal of Environmental Management, 331, 117261.* https://doi.org/10.1016/j.jenvman.2023.117261

European Parliament. (2020). *What is artificial intelligence and how is it used?* European Parliament.

https://www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

Evans, R. (2016, July 20). *DeepMind AI reduces Google Data Centre cooling bill by 40%.* Google DeepMind. https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/

Gaur, L., Afaq, A., Arora, G. K., & Khan, N. (2023). Artificial intelligence for carbon emissions using system of systems theory. *Ecological Informatics,* 76, 102165. https://doi.org/10.1016/j.ecoinf.2023.102165

Goldman Sachs. (2024). *AI is poised to drive 160% increase in data center power demand.* Goldman Sachs. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

Hanbury, P., Wang, J., Brick, P., & Cannarsi, A. (2025). *DeepSeek: A Game Changer in AI Efficiency?*. Bain & Company. https://www.bain.com/insights/deepseek-a-game-changer-in-ai-efficiency/

Heikkilä, M. (2023, December 1). Making an image with generative AI uses as much energy as charging your phone. *MIT Technology Review.* https://www.technologyreview.com/2023/12/01/1084189/making-an-image-with-generative-ai-uses-as-much-energy-as-charging-your-phone/

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). *Towards the systematic reporting of the energy and carbon footprints of machine learning.* arXiv. http://arxiv.org/abs/2002.05651

Khan, W., Daud, A., Khan, K., Muhammad, S., & Haq, R. (2023). Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal, 4,* 100026. https://doi.org/10.1016/j.nlp.2023.100026

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). *Quantifying the carbon emissions of machine learning.* arXiv. http://arxiv.org/abs/1910.09700

Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science, 8*(12). https://doi.org/10.1002/advs.202100707

Leuthe, D., Meyer-Hollatz, T., Plank, T., & Senkmüller, A. (2024). Towards sustainability of AI – Identifying design patterns for sustainable machine learning development. *Information Systems Frontiers.* https://doi.org/10.1007/s10796-024-10526-6

Li, B., Roy, R. B., Wang, D., Samsi, S., Gadepally, V., & Tiwari, D. (2023). Toward sustainable HPC: Carbon footprint estimation and environmental implications of HPC systems. *ACM Digital Library.* https://doi.org/10.1145/3581784.3607035

Liu, V., & Yin, Y. (2024). *Green AI: Exploring carbon footprints, mitigation strategies, and trade-offs in large language model training.* arXiv. http://arxiv.org/abs/2404.01157

Lottick, K., Susai, S., Friedler, S. A., & Wilson, J. P. (2019). *Energy usage reports: Environmental awareness as part of algorithmic accountability.* arXiv. http://arxiv.org/abs/1911.08354

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). *Estimating the carbon footprint of BLOOM, a 176B parameter language model.* arXiv. http://arxiv.org/abs/2211.02001

Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of AI deployment? *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 85–99). https://doi.org/10.1145/3630106.3658542

Luscombe, R. (2024, September 20). Three Mile Island nuclear reactor to restart to power Microsoft AI operations. *The Guardian*. https://www.theguardian.com/environment/2024/sep/20/three-mile-island-nuclear-plant-reopen-microsoft

Microsoft. (n.d.). *2024 Environmental Sustainability Report | Microsoft CSR. Microsoft Sustainability.* Retrieved February 16, 2025, from https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report

McCarthy, J. (2007). *Professor John McCarthy: Father of AI.* http://www-formal.stanford.edu/jmc/

Nvidia. (2024). *NVIDIA sustainability report fiscal year 2024: Climate and efficiency, people, diversity, and inclusion.* https://images.nvidia.com/aem-dam/Solutions/documents/FY2024-NVIDIA-Corporate-Sustainability-Report.pdf

Patterson, D. (2022). *The carbon footprint of machine learning training will plateau, then shrink.* TechRxiv. https://doi.org/10.36227/techrxiv.19139645.v1

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv.* https://doi.org/10.48550/arXiv.2104.10350

Röger, T., Steinle, F., & Schilp, J. (2024). Monitoring Data Quality for AI Models in Industrial Glass Production. *IFAC-PapersOnLine*, *58*(27), 282–287. https://doi.org/10.1016/j.procir.2024.10.088

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2019). *Tackling climate change with machine learning.* arXiv. http://arxiv.org/abs/1906.05433

Rozite, V., Reidenbach, B., & Bertoli, E. (2023). *Data Centres & Networks.* International Energy Agency. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

Sarajcev, P., Kunac, A., Petrovic, G., & Despalatovic, M. (2022). Artificial intelligence techniques for power system transient stability assessment. *Energies,* 15(2), 507. https://doi.org/10.3390/en15020507

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). *Green AI.* arXiv. http://arxiv.org/abs/1907.10597

Shahinfar, S., Meek, P., & Falzon, G. (2020). "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Ecological Informatics*, 58, 101081. https://doi.org/10.1016/j.ecoinf.2020.101085

Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and Policy Considerations for Deep Learning in NLP*. arXiv. http://arxiv.org/abs/1906.02243

Stuart J. Russell and Peter Norvig. (2020). *Artificial Intelligence a Modern Approach Third Edition*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction Second edition, in progress*.

TED AI. (2024). *What are parameters in ai?*. https://tedai-sanfrancisco.ted.com/glossary/parameters/

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). *The Computational Limits of Deep Learning*. arXiv. http://arxiv.org/abs/2007.05558

Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-54271-x

Tornede, T., Tornede, A., Hanselle, J., Mohr, F., Wever, M., & Hüllermeier, E. (2023). Towards Green Automated Machine Learning: Status Quo and Future Directions. *Journal of Artificial Intelligence Research*, *77*, 427–457. https://doi.org/10.1613/jair.1.14340

Trébaol, T. (2020). *École Polytechnique Fédérale de Lausanne CUMULATOR-a tool to quantify and report the carbon footprint of machine learning computations and communication in academia and healthcare*. https://infoscience.epfl.ch/entities/publication/2a903f44-1188-4e03-8204-5d63beeca4d9

United Nations Framework Convention on Climate Change (UNFCC). (2024). *The Paris Agreement*. https://unfccc.int/process-and-meetings/the-paris-agreement

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 13, Issue 4). John Wiley and Sons Inc. https://doi.org/10.1002/widm.1507

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., … Hazelwood, K. (2021). *Sustainable AI: Environmental Implications, Challenges and Opportunities*. arXiv. http://arxiv.org/abs/2111.00364

Xu, J., Zhou, W., Fu, Z., Zhou, H., & Li, L. (2021). *A Survey on Green Deep Learning*. arXiv. http://arxiv.org/abs/2111.05193

You, J., Chung, J.-W., & Chowdhury, M. (2022). *Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training.* arXiv. http://arxiv.org/abs/2208.06102

Zvornicanin, E. (2024). *Comparative analysis of top large language models*. Baeldung. https://www.baeldung.com/cs/top-llm-comparative-analysis

8. APPENDICES

Annex A: Interview Protocol

This annex contains the questions used during the expert consultations with AI/ML researchers. The following questions were included in the Survey Questionnaire aimed at measuring and reducing the carbon footprint of AI models.

Survey Purpose: This survey aims to understand stakeholder engagement and awareness regarding the carbon footprint of AI models, as well as to gather insights on calculation and reduction strategies.

Survey Questions:

1. What is your current role in relation to AI models?
2. Have you ever been involved in the calculation of the carbon footprint of an AI model? What were your observations? [Your perspective is greatly appreciated.]
3. Have you ever been involved in the reduction of the carbon footprint of an AI model? What were your observations? [Your perspective is greatly appreciated.]
4. If your answer to Questions 2 and 3 is no, how aware are you that AI models contribute to carbon emissions? What is your level of knowledge about this issue, and how do you perceive its impact?
5. What is your experience at Zeus regarding carbon reduction and what would be your contribution as the main researcher for those case studies?
6. Is there anything else you would like to add regarding the carbon footprint of AI models or strategies for its reduction?

Confidentiality Statement: Your responses will be kept strictly confidential and used only for academic research purposes. Your participation is greatly appreciated, and your insights will contribute significantly to the thesis.