# Sustainable AI: Rethinking the AI Revolution

Brendan P. Conroy
conro188@morris.umn.edu
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

## Abstract

Artificial intelligence (AI) models have continued to grow more and more advanced, growing both in size and model accuracy. This trend of exponential model growth has some serious environmental implications, however. While models have been growing exponentially, the efficiency of these models has been lagging behind. Furthermore, we have seen a quickly growing carbon footprint of AI. In order to address this growing carbon footprint, we need a fundamental shift in the way we approach developing and designing AI models, a shift towards a "Green AI" mindset that considers the environmental implications of this technology across every phase of development. This shift is very necessary to prevent out of control carbon emissions and the potential damage those emissions will have on our climate.

***Keywords:*** artificial intelligence, neural networks, sustainability, deep learning, carbon footprint

## 1 Background

Artificial intelligence (AI) has rapidly advanced as a technology in recent years, and there is no doubt we greatly benefit from this growing technology in many ways. We see AI applications across several different industries, including, science, medicine, finance, and education. [6] AI refers to any time that machines are performing tasks that would typically require human intelligence. The term was coined in 1956 at a Dartmouth conference, and the field of AI when it was young primarily focused on theoretical models that were greatly limited by a lack of computing power and a lack of available data. Increases in both computing power and available data in recent times have enabled AI to grow rapidly as a technology, and have applications across several different industries and use cases.

Machine learning refers to a subset of artificial intelligence that involves a model that gets trained on some set of data, learns patterns from that data, and is then able to extrapolate to data outside of the data set, and make predictions for that other data. A useful abstraction for envisioning these models is to think of them as a giant math equation with millions of unknown variables or parameters on the inside of the equation. The equation takes some input, and has a goal of producing some expected output given that input. During training, the millions of variables or parameters within the equation are being adjusted/fine-tuned, such that the output is close to what is expected. After training, the model is expected to work on other data outside of the training set.

Neural networks are a subset of machine learning that refers to a specific structure that mimics the system of layers of neurons in the human brain. Deep learning refers to neural networks with multiple layers.

Green or sustainable AI, refers to a specific approach to AI that involves considering the environmental implications of the technology of AI, more specifically, it considers the carbon footprint of AI through the development of AI models. It seeks to change the way that we understand and approach creating AI.

The carbon footprint of AI refers to the emissions of greenhouse gases associated with AI. Greenhouse gases are gases that are involved in the greenhouse effect, a natural phenomenon in which certain gases have the property of trapping heat in our atmosphere, which accelerates global warming and climate change. We want to limit the greenhouse gas emissions/carbon footprint of AI, as we want to limit the contribution of this technology to accelerating climate change.

## 2 Introduction

Artificial Intelligence, or AI for short, is one of the fastest growing technologies in our world today. AI has grown to be broadly used across many sectors, and to solve many different problems that different industries face that require human-level thinking to provide solutions.

While AI has been around for a while, the degree to which we depend upon and use AI as a society is increasing rapidly. As our dependence upon AI continues to increase, the models and infrastructure that enables AI to keep growing increases as well. We have consequentially seen a great increase in the carbon footprint of AI, as bigger and more powerful (but not necessarily more efficient) models require more and more computational power.

Unfortunately, a sustainability approach to AI has not been taken across the machine learning development process, and AI Models have been designed and trained to prioritize accuracy and computational power over efficiency. Additionally, the amount of data that feeds these models has been sharply increasing, which has lead to growth in model size. The environmental footprint of AI includes every step of the AI development process, including experimentation, training, and inference. Beyond this, the footprint of AI also includes the emissions across the entire life cycle of hardware systems. All emissions across the entire system life cycle of AI hardware need to be considered as a necessary part of the transition to "Green" AI.

The need for a transition to a Green AI approach has been made more necessary from several developments in the last decade. For starters, the size of data that AI models use to train (training data) has increased exponentially in the last decade.[6] This increase in data size requires more infrastructure for data storage, and also requires more power consumption. The exponentially increasing data size has lead to a great increase in model size growth. While the size of AI models has been growing rapidly, the memory capacity of the hardware used to run said models has not been increasing at the same rate. In a response to the increasing data sizes and model sizes, we have also seen a great increase in AI infrastructure growth.

While the carbon footprint of AI is largely determined and influenced by these factors, it also goes beyond the operational energy use of AI. The embodied carbon footprint of AI, the carbon footprint of the entire life cycle of infrastructure associated with AI. It includes manufacturing of hardware, production of data and software, transportation, and disposal. Overall, the carbon footprint of AI includes both of these components, operational energy use, and embodied footprint.

The desire to achieve better-performing AI models has led to a trend of models becoming larger and more complex at an increasing rate [6]. This pursuit of higher quality comes with significant environmental consequences. To fully grasp the environmental footprint of AI we have to look beyond training, experimentation, and inference, and consider the broader embodied carbon footprint of AI. We also need to consider the complete machine learning pipeline, and include data collection, model experimentation, model optimization, and run-time inference. We have to consider both the frequency of which each stage in the pipeline is performed, along with the scale of the operation in each stage, and the full life cycle of hardware that is associated with any stage of the machine learning pipeline. It is necessary to not just focus on one element of the footprint, as just focusing on the operational energy use might make a system rely more an a greater embodied carbon footprint. While transitioning data centers to carbon-free and green/renewable energy source may seem like a viable solution, we run into geographic issues of availability of sustainable energy and also the issue of lack of green infrastructure that takes time and financial investment to build. In other words, the availability of access to green and renewable energy everywhere is not uniform and evenly distributed. Therefore we need to focus on reducing the carbon footprint of AI systematically and holistically.

In this paper I will be discussing some key aspects of AI and ML, such as phases of ML model development and and the system life cycle of AI, and how these topics relate to a Green AI mindset. I will then discuss the carbon footprint of AI, including a case study on several ML models developed by Meta. Lastly I will discuss the components of a Green AI mindset, and I will draw conclusions from this work.

## 3 AI and ML Background

The shift towards Green AI requires a holistic focus on every phase of model development for machine learning models, along with a focus on ML infrastructure across the entire system life cycle. The machine learning phases of development include data processing, experimentation, training, and inference. The infrastructure supporting different phases of AI development is tailored to achieve specific goals. Both elements of this holistic focus are described in this section.

### 3.1 Phases of Model Development

Data processing, the first phase of ML model development, is the phase in which raw data is collected and cleaned into a collection of data that works for the given ML algorithms. Data processing includes cleaning the data such that it is ready for training, for example, removing non-integer data, removing outliers in the data, formatting data for consistency, and converting categorical data into numerical data. Data in this stage is also split into training data, and validation data. The data processing stage of model development is essential for training accurate models. Typically, data processing has a negligible effect on carbon footprint. [6]

After the data is processed, it is ready to be fed into a ML for training. Prior to the training phase of development, however, an experimentation phase is required to determine the most efficient model architecture and hyper-parameters for the given problem. Different models and algorithms are considered and tested, and the performance of the different architectures is analysed such that the best architecture can be selected for the given problem. Hyper-parameters, or parameters that are adjusted prior to the learning process, are changed and explored with different models to determine the optimal architecture for the problem at hand. A large collection of different ML structures and hyper-parameters is often explored simultaneously, which can require a great amount of computational power. This implies that this phase of development often contributes greatly to the carbon footprint of AI.

Following experimentation is training. Training is the phase of model development in which the processed data is taken and fed into the selected model architectures following experimentation. This is the phase in which the "learning" happens— where models learn patterns and relationships from the training data in order to make predictions on new data. The learning works by the model adjusted parameters in attempts to achieve desired output. To achieve desired output, the model aims to minimize a loss function, which is a function that captures the different between the models prediction and actual expected output. This phase of model development usually accounts for the largest contribution to carbon footprint. [6]

The last phase of model development is inference. This is the phase in which the trained model is ready to make predictions on data outside of its training data, with parameters adjusted optimally following the patterns learned during training. During inference, the model is given new input, and it returns some type of output or prediction. If trained well with a diverse set of properly cleaned data, the model is expected to make predictions with minimal bias, however it is always a possibility that the model learns some bias from the training data. The inference phase of development also has a significant contribution to the carbon footprint of AI, as unlike the other phases, the power consumption of the phase is not fixed after training. Every time a trained model is used to make a prediction, for example, when a text prompt is entered into a GPT like ChatGPT, the model consumes power to generate the output and that contributes to the carbon footprint. In other words, the carbon footprint of this phase of development often continues to grow after training and model development.

### 3.2 System Life Cycle

Part of the carbon footprint of AI and ML models comes from the system life cycle of hardware and infrastructure associated with models. The system life cycle of hardware can be divided into four major phases: manufacturing, transport, product use, and recycling. The carbon footprint of AI that comes from the manufacturing phase of the system life cycle can be referred to as embodied carbon footprint. The carbon footprint of AI that comes from product use, or emissions that result directly from the use of AI, can be referred to as operational carbon footprint. Both the embodied carbon footprint and operational carbon footprint of AI are major components of AI's overall carbon footprint.

## 4 The Carbon Footprint of AI

To examine the carbon footprint of AI, we will look at the operational carbon emissions for several machine learning models being developed by Meta (see Figure 1). [6] Six deep learning models developed by Meta, LM, RM-1, RM-2, RM-3, RM-4, and RM-5, are compared to 7 open source models, BERT-NAS, T5, Meena, GShard-600B, Switch Transformer, and GPT-3. LM refers to Meta's Transformer-based Universal Language Model for text translation, while the RM models are deep learning recommendation models used for recommending and ranking various Meta products. The operational carbon footprint of all models is divided amongst offline training, which includes experimentation and training models with historical data, online training, which includes training models with recent data, and inference.

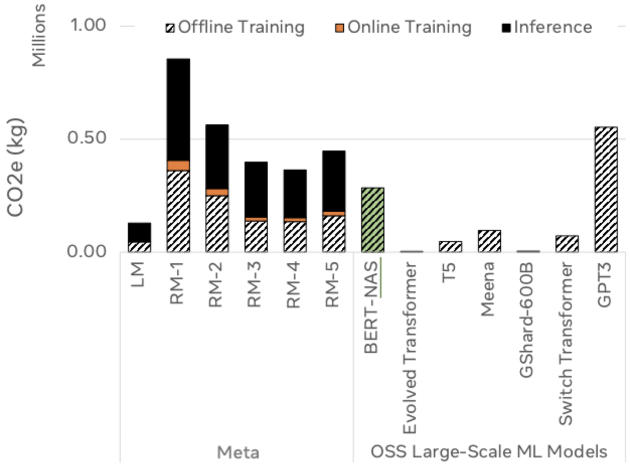

**Figure 1.** Operational Carbon Footprint of Large-Scale ML Tasks

There are several interesting takeaways from the data in Figure 1. For LM, the language model for text translation, the carbon footprint mostly comes form inference, with a smaller footprint coming from offline training. For the recommendation and ranking models, most of the footprint of the models is dominated by inference, however the footprint is more evenly divided among inference and offline training. Online training accounts for the smallest part of the operational footprint of these models. For the Meta models, training and inference seem to be the phases of model development that incur the largest carbon footprint. [6]

Looking at the operational carbon footprint of the open-source models, we observe that interestingly, the size of a model or number of parameters does not directly predict a higher operational carbon footprint. For example, the Switch Transformer model, which is trained with 1.5 trillion parameters, has a significantly smaller operational carbon footprint than the GPT-3 model, which has 750 billion parameters. As we see that larger models don't necessarily correlate with higher operational carbon footprint, it is important to observe that models can achieve a lower carbon footprint if they emphasize an efficiency-based approach to development of models. Models with a large number of parameters can have a significantly reduced operational carbon footprint with an efficient model architecture.

In order to calculate the overall carbon footprint of these different models, we need to first estimate the embodied carbon cost of the models, and then add that to the calculated operational carbon cost. The embodied carbon footprint of the 5 Meta models is estimated in Figure 2, along with the operational carbon footprint of the models, giving us the total carbon footprint of the models altogether. For all the models, the operational carbon cost greatly outweighs the projected embodied carbon cost of each model. Meta uses renewable solar energy to cover the majority of the embodied

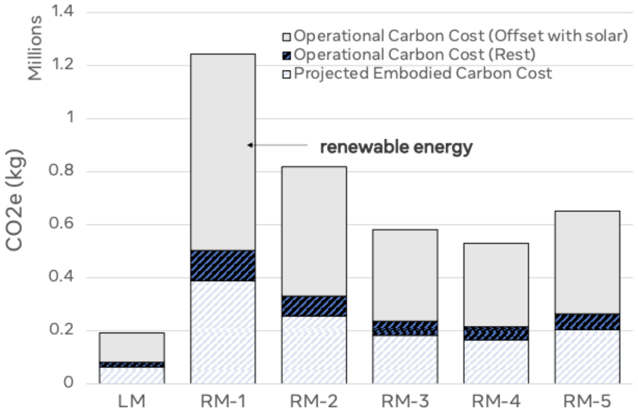

**Figure 2.** Overall Carbon Footprint of Large-Scale ML Tasks

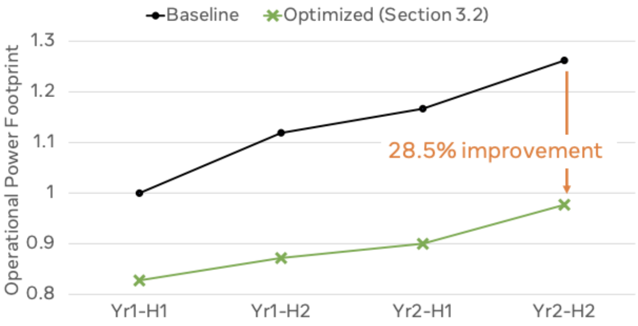

**Figure 3.** Operational Energy Footprint Reduction for Meta's AI Fleet

carbon cost of these models. Accounting for the carbon-free solar energy that is covering the majority of the operational carbon costs, the majority of the overall carbon costs of these models comes from the embodied carbon cost, or the manufacturing of infrastructure and hardware associated with the models. It is important to note that Meta has been able to significantly reduce its overall carbon footprint for these models by adopting renewable solar energy as an energy source for its operational energy needs. [6]

## 5 A Case Study of Meta's Model Optimization

While optimization of models has lead to a reduction in energy consumption, AI infrastructure continues to grow and expand, resulting in overall growing operational energy footprint for Meta's AI fleet (see Figure 3). Regardless, Meta has been able to reduce the overall operational energy footprint for it's AI fleet by 28.5% over the last 2 years.

The optimization of models can be broken down into 4 categories. Model optimization includes designing more resource efficient models. Infrastructure optimization refers to data center optimization, and hardware optimization refers to optimizing hardware to reduce overall operational energy costs. Lastly, another category is platform optimization, which involves enhancing the performance and efficiency of the software platforms or frameworks used to deploy and run machine learning models.

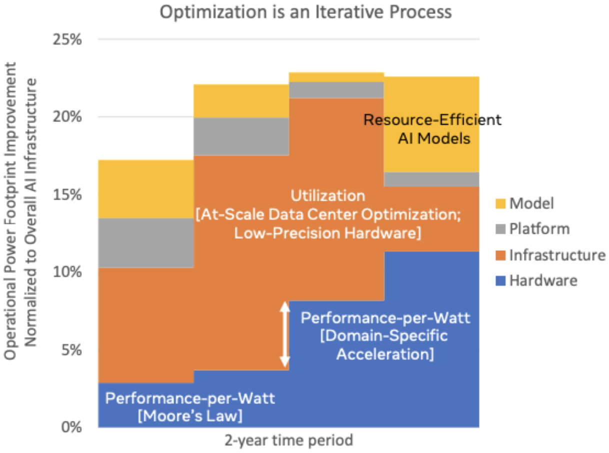

**Figure 4.** Operational Energy Footprint Reduction for Meta's AI Fleet

Figure 4 shows the operational energy footprint reduction over 2 years for Meta's AI fleet divided into the four aforementioned categories. Each bar represents operational power reduction over a 6-month period from each of the optimization categories. Every 6 months sees on average an operational power consumption reduction of about 20%. [6]

While overall operational energy consumption continues to grow, coupling optimization of operational energy uses with a renewable energy approach, as we have seen with Meta, leads to significant reductions in overall carbon footprint. The majority of the total carbon footprint falls on the embodied carbon costs of the infrastructure, which is the component of the total footprint of AI that needs the most work done on minimizing carbon costs. [6]

## 6 A Green AI Mindset

A shift away from our current mindset towards one that considers the environmental implications of the exponential scaling and growth of AI is necessary. While there have been many efforts to look at AI model system efficiency and infrastructure efficiency, optimizing data, experimentation, and training algorithm efficiency hasn't been explored as much. An approach to AI sustainability that encompasses optimization of every phase of the model development cycle is necessary to significantly improve the footprint of AI. This new mindset has three main components: redefining success in the context of ML models, taking a more holistic approach to capturing AI's footprint, and assuming a responsibility to minimize the carbon footprint of AI.

We need a large shift in the way that we define a model's success. The current idea of how successful a model is equates most directly with the accuracy and prediction quality of the model. A model that expands its size and energy consumption drastically to improve accuracy is said to be representative of "progress" in this field. In other words, efficiency isn't looked at as a benchmark of model success in the way that accuracy is. This leads to massive scaling of models in the endless pursuit of accuracy that largely ignores model efficiency and environmental footprint. An AI sustainability mindset says that we need to change this. Instead, our goal should be to improve model efficiency along with model accuracy, such that we see a trend of improving model accuracy with a fixed or decreasing rate of computational cost. In other words, efficiency needs to be used to evaluate models alongside accuracy. We need to introduce a new key question: At what cost do these models achieve accuracy? Encouraging more model transparency is key to incentivize model efficiency, and allowing for an environment of learning from others surrounding designing more efficient models. We need to redefine successful AI as AI that achieves the same new and exciting results and accuracy, while minimizing computational cost with efficient design principles. [4]

A step that can be taken to encourage this shift in our mindset, is to encourage reporting efficiency along with accuracy alongside models. Requiring academic work done surrounding ML models to report and detail efficiency incentivizes developers to take the most efficient approach and optimize model development along every phase of model development. One metric that can be used as a metric of efficiency for models is floating point operators (FPOs). FPOs provide an overall estimation of work performed by the computational process, and furthermore have potential to be a good metric for efficiency.

Taking a more holistic approach to AI involves including both operational and embodied carbon costs in the total carbon footprint for AI, but also examining the embodied footprint of associated hardware and infrastructure along the complete system life cycle. This approach also means optimizing efficiency for every stage of model development.

Lastly, this new approach to AI says that we have a responsibility to make this change. It stems from a broader technological ethics lens, which says that the creators and users of a new technology have a responsibility to reduce the negative implications of the given technology. Even if you are someone not directly involved with the development of models, this can still apply to you. This means being mindful of usage of AI, limiting use to necessary cases, voting, and spreading awareness of the carbon footprint of AI.

### 6.1 Data Scaling

One of the most common approaches to improving model accuracy is to increase the size of data. Rather than optimizing algorithms, accuracy is often attempted to be improved by training models with larger and higher quality training data. The problem with this approach is that with larger datasets, system resources have to be made larger, and more computational power is required. This requires more embodied carbon footprint, as more powerful infrastructure is required, and a larger operational footprint, as more data is stored and used to train the model during the training phase of model development. Even if infrastructure is not upgraded to keep up with the larger datasets, a larger burden is placed on the operational side of things, as it takes much longer to train the model with the larger dataset.

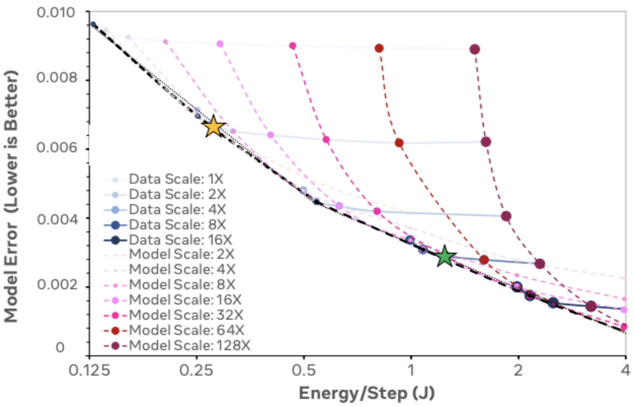

**Figure 5.** Model Quality vs. Energy Consumption

In Figure 5, energy consumption required when training (x-axis), is displayed against model accuracy (y-axis). The model shows us a diminishing rate of model accuracy increase with more energy consumption. In other words, past a certain point, to achieve a only slightly higher accuracy, significantly more power is required. [6]

### 6.2 Memory-Efficient Model Architecture

An efficient solution to optimizing the footprint of experimentation and training phases of model development is to develop resource-efficient model architectures. One approach to this is memory-efficient model architectures. Memory-efficient model architectures require less memory and more efficiently utilizes accelerators. Some memory-efficient techniques include parameter-reduction techniques, in which the number of parameters in the model is reduced without significantly sacrificing performance, and architecture design approaches, which focus on a more memory-efficient model design. [5]

Parameter reduction techniques, such as pruning, quantization, and knowledge distillation, aim to minimize the number of parameters in the model without compromising performance. Pruning identifies and eliminates redundant or less crucial parameters, while quantization reduces the precision of weights and activations. Knowledge distillation involves training a smaller model to mimic the behavior of a larger, more complex model, effectively transferring the knowledge from the larger model to the smaller one.

In addition to parameter reduction techniques, architecture design involves designing models with memory-efficient architectures. Approaches include reducing model depth by using shallower models with fewer layers, and leveraging factorization techniques to decompose large layers into smaller ones.

### 6.3 Efficient use of Hardware

A Green AI approach should include maximizing the efficiency of system resources, while prolonging the life of AI infrastructure. This involves shifting system design from a focus on operational energy optimization to include embodied carbon cost and the life cycle of hardware. [2]

This means that we need to carefully consider how hardware resources are utilized throughout the entire life cycle of AI infrastructure. One key aspect is optimizing the utilization of hardware resources during both training and inference phases of model development. Techniques such as processing multiple data samples simultaneously can lead to more efficient use of computational resources.

Also, extending the lifespan of AI hardware infrastructure through techniques like hardware refurbishment, component reuse, and responsible disposal practices can contribute to reducing the environmental impact of AI systems. By prolonging the life of hardware components and minimizing electronic waste, it is possible to lessen the carbon footprint associated with manufacturing and disposing of hardware.

### 6.4 Carbon-Efficient Scheduling

A solution that has arisen in response to the issue of AI's rapidly growing carbon footprint has been green and clean energy integration into operational energy costs. While this approach has the potential to significantly reduce the carbon footprint of AI, it leads to a new problem— data centers have to be able to adopt to a renewable source of energy that naturally has more fluctuations in production levels. Data centers thus require schedulers that are aware of this, and can smartly and efficiently schedule workloads in a way that can predict patterns of intermittent production and schedule accordingly. [3]

Once renewable energy integration becomes more common, and our grids become more reliant on intermittent renewable energy, AI models can provide a solution to the issue of fluctuations in renewable energy production. Models can be trained to predict these fluctuations, and can inform smart schedulers such that the power grid is more balanced and stable. [1]

### 6.5 Additional Ways to Promote Green AI

There are some additional suggestions to promote Green AI. One approach might be to encourage more reporting surrounding the experimentation phase of model development. Currently this phase is not typically reported. Reporting information about how many different model architectures were explored, and what approaches were successful and unsuccessful can lead to more learning about efficiency during this phase of model development, and help others to optimize models for more efficiency.

Another step that can be taken is to encourage releasing trained models to the public. [6] This would prevent the unnecessary work of retraining models, and lower the carbon footprint overall.

## 7 Conclusion

It is critically important that we make a shift in the way that we approach developing and discussing AI. This shift needs to include holistically examining the environmental footprint of AI along every phase of development, and every life cycle stage of associated hardware. We have a responsibility to manage the negative environmental implications this rapidly growing technology has on our world.

## Acknowledgments

Thank you to all of the Computer Science faculty at Morris, Elena Machkasova, K.K. Lamberty, Peter Dolan, Nic McPhee, Kristofer Schlieper, and Wenkai Guan. Thanks to Joseph Thelen for reviewing my work and providing me with very helpful feedback. Thanks to Udit Gupta, his work inspired my work here and helped my understanding of the material. His website can be found at: ugupta.com. A special thanks goes to my advisor for this project as well, Wenkai Guan.

## References

[1] Tanveer Ahmad, Dongdong Zhang, Chao Huang, Hongcai Zhang, Ningyi Dai, Yonghua Song, and Huanxin Chen. 2021. Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities. *Journal of Cleaner Production* 289 (2021), 125834. https://doi.org/10.1016/j.jclepro.2021.125834

[2] Lieven Eeckhout. 2024. FOCAL: A First-Order Carbon Model to Assess Processor Sustainability. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. Association for Computing Machinery, New York, NY, USA, 401–415. https://doi.org/10.1145/3620665.3640415

[3] Walid Hanafy, Qianlin Liang, Noman Bashir, Abel Souza, David Irwin, and Prashant Shenoy. 2024. Going Green for Less Green: Optimizing the Cost of Reducing Cloud Carbon Emissions. 479–496. https://doi.org/10.1145/3620666.3651374

[4] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. arXiv:1907.10597 [cs.CY]

[5] Tom Veniat and Ludovic Denoyer. 2018. Learning Time/Memory-Efficient Deep Architectures with Budgeted Super Networks. arXiv:1706.00046 [cs.LG]

[6] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities.