*International Journal of Engineering Research and Development*
*e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com*
*Volume 21, Issue 7 (July 2025), PP 33-49*

# Environmental Impacts of Artificial Intelligence Systems

Brahmaleen Kaur Sidhu[1], Gurjit Singh Bhathal[2]
[1,2]*Department of Computer Science and Engineering*
*Punjabi University, Patiala, Punjab, INDIA*
[1]*Corresponding Author*

***ABSTRACT***
*Artificial Intelligence (AI) offers transformative potential across industries, but its development and deployment come with environmental costs. This paper discusses topics such as the carbon footprint of AI models, methods for measuring and reporting environmental impacts, and challenges in estimating the sustainability of AI technologies. It provides insights into energy and carbon accounting, along with case studies demonstrating how AI's environmental footprint is assessed. The paper provides a comprehensive understanding of the relationship between AI and the environment, equipping readers with knowledge to contribute to more sustainable AI practices.*

***KEYWORDS***
*Carbon Accounting, Carbon Emissions, Carbon Intensity, Data Center Energy Use, Eco-friendly AI, Emissions Reporting, Power Usage Effectiveness, Sustainable AI Development*

---

Date of Submission: 02-07-2025 Date of acceptance: 12-07-2025

---

## I. INTRODUCTION

### 1.1 AI Definitions

The term artificial intelligence (AI) was coined in 1955 by Professor of Computer Science, John McCarthy. He defined it as the science and engineering of making intelligent machines. There was another term before AI which has played a foundational role in its development, called cybernetics. Cybernetics is the interdisciplinary study of control and communication in animals, humans, and machines. Coined by Norbert Wiener in the 1940s, it focuses on how systems regulate themselves through feedback loops, enabling stability, adaptation, and goal-directed behavior. At its core, cybernetics deals with how information is transmitted, processed, and used to control actions and make decisions. Cybernetic principles, especially feedback loops, underpin many AI systems such as reinforcement learning, where agents learn by receiving rewards or penalties from the environment. Cybernetics emphasizes how systems can self-regulate and adapt, which is critical for autonomous AI agents (like robots or self-driving cars) that need to make real-time decisions. AI-driven robotics is deeply inspired by cybernetics, where sensors (input), processors (decision), and actuators (output) form a control loop to perform intelligent actions. Cybernetics studies how humans interact with machines, influencing AI development in areas such as natural language processing, human-computer interaction, and cognitive systems. Early AI models, including neural networks, were inspired by cybernetic models of the human brain, emphasizing how biological systems process and act on information. In essence, cybernetics laid the conceptual groundwork for AI by framing intelligence as a system of communication, adaptation, and control—principles still central to modern AI research and applications.

Machine learning is part of AI focused on the development and study of statistical algorithms that can learn from data. Forexample, a machine learning model may be used to analyze weather data then make future predictions of the weather.Professor John McCarthy describes AI as the science and engineering of makingintelligent machines, while machine learning is the statistical use ofalgorithms to make machines seem intelligent.There are a few different types of machine learning algorithms.Supervised machine learning algorithms use training data where the expectedoutput is labeled. A supervised learning algorithm for image classification wouldhave training data including pictures with labels such as hand, cat, traffic lights, etc. Unsupervised machine learning uses unlabeled training data and learns frompatterns inherent in that data.So training data for an unsupervised machine learning algorithm for imageclassification would have images but no labels.Reinforcement learning is when the machine learning model learns by receivingrewards and penalties. Theagent is rewarded for correct moves and punished for the wrong ones.In doing so, the agent tries to minimize the wrong and maximize the right.An algorithm for self-driving cars may utilize reinforcement learning to createlearning policies for how to drive on the highway, how to stay within lanes andchange lanes.

Supervised, unsupervised, and reinforcement learning are all types of machinelearning.And within those categories, there are many types of machine learningalgorithms, from logistic regression, a supervised learning algorithm, toprincipal component analysis, an unsupervised learning algorithm.And of most interest isan algorithm called an artificial neural network.

An artificial neural network is a computational model inspired by the humanbrain.As shown in figure 1, it consists of interconnected nodes, artificial neurons organized in layersthat process input to produce output, learning and adapting through theadjustment of what we call weights in the connections based on training data.So very simplistically, training a neural net is a whole bunch of matrixmultiplication where the nodes are vectors multiplied by weights to createoutput vectors.

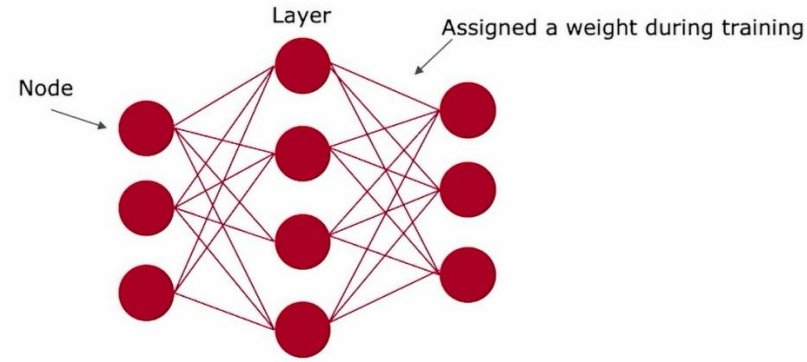


**Figure 1 Artificial Neural Network**

Deep learning is the use of a large multi-layer artificial neural networkthat compute number representations of that training data.Neural networks and deep learning can be supervised, unsupervised, orincorporate reinforcement learning. Oneof the most common components of the engineering architecture of theneural net is the transformer.A transformer is a neural net which incorporates context of data via anintention mechanism, allowing powerful and computationally efficient analysisand generation of sequences, such as words or paragraphs. A foundation model is a neural network trained by self-supervision on large-scale broad data that can be easily adopted to perform a wide range ofdownstream tasks.It often has a transformer architecture. Modelslike GPT-4 and Cloud4 are foundation models.With the advent of foundation models, generative AI has become mainstream.

Generative AI is neural network-based models trained on large amounts of datathat learn the underlying patterns to generate new data mirroring that trainingdata. When asked to make apicture of a dog, it creates a dog, but not the exact dog from any of the training data.Traditional AI, or discriminative AI, is machine learning systems designed tomake specific predictions or decisions based on a particular set of inputs. Thisis the type of AI that has been incorporated in many applications foryears. For example, Google Maps telling the bestway to a destination, or Netflix recommendations. Discriminative AI may be trained on a bunch of animal pictures, and when given a picture, it can tell what is in that picture.Foundation models and generative AI are developed with deep learning.And they can be multimodal, meaning the inputs and outputs can take many forms.The most talked-about form at the moment is text-to-text, wherein input and output are bothtext. Other forms are text-to-image, image-to-text, and even text-to-video.

Parametersare variables in an AI system whose values are adjusted duringtraining to create a desired output.For example, all those connection weights in a neural network pictureare considered parameters. Theseparameters are adjusted during the model training process.There is a pre-training phase, which refers to the step where the model istrained on a whole bunch of general data, inthe world of large models, a few billion words, generally from theinternet.Then there is fine-tuning, where the model is further trained on a smaller setof datawhich is context specific.Inference is the process when the model processes inputs and then produces outputs. Every time a question is typed into ChatGPT and it provides ananswer, it just performs inference.

Training these AI models requiresa lot of compute. Compute is measured in floating-point operations, or flops.Flops are the number of calculations, so additions, multiplications, performed. Since training a neural net is very simplistically a bunch of matrix multiplication, calculation steps are needed to perform this multiplication, which translate into flops. Flops are often used as an approximation of computational resources used in model training or used to discuss the efficiency of a model. The less flops per parameter means the model is more efficient at doing thecalculations. A Graphics Processing Unit (GPU) is a type of hardware on which neural networks are trained.It is a specialized version of a Central Processing Unit, the processorin a computer that executes instructions or algorithms.Lastly, the data center is where the compute infrastructure is housed.

Flops, GPUs, CPUs, and data centers are important whenconsidering the environmental impacts of AI. For instance, data centerstake a significant number of resources to function, between the electricity torun and keep the GPUs cool to the water needed to keep the cooling systemrunning.

**1.2 A Brief History of AI**

As mentioned earlier, many trace the origins of AI toProfessor John McCarthy, who coined the term in the 1950s.There were others around that time.Alan Turing, who wrote about the possibility of machine intelligence in the1940s and 50s, or Norbert Wiener, a founder of cybernetics.But the field extends back earlier.Traditions of machines that imitate intelligence stretch back centuries.An earlier word for artificial humans and animals, automaton, stemmed fromGreek roots meaning self-moving.Self-moving machines were inanimate objects that seemed to borrow the definingfeatures of living creatures.Vokosin's duck was an example of this.It was a mechanical duck which appeared to have the ability to eat kernels ofgrain.And Manzetti's flute playing automation was in the shape of a man.He was life-sized, seated in a chair, and hidden inside the chair were levers connecting rods and compressed airtubes, which made the lips and fingers move.Dendrel was an AI project started in the 1960s at Stanford University by EdwardPhenenbaum.It was a computer system which automated the decision-making process and problem-solving behavior of organic chemists. It utilized expert systems, which is an algorithm designed to solve complexproblems by reasoning, mainly through if-then rules.
Inthe 1970s, backpropagation was invented.This is gradient estimation commonly used for training neural networks.And between the 1970s and the 1990s, there were two AI winters, a period ofreduced funding and interest in AI research, though therewere still many who were working on it.And in the mid-90s, the second AI winter begins to thaw.And in 1997, Deep Blue, an IBM supercomputer expert system trained to playchess, defeats the then world champion chess player.And in 2005, Stanley, a self-driving car developed at Stanford University, wonthe DARPA Grand Challenge, which is a self-driving car competition.In 2011, IBM Watson wins Jeopardy.This is the leading edge of a new generation of computers capable ofunderstanding questions posed in natural language and answering them far moreaccurately than any standard search technology.This represented a big leap in natural language processing.And in 2012, a convolutional neural network called AlexNet won the ImageNetChallenge.The ImageNet Challenge is where researchers run image classification algorithmson ImageNet, a database of over 14,000 images.The goal is to create an algorithm that correctly identifies as many of theimages as possible.

AlexNet was the first time a neural network competed in the ImageNet Challenge.It made the community realize the power of neural nets when there was asufficient amount of training data.In the ImageNet database, an AlexNet winning that challenge was part of whatkicked off the next era of AI systems.Systems based on neural networks trained on large amounts of data. In 2016, AlphaGo, a deep neural network trained by Google DeepMind,mastered the ancient game of Go.Defeating a Go world champion and built upon this new era of AI, systems basedon neural nets trained on a lot of data, this time for playing games.In 2019 and 2020, OpenAI developed GPT-2 and GPT-3, generative pre-trainedtransformer, a deep neural network trained using the internet to generate whatseems like any type of text.And in 2022, they released it with a nice user interface to the public andcalled it ChatGPT. Themoment ChatGPT was released publicly, the moment thepublic became aware of the progress and opportunities in the field ofAI that had been happening for decades.Since then, many similar models have been released.

**1.3 Compute, Data, and Algorithms**

As mentioned earlier, there are three key ingredientsthat have enabled recent progress on AI: Compute, data, and algorithms.The chart in Figure 2 is from Stanford University's AI Index and shows the compute andteraflops needed to train notable machine learning systems.The x-axis is time and the y-axis is teraflops. They-axis is a log scale, meaning the amount of compute has beengrowing exponentially over time.And this has only been feasible due to the new generation of GPUs that processthese large amounts of data.

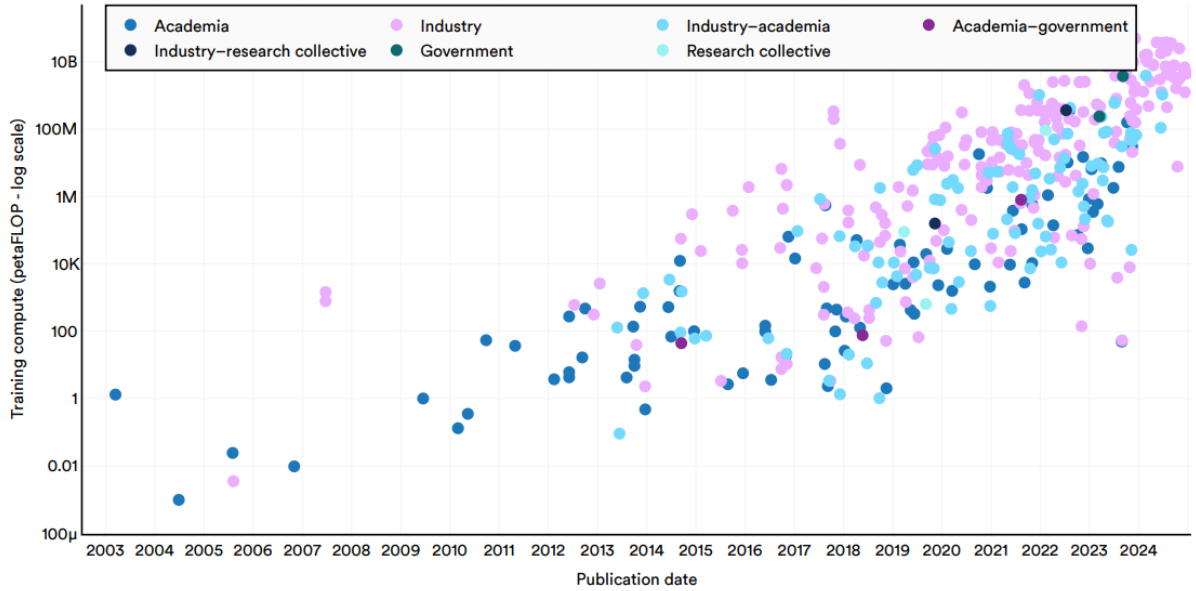


**Figure 2 Training compute of notable AI models by sector, 2003–24[1]**

The chart in Figure 3 is from a research organization called Epoch AI (https://epoch.ai/). It shows the peak computational performance of common ML accelerators at a given precision. Time is on the x-axis and flops is on the y-axis. It may be noted that GPU performance is improving, especially in machine learning specific GPUs. The chart shows trendlines for number formats with eight or more accelerators: FP32, FP16 (FP = floating-point, tensor-* = processed by a tensor core, TF = Nvidia tensor floating-point, INT = integer) compute performance for various GPUs over time.


**Figure 3 Peak computational performance of common ML accelerators at a given precision [2]**

The second ingredient that has enabled recent progress on AI is data.Epoch AI reports that models are using more and more training data across all domains of ML. In language modeling, datasets are growing at a rate of 3.5 times per year (Figure 4). The largest models currently use datasets with tens of trillions of words. The largest public datasets are about ten times larger than this, for example Common Crawl contains hundreds of trillions of words before filtering.

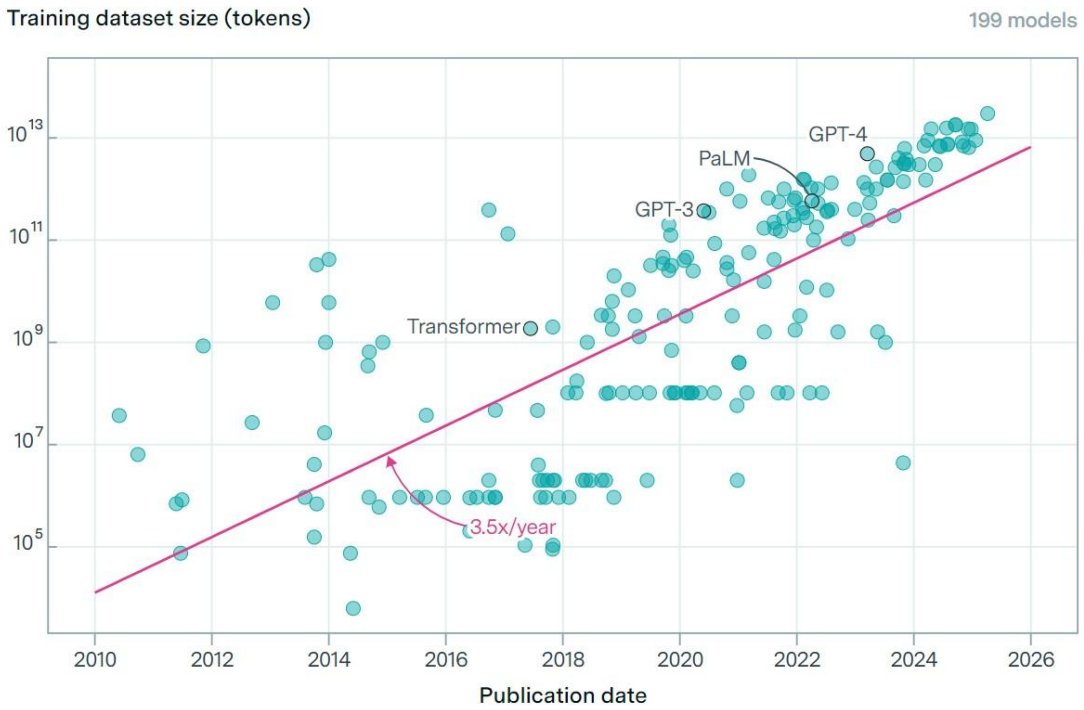


**Figure 4 Size of datasets used to train language models [1]**

The third ingredient that has enabled recent progress on AI is algorithms.In recent years, the performance of AI algorithms has been improving. An AI benchmark is a standardized test used to evaluate the performance and capabilities of AI systems on specific tasks. For example, ImageNet is a canonical AI benchmark that featuresa large collection of labeled images, and AI systems are tasked with classifying these images accurately. Tracking progress on benchmarks has been a standard way for the AI community tomonitor the advancement of AI systems. Figure 5 illustrates the progress of AI systems relative to human baselines for eight AI benchmarks corresponding to 11 tasks (e.g., image classification or basic-level reading comprehension). As of 2024, there are very few task categories where human ability surpasses AI. Even in these areas, the performance gap between AI and humans is shrinking rapidly. For example, on MATH, a benchmark for competition-level mathematics, state-of-the-art AI systems are now 7.9 percentage points ahead of human performance, a significant improvement from the 0.3-point gap in 2024.

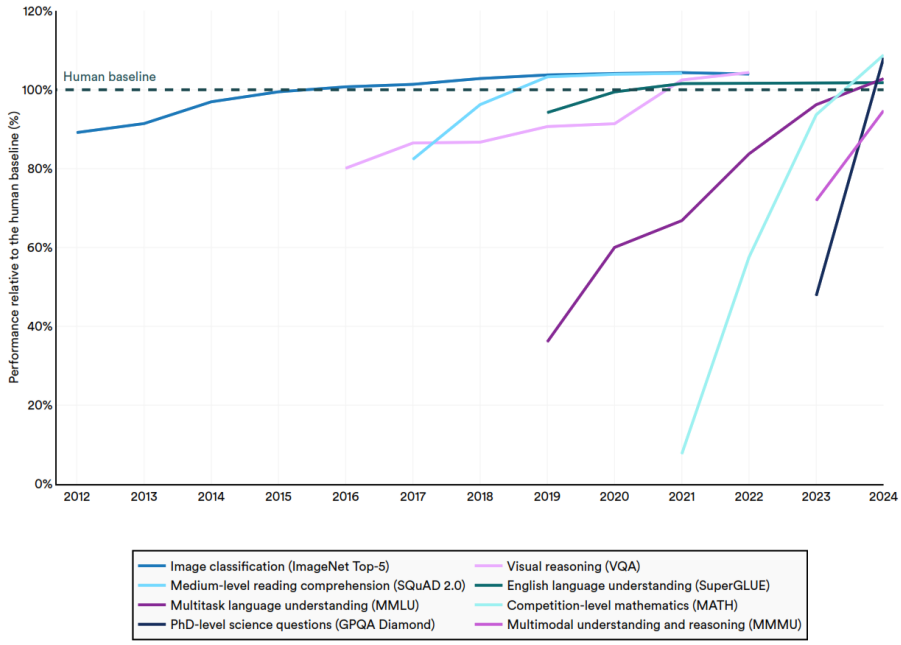


**Figure 5 Select AI Index technical performance benchmarks vs. human performance [1]**

**1.4 AI Benchmarks**

Historically, most AI benchmarks have focused on performance and accuracy. However, as these tools start to have a larger and larger impact on society, there have been new benchmarks developed and calls to develop even more to focus on aspects like toxicity, bias, and even energy efficiency. This section highlights a few such benchmarks.

Stanford researchers introduced HELM, Holistic Evaluation of Language Models in 2022. It is designed to evaluate LLMs across diverse scenarios, such as readingcomprehension, language understanding, mathematical reasoning, and more.They look at this versus diverse metrics, such as accuracy, collaboration,fairness, bias,toxicity.HELM assessed 142 models from severalleading companies and used what they call a min-win rate to track averageperformance across all scenarios.The team also developed HIME, Holistic Evaluation of Image Models.

MMMU is the Massive Multidiscipline Multimodal Understanding Benchmark.This benchmark comprises of about 11,000 college-level questions from six coredisciplines, art, business, science, health, medicine, social science,technology, and others.In addition to text, the question formats include charts, maps, tables, and chemical structures.MMMU is one of the most demanding tests of AI to date.GPT-4.0 won with a score of 69.1%.

MLPerf was developed by MLCommons, a consortium of AI leaders from academia,research labs, and industry.Within MLPerf, there are various benchmarks: ML Perf Training, ML Perf Inference, AI Safety, ML Perf Storage, etc.In the training challenge, participants train ML systems to execute varioustasks using common architecture.Tasks include image recognition, recommendation, natural language processing,and more. Entrantsare ranked on their absolute wall clock time.The inference challenge measures how fast a trained AI system can processinputs and produce outputs, across similar tasks, image recognition, recommendation, natural language processing. TheMLPerf inference includes power measurement.These tools and techniques complement the performance benchmarksenable reporting and comparing energy consumption,performance, and power for submitting systems. Similarly, HULKis a multitask energy efficiency benchmark for natural languageprocessing used to evaluate energy efficiency based on the time and the cost inpre-training, fine-tuning, and inference.

**1.5 Applications of AI**

The previous section discussed the progress AI has made in the past decade due to larger datasets, better performing GPUs, and better performing algorithms. This section will explore cutting-edge applications of this emerging technology across various fields such as healthcare, robotics, education, and sustainability. The aim is to provide essential context on the technology's potential to better society.

Two significant applications of AI in healthcare are in the field of medical imaging and the field of biological discovery. AI algorithms, particularly deep learning models, have shown remarkable accuracy in analyzing medical images such as X-rays, CT scans, and MRIs [3]. These AI systems can detect abnormalities, assist in early diagnosis of diseases like cancer, and even predict patient outcomes. AI is accelerating the drug discovery process by analyzing vast amounts ofbiological and chemical data to identify potential drug candidates [4].Machine learning algorithms can predict how different compounds might interactwith specific proteins or disease targets, reducing the time and cost ofdeveloping new medications [5].

A notable example, Google DeepMind made history when its co-founder and CEO, Demis Hassabis, and research director John Jumper were awarded the Nobel Prize in Chemistry in 2024 for their groundbreaking work on AlphaFold, an AI system that predicts the 3D structure of proteins from their amino acid sequences. AlphaFold addressed a 50-year-old grand challenge in biology: accurately determining protein structures, which are crucial for understanding biological processes and developing new medications. By leveraging deep learning techniques, AlphaFold achieved unprecedented accuracy in predicting protein folding, significantly accelerating research in fields like drug discovery and disease understanding.

In education, AI has the potential to enhance the educational experience byproviding more personalized, accessible, and effective learning opportunities.There have been many studies looking at the positive impact of a good personaltutor on student learning.However, not every classroom or family has the resources to obtain such aperson. AI-based tutoring systems can provide students with one-on-one guidance andsupport, simulating the experience of working with a human tutor. They can tailor educational content and pacing to individualstudent needs, by analyzing student performance data, learning patterns, and preferences tocreate customized learning paths.These systems use natural language processing and machine learning tounderstand student queries, provide explanations, and offer targeted feedback. This approach can help address the diverse learning needs of students andpotentially improve engagement.They can also extend access to high-quality tutoring and provide additionalsupport outside of classroom hours.Similarly, these tools could be used to augment teachers in creating a morepersonalized experience for their students.

AI has significantlyprogressed in the field of robotics.Two examples that stand out for their transformative potential are roboticmanipulation through reinforcement learning and robotic-assisted surgery.Robotic manipulation involves enabling robots to interact with objects in theirenvironment, a task that is inherently challenging due to the complexity ofphysics, object variability, and the need for precision. Forexample, it is very difficult for a robot to know how much force toapply when picking up a blueberry versus a ball.Recent advancements in AI, particularly in reinforcement learning, haveempowered robots to learn and refine manipulation tasks through trial anderror, leading to more adaptable and generalizable skills.Although, the field of robotics tends to have less data, many are working to changethis.AI-powered surgical robots assist surgeons in performing complex procedureswith greater accuracy than human hands alone can achieve, especially inminimally invasive surgeries.There is also great promise for robotics to bring certain healthcare proceduresto locations that do not have physicians with the expertise.So, physicians in locations far away could control a robot assisted by AI inrural locations.

AI is playing a crucial role in combating climate change by enabling moreefficient resource management, better predictions, and innovative solutions toreduce greenhouse gas emissions.Two of the most exciting applications of AI in this area are climate modelingand prediction and renewable energy optimization.Accurate climate modeling is essential for understanding the potential impactsof climate change and developing strategies to mitigate them.AI enhances traditional climate models by analyzing vast amounts of data fromvarious sources, including satellites, weather stations, and ocean buoys tocreate more precise and timely predictions.The transition to renewable energy is crucial for reducing greenhouse gasemissions, and AI plays a vital role in optimizing the production, storage, anddistribution of renewable energy, making it more efficient and reliable. One behavior to reduce greenhouse gas emissions is the electrification of ourvehicles.In this case, the transition to electric vehicles will challenge the electricgrid, for instance, if everyone went home at 5 p.m. to charge their car.Researchers at Stanford developed a probabilistic framework that models EVcharging demand by analyzing drivers' behavior.They develop a use case study in California to present scenarios forelectricity demand in 2030, making this a valuable tool for planners andpolicymakers as they prepare for the growing EV market and energy transition.These innovations underscore the power of AI to revolutionize the world, butthey also bring to light important questions and challenges that mustbe addressed.

### 1.5 Ethical and Societal Implications

As AI is integrated into critical sectors like health care, education, andenvironmental management, it is crucial to consider the ethical, social, andenvironmental implications of these technologies.Multiple studies have revealed significant bias due to non-diverse trainingsets.A study published in 2020 found data used to train health care models primarilycame from just three states in the United States, California, Massachusetts,and New York, and these data sets lacked representation from the broader U.S. population [6].This can have many downstream consequences, one of which is that the model willnot work as well on populations not represented.This can impact health outcomes of these populations.In general, the data sets used to train AI tend to be mostly from westerncountries.

As mentioned earlier, the increasing size of models has created an increasing demand for compute, which is housed in data centers. Figure 6 shows the shows a map of data centers in different countries around the world. Table 1 lists the number of data centers in each country (only for countries having more than 50 data centers). DataCenterMap lists 9649 data centers from 164 countries worldwide.

**Table 1 Number of data centers in different countries (src: datacentermaps.com)**

| Country | Data centers | Country | Data centers | Country | Data centers |
|---|---|---|---|---|---|
| USA | 3680 | Brazil | 171 | Singapore | 74 |
| Germany | 424 | Italy | 169 | South Korea | 73 |
| United Kingdom | 418 | Spain | 156 | Norway | 65 |
| China | 346 | Indonesia | 139 | Chile | 62 |
| France | 264 | Ireland | 121 | Mexico | 59 |
| Canada | 264 | Switzerland | 109 | Romania | 59 |
| India | 262 | Malaysia | 102 | Israel | 56 |
| Australia | 256 | Sweden | 99 | Denmark | 55 |
| The Netherlands | 192 | Hong Kong | 93 | New Zealand | 54 |
| Japan | 182 | Poland | 84 | United Arab Emirates | 54 |
| Russia | 173 | Turkey | 82 | Finland | 50 |

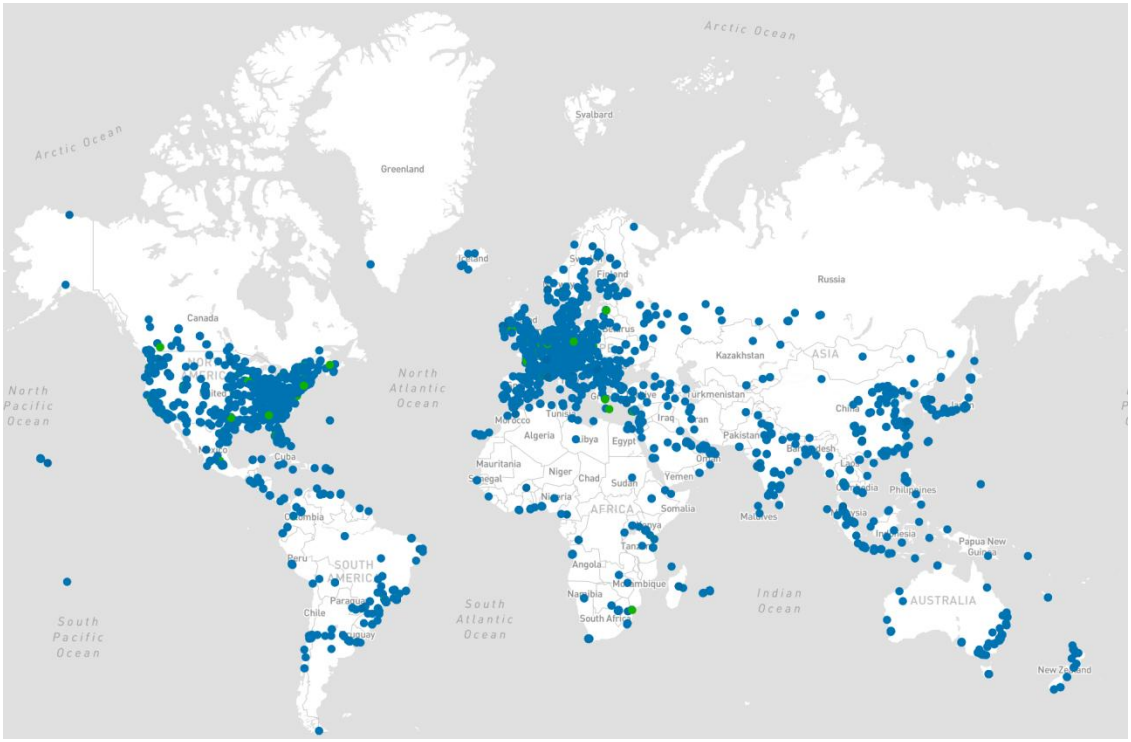


**Figure 6Data centers in different countries (src: datacentermaps.com)**

These data centers require significant resources torun:land, energy, and water for cooling.This raises concerns, especially in locations where these resources arelimited.Google's data centers worldwide consumed nearly 6 billion gallons (22.7 billion liters) of water in 2024 [7]. The company's '2024 Environmental Report' showed an 8% annual increase in water consumption, driven by advancements in search functions, AI, and other projects. AI remains the primary factor behind the surge, with Google's water consumption having jumped 20% in 2022. For comparison, Türkiye consumed 16.6 billion gallons (63 billion liters) of water in 2022, according to the Turkish Ministry of Environment, Urbanization, and Climate Change. Google's data centers alone accounted for nearly one-third of that total [8].TheInternational Energy Agency forecasts that global data centerelectricity demands will more than double from 2022 to 2026, with AI playing amajor role in that increase [9].

The opportunities for AI to better society, as discussed in this section, areimmense.In2022, Google DeepMind released the results ofexperiments inwhich it trained a reinforcement learning agent called BCOOLER to optimizecooling procedures for Google's data centers.At the end of one particular three-month experiment, BCOOLER achieved roughly12.7% energy savings [10]. Figure 7 shows the energy savings results over time for one of the live experiments. Each point on the curve represents the cumulative normalized savings since the beginning of the experiment.


**Figure 7 Energy savings results over time for one of the live experiments of DeepMind's BCOOLER [10]**

## II. BACKGROUND AND PRIMER ON ENERGY & CARBON ACCOUNTING

Thissection discusses measuring and reportingthe carbon emissions of AI systems. The carbon emissions from training frontier AI models have steadily increased over time. Figure 8 shows the carbon emissions of selected AI models over the years. While AlexNet's emissions were negligible, GPT-3 (released in 2020) reportedly emitted around 588 tons of carbon during training, GPT-4 (2023) emitted 5,184 tons, and Llama 3.1 405B (2024) emitted 8,930 tons. DeepSeek V3, released in 2024, and whose performance is comparable to OpenAI's o1, is estimated to have emissions comparable to the GPT-3, released five years ago.

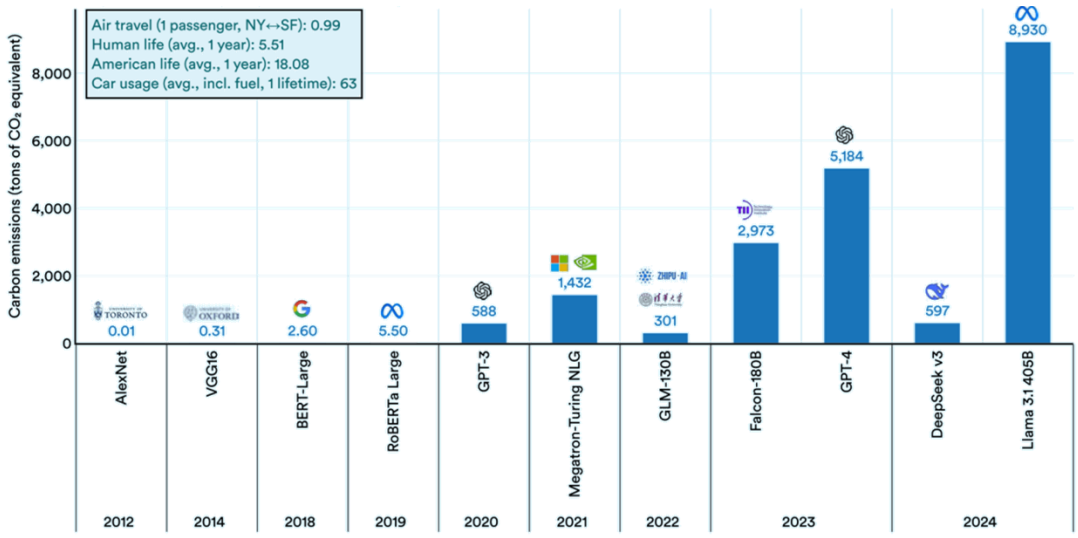


**Figure 8 Estimated carbon emissions from training select AI models and real-life activities, 2012–24 [1]**

This section introduces the terms and concepts required as a primer for energy accounting.In recent years, many cloud providers have set ambitious sustainability targetsand are working towards carbon neutrality by offsetting their emissions withRenewable Energy Certificates (RECs).Achieving carbon neutrality means an organization has a net zero carbonfootprint.Each purchased REC certifies that one megawatt of renewable energy has beenadded to the grid by the organization, offsetting an equivalent amount of non-renewable energy.

Google claims that it became the first major company to become carbon neutral, in 2007. And in 2017, it became the first company to match 100% of its electricity consumption with renewable energy [11]. Google Cloud purchases enough renewable energy, i.e. wind and solar power to match its data centers' electricity consumption. Since the power is averaged annually, a particular data center, at any given time, may have too much renewable power, or too little. Google feeds the extra power into the local grid and draws power from the local grid when renewable generation is lacking. Google aims to run its business on carbon-free energy everywhere, at all times, by 2030.

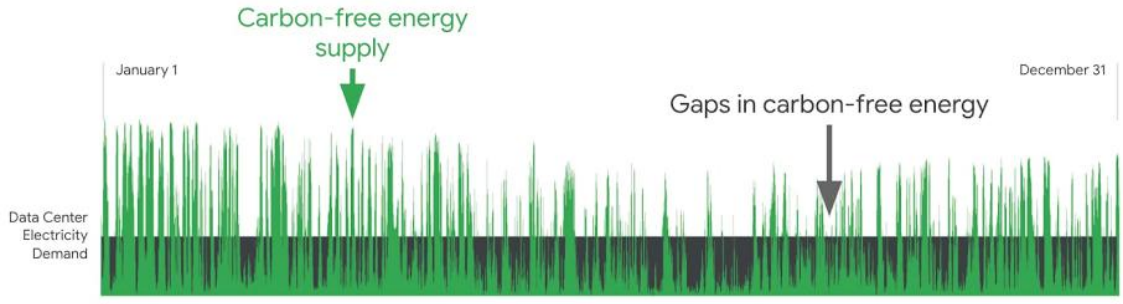


**Figure 9 Hourly carbon-free performance of a Google data center [11]**

Energy for data centers in many locations are not currently derived from carbon-neutral sources, and whenrenewable energy is available, it is still limited tothe equipment available to produce and store it.The energy consumption of a system can be measured in joules or watt-hours.A watt is a unit of power, and one watt is equivalent to one joule per secondand represents the amount of energy needed to power the system.Lifecycle accounting refers to all stages of the product lifecycle.In the case of AI models, this would include activities like materialextraction and manufacturing to end-of-life disposal.However, it is currently quite difficult to attribute

hardware manufacturingand disposal on a per-equipment basis. Most researchers focus on the energyconsumption of model training and deployment.

Pre-training an AI model refers to the steps where a whole bunch ofgeneral data is fed into the model.Fine-tuning is when the model is further trained on a smaller set of context-specific data, and inference is when the model processes inputs and producesoutputs.Energy impacts of pre-training and fine-tuning are generally accounted for inthe model training step, and inference is generally accounted for in thedeployment step.

When measuring energy used during training and deployment, the focusis mainly onthe data center.This includes cooling, lighting, power conversion, networkhardware, and storage. Storage can include aspects like CPUs and DRAM.DRAM, dynamic random access memory, is a type of memory that is typically usedfor the data or program code needed by the computer to function.

Another important term is power usage effectiveness, or power unitefficiency (PUE).This is a ratio of Total Facility Energy (all energy used in the data center including lighting, cooling, power systems, etc.) to IT Equipment Energy (energy used only by computing equipment such as servers, storage, networking devices). PUE describes how efficiently a computer data center uses itsenergy.An ideal PUE is one.An accurate accounting for all of the components of this ratio requires complex modelingand varies depending on workload or utilization of the CPUs and GPUs (which is usually reported as a percentage).

Carbon accounting at project scale can be defined as measuring thevaluation of carbon and green house gas (GHG) emissions and offsetting from projects.This assessment informs project owners and investors and establish standardizedmethodologies.Carbon and GHG emissions are typically measured in $CO_2$ equivalents.This is the amount of carbon and other GHG converted to carbon amounts releasedinto the atmosphere as a result of the project. Carbon offsetting is when organizations invest in green initiatives thatbalance out the carbon emissions as a result of the project. Forexample, an organization might support solar or wind energy, whichproduce more renewable energy than the energy needed to power their datacenters and used to train their AI models. U.S. Environmental Protection Agency defines the social cost of carbon ($SCCO_2$)as the measure in dollars of the long-term damage done by a ton of carbon dioxide emissions in a given year.This dollar figure also represents the value of damage avoided for a smallemission reduction.

Carbon emissions for a compute system can be estimated by understanding thecarbon intensity of the local energy grid and the energy consumption of thesystem.Any given energy grid will have a carbon intensity, the grams of $CO_2$ equivalentemitted per kilowatt hour of energy used.This carbon intensity is determined based on the energy sources supplying thegrid, and each energy source has its own carbon intensity.

Electricity Maps [12] is a powerful tool for understanding and acting on electricity carbon footprints. It provides a live 24/7 visualization of where electricity comes from and how much $CO_2$ equivalent is emitted during generation. In addition to current emissions data and historical trends, it provides forecasts up to 72 hours ahead, covering over 200 zones worldwide, with granular data on electricity mix, carbon intensity, and prices. The tool also offers a developer API and an open-source codebase (GitHub licensed), enabling integration into apps and dashboards. Web and mobile app display intuitive maps, hourly $CO_2$ readings, breakdowns of generation sources, emissions, and cross-border flows make Electricity Maps a significant tool in the research for green energy.

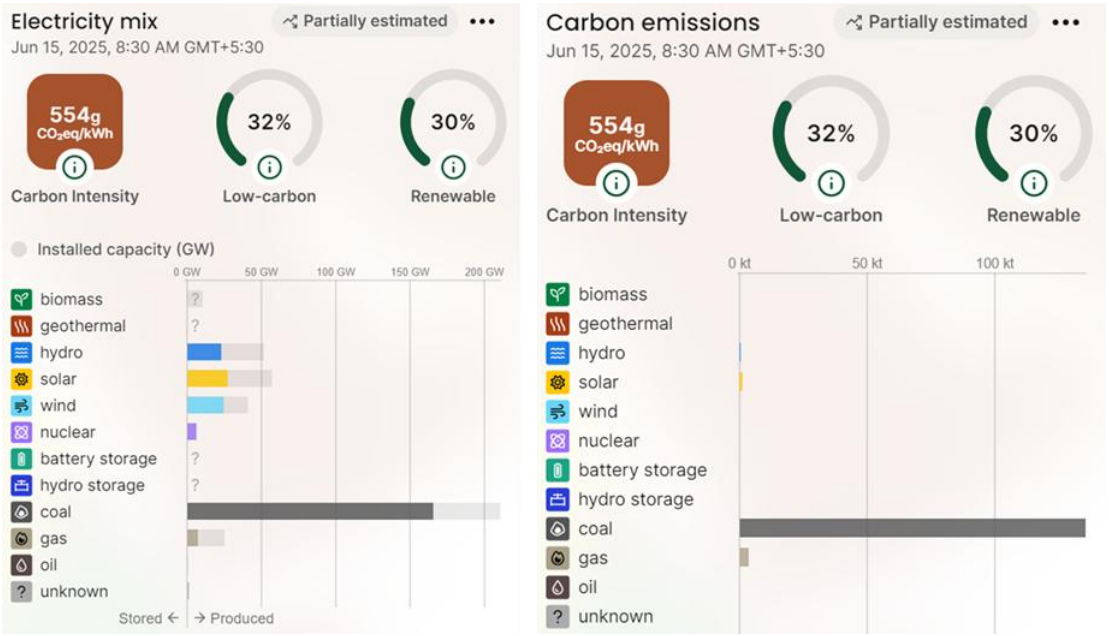


**Figure 10 Live Electricity Mix and Carbon Emissions for India on July 15, 2025 (src: https://app.electricitymaps.com/zone/IN/72h/hourly)**

### 2.1 AI Carbon Accounting Methodologies

This section discusses two methodsused to estimate the carbon emissions of AI models.

Strubell et al. [13] quantify the approximate computational, financial andenvironmental costs of training a variety of recently successful neural network models forNLP. Authors performed an analysis of the energy required to train a variety of NLP models such as transformer, ELMo (Embeddings from Language Models), BERT (Bidirectional Encoder Representations from Transformers), GPT-2 (Generative Pre-trained Transformer). Themodelswere trained using the default settingsprovided in table 2.

**Table 2 Computational requirements of popular NLP models**

| Model | Parameters | Hardware Used | Training Time | Notes |
|---|---|---|---|---|
| **Transformer (T2Tbase)** | 65M | 8× NVIDIA P100 GPUs | 12 hours | 300k steps reported |
| **Transformer (T2Tbig)** | 213M | 8× NVIDIA P100 GPUs | 84 hours (3.5 days) | 300k steps |
| **NAS-Transformer** | — | 1× TPUv2 core | 10 hours (300k steps) | Full NAS search: 979M steps = 32,623 TPU hours or 274,120 GPU hours |
| **ELMo** | — | 3× NVIDIA GTX 1080 GPUs | 336 hours (14 days) | Based on stacked LSTMs |
| **BERT (Base)** | 110M | 16× TPU chips | 96 hours (4 days) | Devlin et al. (2019) |
| **BERT (Base)** | 110M | 64× V100 GPUs (4 DGX-2H) | 79.2 hours (3.3 days) | NVIDIA implementation |
| **GPT-2 (Large)** | 1542M | 32× TPUv3 chips | 168 hours (7 days) | Trained on massive unsupervised data |

Each model was trainedfor a maximum of 1 dayona NVIDIA Titan X GPU, except forELMo which was trained on 3 NVIDIAGTX 1080 Ti GPUs, sampling the GPU power consumption repeatedly. CPU power consumption was sampled using Intel's RunningAverage Power Limit interface. The total time expected for models to train to completion was estimated using training times andhardware reported in the original papers of the models. Calculation of the power consumption in kilowatt-hours(kWh) was done as follows.

$$p_t = 1.58t \times \left(p_c + p_r + p_g\right) \div 1000$$

Where,

$p_c$:average powerdraw (in watts) from all CPU sockets during training

$p_r$:average power draw from allDRAM (main memory) sockets

$p_g$: average power draw of a GPU during training

$g$:number of GPUs used to train

$t$ : time taken to train the model

PUE coefficient: 1.58 (2018 global average for data centers)

Thus, the total power consumption as combined GPU,CPU and DRAM consumption is multiplied by PUE, which accounts for the additional energy required to support the compute infrastructure (mainly cooling).
Authors provide the following formula to convert power toestimated CO2 emissions:

$$CO_2e = 0.954\ p_t$$

This conversion considers the relative proportions of different energy sources (primarily natural gas, coal, nuclear and renewable) consumedto produce energy in the United States. Authors believe that the U.S. breakdown of energy provides a reasonable estimate of $CO_2$ emissions per kilowatt hour of compute energy used as it is comparable to that of the mostpopular cloud compute service, Amazon Web Services.
The second methodology [14] builds a framework to track and report the environmental impact of machine learning experiments that aims to promote transparency and accountability in research by encouraging authors to document energy usage and emissions alongside traditional performance metrics.

At the core of the framework is the Experiment Impact Tracker, a lightweight, Python-based tool that logs system-level metrics during machine learning training. With just a few lines of code, it begins collecting information such as CPU and GPU usage, power draw, memory usage, disk activity, and training duration. The tracker identifies the energy grid region of the machine running the experiment (via IP address) and links it with regional carbon intensity data, which is used to estimate the experiment's emissions. It can even track real-time carbon intensity in California by polling data from CAISO, highlighting how carbon output can vary depending on time of day and grid energy sources. The framework employs a fair accounting method to assign energy consumption to individual experiments, especially in shared environments. It does this by monitoring per-process resource utilization and calculating total energy using a PUE factor to account for infrastructure energy overhead. For example, if a training process uses 25% of the total CPU time during an experiment, it is credited with 25% of the CPU energy consumption. This ensures accurate measurement even when multiple jobs run concurrently on the same machine. The resulting energy usage is then converted into carbon emissions using the following formula.

$$e_{total} = PUE \sum_p \left(p_{dram} e_{dram} + p_{cpu} e_{cpu} + p_{gpu} e_{gpu}\right)$$

Where,
$p_{resource}$: percentage of system resource used by the attributable processes relative to the total in-use resources
$e_{resource}$: energy usage of that resource
To support reproducible and accessible reporting, the framework includes a script that automatically generates HTML appendices with graphs, tables, and summaries of energy and carbon metrics. These appendices can be published alongside research papers, making it easier for readers to understand the environmental costs of model training. The broader goal of the framework is to foster a culture of sustainable AI research, where energy efficiency and carbon awareness become standard parts of model evaluation and scientific communication.

## III. REPORTING CARBON EMISSIONS

The previous section discussed mechanisms to report carbon impact and this section discusses some tools used. TheCodeCarbon tool (https://codecarbon.io/)considers the location of the datacenter of the cloud provider and the $CO_2$ emissions of the power grid at thatlocation. This information is generally available on one of the websites of the cloudproviders, such as AWS, Google, and Azure.If the carbon intensity in a certain location is not available, the CodeCarbonteam uses the electricity mix as a weighted average with the carbon intensityof each component. Forexample, if the energy mix for Germany is 9% biomass, 20% coal, 3% wind,56% solar, etc., the carbon intensity of those mechanisms can be obtained based on the data from the website electricitymaps.com and an energy mix can be derived. Ifthese mixes are not available, theaverage of 475 grams of $CO_2$equivalents per kilowatt hour would be applied.CodeCarbon uses a scheduler that by default calls for a measure every 15seconds, which the user can adjust. Likeother methodologies, power usage from GPU, CPU, and RAM are tracked. Foreach experiment, output is provided through a CSV or a webapplication. Users can see the net power consumption and carbon equivalents for the projectand comparisons to everyday activities, like driving a car or watching TV.

The Microsoft Emissions Impact Dashboard is a Power BI–based tool designed to help Azure and Microsoft 365 customers monitor, analyze, and report the greenhouse gas emissions associated with their cloud usage Originally launched in 2020 as the 'Microsoft Sustainability Calculator', the solution was relaunched as a fully featured dashboard by late 2021. The Emissions Impact Dashboard provides transparency into greenhouse gas emissions associated with using Microsoft cloud services and enables a better understanding of the root causes of emissions changes. Organizations can measure the impact of Microsoft cloud usage on their carbon footprint, and they can drill down into emissions by month, service, and datacenter region. The tool also enables customers to enter un-migrated workloads and get an estimate of emissions savings from migrating to the Microsoft cloud. Newly added data protection allows Emissions Impact Dashboard administrators within an organization to control who can see their company data in the tool.

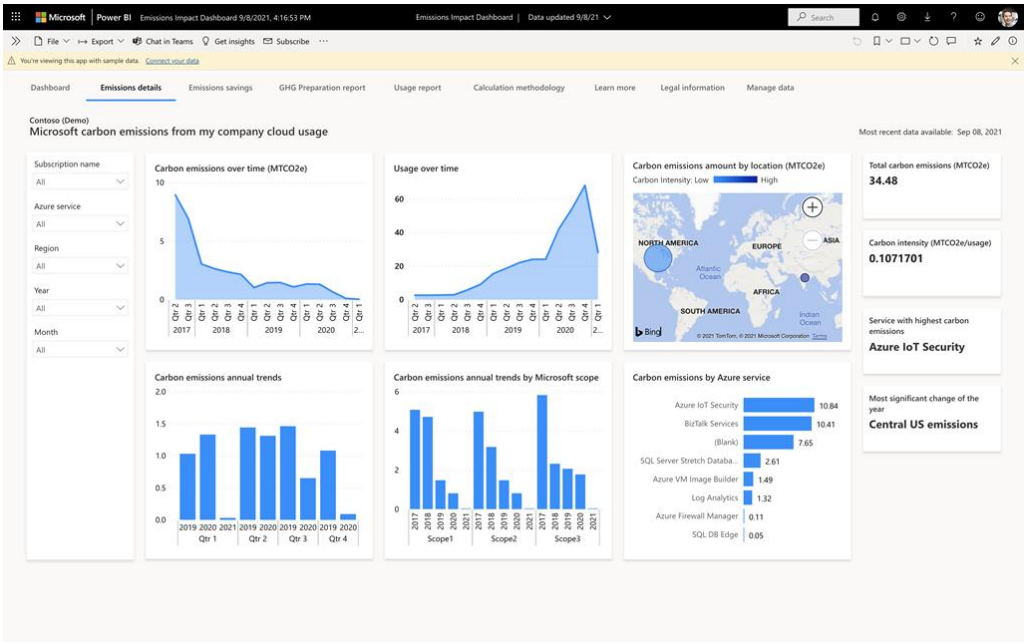


**Figure 11 The main Microsoft Emissions Impact Dashboard focuses on showcasing overall emissions and usage over time, as well as carbon intensity, which is a metric of carbon efficiency specific to cloud usage. (src: https://azure.microsoft.com/)**

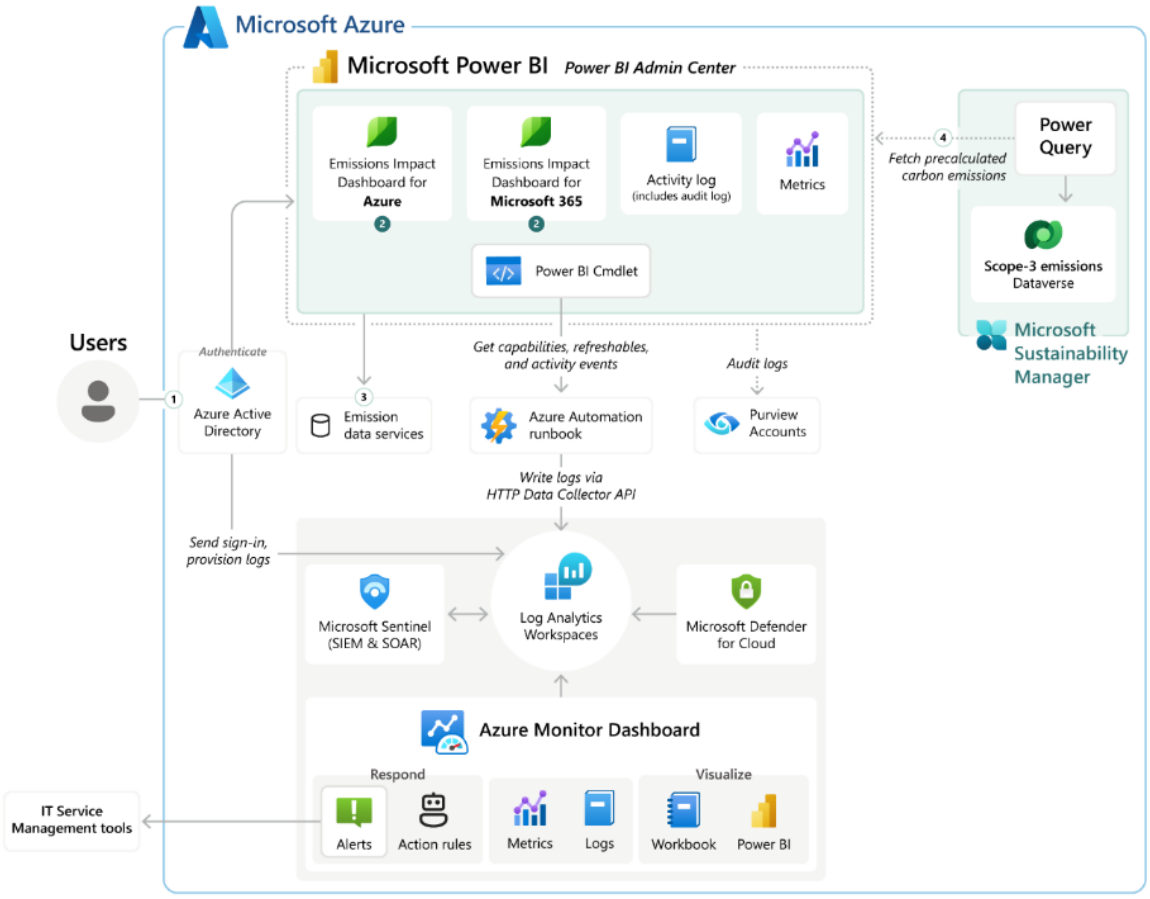


**Figure 12 A custom Azure Automation playbook collects activity logs in Power BI via Power BI SDK. It then transforms and feeds this log into an Azure Log Analytics workspace so that organizations can use Azure Monitor for monitoring, query, and alert purposes. (src: https://learn.microsoft.com/)**

### 3.1 Importance of Carbon Reporting

Many cloud providers are striving forcarbon neutrality by using RECs, to offset theiremissions.

- One REC is issued when one megawatt hour of electricity generated from arenewable energy source is delivered to the grid.
- Users can buy RECs from renewable energy companies to offset their electricityconsumption.For instance, a company might purchase RECs and then sell the associatedrenewable energy back to the grid that supplies their data centers.
- However, this approach supports future renewable energy production rather thanaltering the current energy mix on the grid.
- It is important to note that even if a cloud provider claims carbon neutralityfor their data centers, the actual $CO_2$ emissions can significantly varydepending on the region and even the time of day. Solarenergy cannot be generated at night, for example.
- Forthis reason, most of the methods and trackers do not consider offsetswhen estimating the carbon emissions of AI models.

Most of the research works advocate for the reporting of carbonemissions in the world of machine learning.No methods have been discussed to reduce those emissions.However, the first step in reducing emissions is understanding them.Without consistent and accurate accounting, theimpacts of these models and what consequences they can have on society will be unknown.

In 2020, Hulk, a multitask energy efficiency benchmark for natural languageprocessing, evaluated energy efficiency based on the time and the cost inpre-training, fine-tuning, and inference. In2021, MLPerf, one of the most popular benchmarks to measure training andinference performance for hardware, software, and services, added a systempower measurement to complement performance measurements. Apart from these, there haven't been any broad energybenchmarks discussed within the AI community.Pairing energy and carbon emissions benchmarks directly in addition toperformance or accuracy benchmarks can institute a climate-friendly culturewithin the AI community and spread information about the most energy andclimate-friendly combinations of hardware, software, and algorithms.

Figure 13 shows an evaluation of four baseline reinforcement learning algorithms, namely, Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), A2C+Vtraces, and Deep Q Networks (DQN), in two evaluation environments, PongNoFrameskip-v4 and BreakoutNoFrameskip-v4[14]. The models are trained for only 5M timesteps, less than prior work, to encourage energy efficiency and evaluate for 25 episodes every 250k timesteps. The Average Return is plotted across all evaluations throughout training (giving some measure of both ability and speed of convergence of an algorithm) as compared to the total energy in kWh. Weighted rankings of Average Return per kWh place A2C+Vtrace first on Pong and PPO first on Breakout. Using PPO versus DQN can yield significant energy savings, while retaining performance on both environments(in the 5M samples regime). The experiment shows that while no algorithm is the energy efficiency winner,the light blue dots attain balance between efficiency and performance. Thelight blue dots are the PPO dots, and they're relatively low on the x-axis,power, and high on the y-axis, asymptotic return.

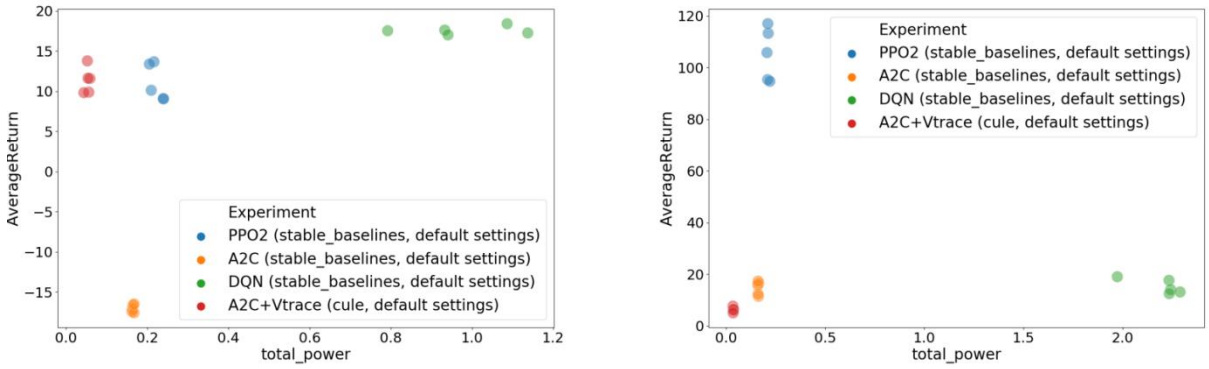


**Figure 13 Average Return versus total power of A2C, PPO, DQN, and A2C+VTraces on PongNoFrameskip-v4 (left) and BreakoutNoFrameskip-v4 (right) [14]**

Accurate reporting of energy metrics also enables cost-benefit analysis thatwould otherwise be impossible.For instance, the estimated revenue generated by a model could be comparedagainst its electricity costs, or the carbon emissions saved by a model couldbe weighed against the emissions it generates.

## IV. CASE STUDY: ENVIRONMENTAL IMPACT OF BLOOM 176B

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) is a 176-billion-parameter autoregressive large language model developed as part of the BigScience Project [15], a collaborative international effort to democratize large-scale AI research. It represents not only a scientific and technical milestone, but also one of the most transparent large-scale model development projects in terms of environmental reporting. Unlike many proprietary models, BLOOM was accompanied by a full lifecycle carbon footprint assessment, which provides a rare and instructive insight into the environmental costs of developing and deploying such models.

### 4.1 Lifecycle Emissions Breakdown

The training phase of BLOOM took place on the Jean Zay supercomputer in France, a system partially powered by nuclear energy, which significantly reduced its carbon intensity compared to fossil-fuel-heavy grids. The dynamic emissions—i.e., the electricity consumed directly during training—were estimated at 24.7 tonnes of $CO_2$ equivalent ($tCO_2eq$). However, a more holistic environmental analysis reveals that training emissions are only one part of the model's overall footprint [16].

When taking into account hardware manufacturing (embodied emissions) and idle energy consumption, the total emissions nearly double. Specifically:

- Idle energy—the electricity used by the infrastructure (such as GPUs and system components) when not actively training—contributed an estimated 14.6 $tCO_2eq$.
- Hardware manufacturing and embodied emissions added another 11.2 $tCO_2eq$, which includes the emissions from producing and shipping high-performance GPUs, CPUs, and servers used during the training.

In total, the end-to-end emissions from training BLOOM are estimated at approximately 50.5 $tCO_2eq$, demonstrating that focusing only on compute-time electricity underestimates the environmental burden by more than 50%.

### 4.2 Inference and Deployment Emissions

The BLOOM team also undertook a post-training evaluation of energy use during inference, which is particularly important since LLMs are often deployed at scale, serving millions of queries. Power measurements were conducted during real-time API calls, showing that per-GPU energy draw ranged from 78W to 171W depending on the load and batch size. Although each individual inference may only consume a small amount of energy, sustained low-utilization (such as idle servers during off-peak hours) can lead to significant emissions over time.

To mitigate this, the team recommended deploying batching strategies, improving resource sharing, and using dynamic instance scaling to reduce the environmental cost of hosting such models. Their findings suggest that optimization during inference may be just as important as reducing training time, especially when the model serves a large user base.

### 4.3 Broader Implications

The BLOOM case study reinforces the argument that AI sustainability efforts must move beyond training emissions. It highlights the importance of including idle power, hardware lifecycle emissions, and geographical energy grid carbon intensities in any serious environmental evaluation. One important insight is that even if a supercomputer is relatively clean (due to low-carbon energy sources like nuclear or hydro), the overall impact can still be high when idle energy and hardware manufacturing are factored in.

Moreover, BLOOM sets a precedent in environmental transparency for the AI community. By publishing detailed energy metrics, carbon accounting methodology, and the breakdown of hardware usage, the BLOOM project provides a model for responsible and open AI development. It also underscores the need for future models to be optimized not only for performance but also for efficiency, carbon-awareness, and sustainability over the entire AI lifecycle—from data curation to long-term deployment.

Figure 14 provides a comparison chart of the estimated carbon emissions for three major AI models—BERT Base, GPT-3, and BLOOM. BERT Base emits very little $CO_2$ (around 0.65 tonnes) during training due to its relatively small size (110M parameters). GPT-3, with 175 billion parameters, has extremely high training emissions (~552 tonnes$CO_2eq$), and this doesn't account for idle energy or hardware emissions. BLOOM, while also a 176B-parameter model, shows significantly lower training emissions (24.7 tonnes)—thanks to the low-carbon energy mix used (nuclear/hydro). However, when idle energy and hardware manufacturing emissions are included, the total rises to 50.5 tonnes, still far below GPT-3.

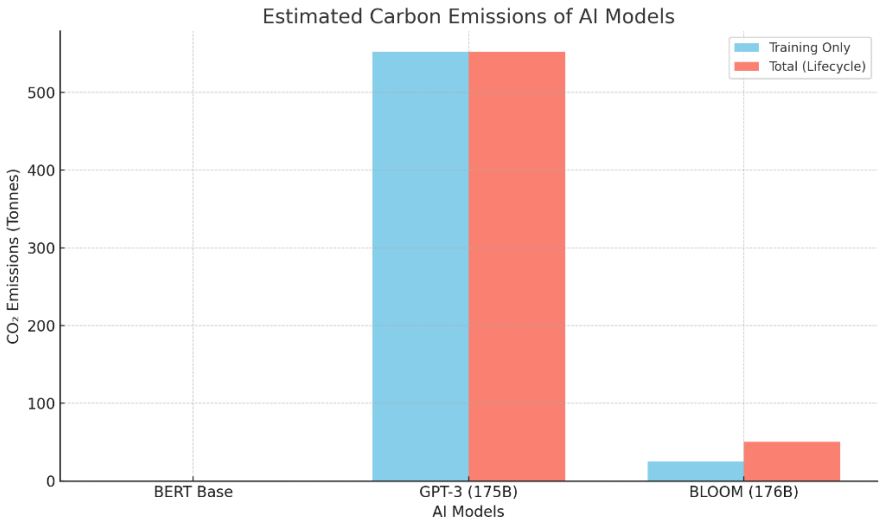


**Figure 14 Estimated carbon emissions for BERT Base, GPT-3, and BLOOM**

## V. RECOMMENDATIONS

The environmental implications of developing and deploying AI systems are becoming increasingly significant as models grow in size, complexity, and usage. Addressing this issue requires a multi-pronged strategy involving improved transparency, conscientious design and deployment practices by developers, and enhanced user awareness. The following recommendations are structured across three key areas:

1. Reporting and Transparency

To foster accountability and enable informed comparisons across models, standardized reporting of environmental metrics should become an integral part of AI research and deployment. Publications and benchmarks should include comprehensive documentation of energy consumption, carbon emissions, hardware specifications, model size, and training duration. Existing frameworks such as the Experiment Impact Tracker and corporate tools like Microsoft's Emissions Impact Dashboard provide viable mechanisms for such reporting. Furthermore, environmental metrics should be extended beyond the training phase to include fine-tuning, inference, and real-world deployment scenarios. The introduction of sustainability-oriented leaderboards that rank models based on carbon efficiency in addition to task performance may further incentivize environmentally responsible innovation.

2. Developer-Led Design and Deployment Decisions

Developers play a crucial role in minimizing the environmental footprint of AI. This can be achieved by prioritizing computationally efficient model architectures and training methodologies. Techniques such as model pruning, knowledge distillation, early stopping, and parameter sharing have shown promise in reducing energy usage while maintaining competitive accuracy. Moreover, selecting cleaner compute options—such as data centers powered by renewable energy or carbon-aware cloud scheduling—can substantially lower associated emissions. Where feasible, AI workloads should be scheduled during periods of lower grid carbon intensity or run on energy-efficient hardware platforms like optimized TPUs or next-generation GPUs. These practices collectively enable a more sustainable AI development lifecycle.

3. User Awareness and Informed Usage

End-users, often unaware of the environmental impact of their interactions with AI systems, should be equipped with the tools and information necessary to make more sustainable choices. One approach involves embedding energy or carbon intensity indicators into user interfaces, particularly for high-consumption features such as generative AI queries or continuous model-driven recommendations. Additionally, application-level options such as “eco-mode” or “low-impact inference” could be offered to reduce emissions in non-critical use cases. Educational outreach—including public campaigns, documentation, and curricular integration—can also play a vital role in raising awareness and fostering responsible behavior among users and developers alike.

These combined strategies—grounded in transparency, thoughtful design, and collective awareness—can contribute meaningfully to the sustainable advancement of AI technologies. Future research should further explore and validate low-carbon methodologies, while policy and industry standards can help institutionalize these practices.

## VI. CONCLUSION

As artificial intelligence systems continue to scale in complexity and pervasiveness, their environmental impact becomes an increasingly urgent concern. From the energy-intensive process of training large language models to the cumulative emissions generated during deployment and inference at scale, AI systems contribute meaningfully to global carbon emissions. These impacts are often obscured by a lack of standardized reporting and insufficient awareness among both developers and users.

Addressing these challenges requires a holistic approach that integrates transparency, efficiency, and accountability. Reporting mechanisms must be standardized to include energy use, carbon footprint, and hardware details, enabling more informed comparisons and sustainable design choices. Developers must adopt energy-efficient architectures, leverage cleaner computational infrastructure, and make thoughtful decisions about model size and training frequency. At the same time, users should be made aware of the environmental cost of AI-driven features and empowered to make lower-impact choices where possible.

Sustainability must become a foundational consideration in the development and deployment of AI, rather than an afterthought. Through collaborative efforts between researchers, industry stakeholders, policymakers, and the broader public, we can guide AI innovation toward a future that is not only intelligent and impactful—but also environmentally responsible.

## REFERENCES


[1] N. Maslej, "The AI Index 2025 Annual Report," Stanford University, 2025.

[2] M. Hobbhahn, L. Heim and G. Aydos, "Trends in Machine Learning Hardware," 9 November 2023. [Online]. Available: https://epoch.ai/blog/trends-in-machine-learning-hardware. [Accessed 13 June 2025].

[3] L. Guo, C. Zhou, J. Xu, C. Huan, Y. Yu and G. Lu, "Deep Learning for Chest X-ray Diagnosis: Competition Between Radiologists with or Without Artificial Intelligence Assistance," *Journal of Imaging Informatics in Medicine,* vol. 37, pp. 922-934, 2024.

[4] M. K. G. Abbas, A. Rassam, F. Karamshahi, R. Abunora and M. Abouseada, "The Role of AI in Drug Discovery," *ChemBioChem,* vol. 25, no. 14, 2024.

[5] Y. Zhang, Y. Wang and C. Wu, *Drug-target interaction prediction by integrating heterogeneous information with mutual attention network,* arXiv, 2024.

[6] A. Kaushal, R. Altman and C. Langlotz, "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms," *JAMA,* vol. 324, no. 12, pp. 1212-1213, 2020.

[7] Google, "2024 Environmental Report," Juky 2024. [Online]. Available: https://sustainability.google/reports/google-2024-environmental-report/. [Accessed 1 May 2025].

[8] A. Günyol, "Google data centers used nearly 6B gallons of water in 2024," Anadolu Ajansi, 2025.

[9] International Energy Agency, "AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works," International Energy Agency, 2025.

[10] J. Luo, C. Paduraru and O. Voicu, "Controlling Commercial Cooling Systems Using Reinforcement Learning," DeepMind, 2022.

[11] U. Hölzle, "Announcing 'round-the-clock clean energy for cloud," 14 September 2020. [Online]. Available: https://cloud.google.com/blog/topics/inside-google-cloud/announcing-round-the-clock-clean-energy-for-cloud. [Accessed 14 June 2025].

[12] Tomorrow, Electricity Maps ApS, [Online]. Available: https://www.electricitymaps.com/. [Accessed 15 June 2025].

[13] E. Strubell, A. Ganesh and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, 2019.

[14] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky and J. Pineau, "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning," *Journal of Machine Learning Research,* vol. 21, no. 248, pp. 1-43, 2020.

[15] BigScience, "BigScience Large Open-science Open-access Multilingual Language Model," 6 July 2022. [Online]. Available: https://huggingface.co/bigscience/bloom. [Accessed 21 June 2025].

[16] A. S. Luccioni, S. Viguier and A.-L. Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," *Journal of Machine Learning Research ,* vol. 24, pp. 1-15, 2023.