

Carbon Footprint of Training Large AI Models: Measuring and Proposing Mitigation Strategies for the Environmental Cost of AI

Meet Dasalaniya ¹

Aastha Jani ²

Vivek Gajjar ³

Urmi Gosaliya ⁴

¹ Computer Engineering Department, Gandhinagar University, Gandhinagar, Gujarat, India.

² Computer Engineering Department, Gandhinagar University, Gandhinagar, Gujarat, India.

³ Computer Engineering Department, Gandhinagar University, Gandhinagar, Gujarat, India.

⁴ Computer Engineering Department, Gandhinagar University, Gandhinagar, Gujarat, India.

Abstract:

Over the past decade, artificial intelligence has quietly shifted from a research curiosity to infrastructure that billions of people depend on every day. Large language models write code, assist doctors, translate languages, and hold conversations in ways that once seemed firmly out of reach. But behind every chatbot interaction and AI-assisted search result, there is a data center consuming electricity and in most parts of the world, that electricity still comes largely from fossil fuels. The carbon cost of building and running these systems is real, growing, and almost entirely invisible to the people using them.

This paper investigates that hidden cost in depth. We examine how energy demands have grown alongside model size from BERT's modest 652 kg of CO₂ in 2018 to independent estimates placing GPT-4's training footprint in the thousands of tonnes. We survey the measurement tools researchers use to quantify these emissions and expose the significant gaps that remain. We then present a comprehensive set of mitigation strategies spanning algorithmic efficiency, hardware improvements, geographic scheduling, renewable energy, and transparency standards. Finally, we honestly confront the structural barriers competitive scaling incentives, the Jevons Paradox, and the near-total absence of embodied carbon from current estimates that make progress harder than the technical solutions alone would suggest. The central argument is straightforward: sustainable AI is technically achievable. What it requires is the collective will to treat it as a genuine priority.

Keywords: Artificial Intelligence, Carbon Footprint, Large Language Models, Energy Consumption, Green AI, Sustainable Computing, CO₂ Emissions, Machine Learning, Inference Efficiency, Embodied Carbon.

1. Introduction

1.1. Background and Motivation:

There is a version of the AI story almost everyone has heard: a technology once confined to university labs has, in just a few years, become woven into daily life. Models like GPT-4, Google Gemini, Anthropic's Claude, and Meta's LLaMA now power tools used by hundreds of millions of people writing assistants, code editors, medical information services, language translation engines, and much more. The pace of this transformation has been remarkable, and it shows no sign of slowing.

But there is another version of the AI story that gets far less attention: the version about what it costs to build these systems. Training a frontier language model is not a task that runs on a laptop overnight. It is a months-long industrial process that burns through electricity at a scale comparable to small towns, running across clusters of thousands of specialized chips in data centers around the world. When that electricity comes from fossil fuels which it still largely does in most regions the result is carbon dioxide entering the atmosphere.

The scale of this was brought into sharp focus by a 2019 study from Emma Strubell and colleagues at the University of Massachusetts Amherst. They calculated that training a single large NLP model could emit as much CO₂ as five average American cars over their entire operational lifetimes roughly 284 tonnes. That finding caught the tech world's attention. And since then, the models being trained have only grown dramatically larger.

1.2. Study Objectives:

This paper sets out to accomplish four interconnected things:

- Measure and contextualize the energy consumption and carbon emissions associated with training state-of-the-art AI models, using documented case studies wherever possible.
- Survey and evaluate the tools and methodologies researchers currently use to track AI's environmental footprint, identifying both their strengths and their significant gaps.
- Propose a concrete, prioritized set of evidence-based mitigation strategies spanning algorithmic design, infrastructure choices, energy sourcing, and policy.
- Honestly confront the structural and economic barriers that make progress toward sustainable AI harder than the technical solutions alone would suggest.

1.3. Scope and Limitations:

A few boundaries are worth establishing upfront. This paper focuses primarily on the training phase of large language models and other large-scale deep learning systems, though inference costs are addressed in detail in Section 4. Exact energy and emissions data for frontier model training is rarely disclosed by the companies conducting it; where figures are disputed or uncertain, this is stated clearly. We also acknowledge that embodied carbon the emissions from manufacturing hardware is flagged as a critical blind spot rather than fully quantified, given the absence of reliable published data.

1.4. Paper Organization:

Section 2 explains how large AI models are trained and why that process is so energy-intensive, including the hardware involved and the often-overlooked costs beyond the main training run. Section 3 covers how researchers measure carbon footprints, including the standard formula and tools, alongside the challenges that limit accuracy. Section 4 presents documented case studies and comparative emissions data, including a discussion of inference costs that are frequently underweighted. Section 5 lays out a comprehensive set of mitigation strategies. Section 6 addresses structural challenges that slow adoption. Section 7 concludes with a synthesis of findings and a call to action.

2. Background: How AI Models Are Trained and Why It Costs So Much

2.1. The Training Process:

At its core, training an AI model is a process of learning from data through repetition. For a language model, that means feeding it billions of words and asking it, over and over, to predict what comes next then adjusting its internal settings (called “parameters” or “weights”) each time it gets something wrong. This process called stochastic gradient descent requires computing a “forward pass” to generate a prediction and a “backward pass” to calculate how to adjust every weight. Each step involves millions of matrix multiplications. Modern frontier models can have hundreds of billions or even trillions of parameters, and training requires running this cycle billions of times over enormous datasets.

This process runs on specialized hardware primarily Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs). These chips are purpose-built for the parallel computation AI training demands. A single high-end GPU, such as NVIDIA’s A100 or H100, draws between 300 and 700 watts under load. In a large training run, thousands of these chips operate in parallel for weeks or months. The aggregate power draw of such a cluster can reach tens of megawatts enough to power a small city.

2.2. The Scale of Modern AI Models:

The scale of the problem becomes clearer when you trace the growth in model size over just a few years. In 2018, BERT then considered a landmark had 340 million parameters. By 2020, GPT-3 reached 175 billion. Google’s PaLM, released in 2022, reportedly had 540 billion. Today’s frontier models are estimated to operate in the trillions. This growth is not linear in its energy impact: the compute required scales roughly with the product of parameter count and training tokens, meaning each jump in model size brings a disproportionate increase in energy consumption. The 2022 “Chinchilla” paper by Hoffmann et al. offered a corrective, showing that many models had been over-parameterized relative to training data and that more efficient scaling could achieve comparable quality at lower compute cost. Even so, the general direction of travel in frontier AI remains toward larger, not smaller, models.

It is also worth noting a less-discussed but significant cost: the research process that precedes any major training run. Before committing to a full-scale job, teams typically run dozens or

hundreds of smaller experiments testing architectures, adjusting hyperparameters, exploring data mixtures. These “ablation” runs collectively amount to a substantial fraction of the final run’s cost. They almost never appear in official energy reports, meaning published figures systematically underestimate the true cost of producing a model.

2.3. The Energy Chain: From Computation to Carbon:

The link between a GPU’s power draw and carbon in the atmosphere depends heavily on where the electricity comes from. Data centers draw power from their local grid, and that grid’s carbon intensity measured in grams of CO₂ per kilowatt-hour varies enormously by region. In France, where nuclear power dominates, grid intensity can fall below 100 gCO₂/kWh. In coal-dependent regions, it can exceed 600 gCO₂/kWh. This means an identical training run can have a carbon footprint six times higher depending purely on its geographic location before any algorithmic or hardware differences are considered.

There is also a second energy cost that often goes unmentioned: cooling. Every chip generates heat, and cooling systems whether air conditioning, liquid cooling, or evaporative towers typically account for 30 to 40 percent of a data center’s total energy use, on top of computation itself. The Power Usage Effectiveness (PUE) ratio describes this overhead: a PUE of 1.0 means all power goes to computation; a PUE of 2.0 means half is lost to overhead. Most modern hyperscale facilities operate between 1.1 and 1.4, while older centers can exceed 2.0.

3. Measuring the Carbon Footprint of AI

3.1. The Core Formula:

The fundamental calculation for an AI training run’s carbon footprint is a multiplication: total energy consumed, multiplied by the carbon intensity of the electricity source. Carbon intensity measured in kilograms of CO₂ per kilowatt-hour varies dramatically depending on where and when training takes place.

$$\text{Carbon Emissions (kg CO}_2\text{eq)} = \text{Energy Consumed (kWh)} \times \text{Carbon Intensity (kg CO}_2\text{eq / kWh)}$$

Running the same training job in a coal-heavy region versus one powered by hydro or nuclear can produce emissions totals differing by a factor of ten or more. A more complete version, used by modern measurement tools, also incorporates the data center’s PUE to account for cooling and infrastructure overhead: Total Energy = GPU Energy × PUE + Network Energy + Storage Energy. This makes geographic location one of the most powerful variables in the sustainability equation even before any algorithmic or hardware changes are considered.

3.2. Measurement Tools and Frameworks:

A growing ecosystem of tools has emerged to make emissions measurement more accessible and consistent:

- **CodeCarbon:** An open-source Python library developed by researchers at Mila and McGill University. It integrates directly into training code, monitors GPU and CPU power draw in real time, queries live grid carbon intensity from services like

ElectricityMap and WattTime, and generates CO2 emissions reports. It is increasingly regarded as the standard tool for ML carbon tracking.

- **ML CO2 Impact Calculator:** A web-based tool by Lacoste et al. (2019) that estimates training emissions from hardware type, training duration, and cloud provider region. Less precise than CodeCarbon but requires no code integration, making it accessible for quick estimates.
- **Experiment Impact Tracker:** By Henderson et al. (2020), this tool monitors GPU power draw, CPU usage, and memory consumption in real time, maps those figures to grid carbon intensity, and is designed for reproducibility enabling researchers to include energy data as a standard part of published results.
- **Green Algorithms Calculator:** A broader tool covering computational research across AI, bioinformatics, genomics, and other fields. Particularly useful for cross-disciplinary comparisons and for demonstrating that computational efficiency is a universal research concern.

3.3. Challenges in Accurate Measurement:

Despite these tools, the gap between what can be measured and what needs to be measured is substantial. Three specific problems deserve attention.

First, company disclosure remains almost entirely voluntary. The major AI labs rarely publish detailed energy data for large training runs, partly for competitive reasons and partly because no regulatory requirement exists. Without this data, independent researchers are forced to work from estimates based on hardware configurations and training durations that are themselves often only partially disclosed.

Second, inference costs are systematically underweighted. The energy from a training run is a one-time cost. But every query a deployed model answers consumes energy too, and that accumulates continuously. For popular models, the cumulative inference cost over a model's operational lifetime can exceed training costs by an order of magnitude. Current measurement frameworks overwhelmingly focus on training, leaving the larger part of the picture largely unexplored.

Third, embodied carbon the emissions from manufacturing GPUs, servers, and data center infrastructure is almost entirely absent from standard accounting. Estimates for the embodied carbon of a single high-end GPU range from 150 to 500 kg CO₂e. Across a training cluster of tens of thousands of chips, this adds up to thousands of tonnes before the first training step runs. As global demand for AI hardware accelerates, this blind spot becomes increasingly difficult to justify.

4. Real-World Data and Case Studies

4.1. Notable Training Runs and Their Emissions:

The following case studies represent the best-documented estimates available. Many figures are estimates based on hardware configurations, training durations, and regional grid data because exact data is rarely published by the organizations conducting the training.

- **BERT (2018):** Consumed an estimated 1,507 kWh during training, producing roughly 652 kg of CO₂ comparable to a round-trip flight between New York and San Francisco. Modest by later standards, but BERT was one of the first models whose energy cost was publicly calculated, making it a landmark in AI sustainability awareness.
- **GPT-2 (2019):** With 1.5 billion parameters, it consumed approximately 27 MWh roughly 18 times more than BERT for a model about 4.4 times larger, illustrating the non-linear relationship between model scale and energy cost.
- **GPT-3 (2020):** OpenAI's GPT-3 consumed an estimated 1,287 MWh of electricity during training, producing approximately 552 tonnes of CO₂ roughly equivalent to the annual carbon footprint of 120 average Americans.
- **Gopher (2021):** DeepMind's 280-billion-parameter model reportedly produced around 380 tonnes of CO₂ equivalent during training. DeepMind used carbon offset programs to address this, though the adequacy of offsets as a long-term sustainability strategy remains genuinely debated.
- **PaLM (2022):** Google's 540-billion-parameter model had estimated training emissions of approximately 1,460 tonnes of CO₂ equivalent, consuming around 3,400 MWh. Google stated it powered training on carbon-free energy where possible, though full details have not been independently verified.
- **GPT-4 and Beyond (2023–present):** OpenAI has not published official training figures for GPT-4. Independent estimates based on known hardware configurations and reported training timelines place its carbon footprint in the range of 2,000 to 5,000+ tonnes of CO₂ equivalent though these figures carry significant uncertainty.

4.2. Comparative Performance Data:

The table below summarizes what we know or can reasonably estimate about training emissions across landmark models. The scaling trend it reveals is difficult to ignore:

Model	Year	Parameters	Energy (MWh)	CO ₂ (tonnes)	Equivalent To
BERT	2018	340M	~0.0015	~0.00065	NYC–SF round-trip flight
GPT-2	2019	1.5B	~0.027	~0.011	Several long car trips
GPT-3	2020	175B	~1,287	~552	120 Americans (annual)
Gopher	2021	280B	~unknown	~380	80+ Americans (annual)
PaLM	2022	540B	~3,400	~1,460	315+ Americans (annual)
GPT-4 (est.)	2023	est. ~1T+	~5,000+	~2,000+	430+ Americans (annual)

Table 1: Estimated training energy and carbon emissions for notable AI models. Equivalencies use U.S. EPA figures for average American annual carbon footprint (~14.5 tonnes CO₂e). Figures marked “est.” are independent estimates.

4.3. The Inference Problem:

Training costs grab the headlines, but there is a quieter energy drain that often goes unexamined: inference the energy consumed every time someone actually uses a trained model. A 2023 analysis by researchers at Carnegie Mellon University found that for highly popular models, inference can account for 80 to 90 percent of the total lifetime energy cost, far outstripping the one-time training expense. Every chatbot message, every AI-generated image, every algorithmic recommendation draws from that running total. At the scale of hundreds of millions of daily users, this cumulative figure is the dominant environmental story, not training.

The inference problem is also harder to address than training efficiency. Training runs can be carefully optimized, scheduled during low-carbon periods, and consolidated onto efficient hardware. Inference is distributed, demand-driven, and unpredictable popular models may serve tens of thousands of simultaneous requests from all regions at all hours. The optimization challenges are different in kind, not just in degree, and they require deliberate engineering investment that deployment teams have not historically prioritized.

4.4. Water Consumption: The Other Hidden Cost:

Energy and carbon are not the only environmental costs worth noting. Many data centers particularly those using evaporative cooling towers consume substantial quantities of freshwater. Research by Shaolei Ren and colleagues at UC Riverside estimated that training GPT-3 consumed approximately 700,000 liters of water for cooling. A single large data center can use millions of gallons per day. In regions already experiencing water stress, the expansion of AI data center infrastructure places additional pressure on a resource that is becoming scarcer in many parts of the world.

5. Mitigation Strategies

There is no single fix for AI's environmental footprint. Meaningful progress requires action at every level from individual researchers choosing their tools to cloud providers redesigning infrastructure to policymakers setting expectations for the industry. What follows is an overview of the most evidence-supported approaches currently available.

5.1. Algorithmic Efficiency:

Perhaps the most impactful lever available is designing models that require less computation to achieve strong results. This does not have to mean sacrificing capability research has consistently shown that thoughtful model design can match or approach the performance of larger models at a fraction of the energy cost. Key techniques include:

- **Pruning:** Removing parameters from a trained model that contribute minimally to its outputs. Large networks are often dramatically over-parameterized removing 50 to 90 percent of weights while preserving most accuracy is frequently possible. Structured pruning (removing entire neurons or attention heads) is especially practical because it produces models that run faster on standard hardware.
- **Quantization:** Replacing 32-bit floating-point numbers used to represent model weights with lower-precision formats (8-bit integers or even 4-bit). This reduces memory footprint and speeds up inference significantly. NVIDIA's H100 GPU is specifically

optimized for FP8 operations, and models like LLaMA 2 have official 4-bit quantized releases.

- **Knowledge Distillation:** Training a smaller “student” model to mimic the behavior of a larger “teacher” model. The student learns from the teacher’s soft probability distributions, which carry richer information than hard labels alone. DistilBERT, for example, retains approximately 97 percent of BERT’s performance while using 40 percent fewer parameters and running 60 percent faster.
- **Mixture of Experts (MoE):** An architectural approach that routes each input token to only a subset of “expert” sub-networks rather than the entire model. This allows total parameter counts to be very large while keeping per-token compute manageable. Used in Google’s Switch Transformer and reportedly in some configurations of GPT-4.
- **NAS with Efficiency Constraints:** Neural Architecture Search with efficiency constraints: Automating the process of finding model designs that optimize for both accuracy and computational efficiency simultaneously.
- **Chinchilla-Optimal Scaling:** The 2022 “Chinchilla” paper by Hoffmann et al. demonstrated that many large models had been trained sub-optimally too large for the amount of training data they received. Training models at the compute-optimal ratio of parameters to tokens produces equivalent or better quality at lower total energy consumption.

5.2. Geographic and Temporal Optimization:

One of the most striking findings in recent AI sustainability research is how much carbon emissions can be reduced simply by choosing where and when to run training jobs without any change to the model, hardware, or algorithm. Research by Dodge et al. (2022) found that the same training run could vary in carbon emissions by a factor of roughly 30x depending on the geographic location of the data center. Even within a single cloud provider’s infrastructure, choosing the right region can reduce emissions by 80 percent or more.

Temporal optimization adds another dimension. Grid carbon intensity varies throughout the day as demand and generation mix shift. Tools like WattTime and ElectricityMap provide real-time and forecast data on grid intensity, enabling training jobs to be scheduled when the carbon cost is lowest. For jobs that can tolerate a modest delay in start time which includes most research training runs this requires only a scheduling change. Major cloud providers including Google, Microsoft Azure, and AWS now offer carbon-aware scheduling features, though they remain opt-in rather than the default.

5.3. Hardware Improvements:

Hardware efficiency improvements continue to deliver meaningful gains. NVIDIA’s H100 GPU offers approximately 3x better performance-per-watt than the A100 it replaced, and around 9x better than the V100 generation. A training run on H100s may consume substantially less energy than the same run on older hardware, even with identical model and algorithm. Custom AI accelerators Google’s fifth-generation TPUs, Cerebras’s Wafer-Scale Engine, Graphcore’s IPU represent different architectural approaches to further improving efficiency. On the longer horizon, neuromorphic chips designed to mimic the energy-efficient signaling of biological

neurons hold significant promise. Data centers are also improving their PUE ratios through direct liquid cooling, outside-air cooling in appropriate climates, and better heat management architectures.

5.4. Renewable Energy Procurement:

Major tech companies have made significant public commitments to renewable energy. Google claims to have matched its global electricity consumption with renewable energy purchases since 2017; Microsoft has pledged to be carbon-negative by 2030. These commitments are meaningful they drive real investment in renewable infrastructure but they come with an important caveat. Annual matching through Renewable Energy Certificates (RECs) is not the same as running on clean power at every hour. A company can consume coal-powered electricity at night and purchase solar RECs generated during the day, without clean electrons ever reaching the facility that needed them.

The more meaningful standard is 24/7 Carbon-Free Energy (CFE) matching consumption with clean generation at every hour, in every location. Google's data centers achieved 64 percent 24/7 CFE as of 2022 and has set a goal of 100 percent by 2030. This requires on-site generation, long-duration energy storage, or time-matched power purchase agreements a fundamentally harder engineering and commercial challenge than annual matching, and the right target for the industry.

5.5. Transparency and Reporting Standards:

Right now, there is no accepted standard for disclosing the energy and carbon cost of AI training. Most academic papers do not report it; most companies do not publish it. This makes independent accountability nearly impossible. A researcher publishing a new model can describe its architecture, benchmark performance, and training data in detail, then say nothing at all about how much energy it consumed. This happens routinely, and it needs to change.

Establishing transparency standards analogous to financial reporting requirements would change the incentive structure. If organizations had to report training emissions alongside benchmark results, the pressure to improve would increase considerably. At minimum, researchers should report hardware used, training duration, geographic location, estimated energy consumption, and estimated carbon emissions. Conference program committees at venues like NeurIPS, ICML, ICLR, and ACL should consider making this a condition of acceptance.

5.6. Lifecycle Thinking and the Importance of Inference:

Given that inference often accounts for the majority of a model's total lifetime energy cost, optimizing deployed systems is just as important as optimizing training. Effective approaches include model caching (storing and reusing results for common queries), batching (grouping multiple requests together to improve GPU utilization), query routing (directing simple requests to smaller specialized models rather than routing everything through a large general-purpose system), and speculative decoding (where a small draft model generates candidate tokens that a larger model verifies in parallel, reducing average compute per generated token).

Edge AI deploying optimized models directly on local devices rather than in cloud data centers offers another dimension of inference efficiency. On-device inference eliminates round-trip energy to a remote data center and can leverage locally available renewable energy. The rapid improvement of on-device AI chips, including Apple's Neural Engine and Qualcomm's Hexagon DSP, is making edge deployment increasingly viable for a broader range of tasks.

6. Challenges and Limitations

6.1. *The Scaling Incentive:*

Even with a solid toolkit of mitigation strategies available, systemic forces keep pushing in the wrong direction. The economic incentive to build bigger models is real and persistent larger models tend to perform better on benchmarks, and competitive pressure between labs has created something of an arms race in model scale. For any individual organization to unilaterally prioritize efficiency over capability requires either strong internal values or external regulation. Neither has yet fully materialized. The result is a coordination problem: each actor behaves rationally in its own interest while producing an outcome that is collectively unsustainable.

6.2. *Lack of Transparency:*

Without disclosure, there is no accountability, and without accountability, there is no pressure to improve. When companies do not publish training energy data, researchers cannot estimate the industry's total environmental impact or track whether it is getting better or worse. The opacity is also self-reinforcing: because there is no norm of disclosure, organizations that do provide data voluntarily expose themselves to scrutiny that their competitors avoid. Until disclosure is universal through regulation or strong community norms this gap will persist.

6.3. *The Jevons Paradox:*

There is a well-established economic pattern worth flagging: the Jevons Paradox. When a technology becomes more efficient and therefore cheaper to use, demand typically rises sometimes enough to increase total resource consumption even as per-unit consumption falls. First observed by economist William Jevons in the context of coal-efficient steam engines in the 19th century, the dynamic is highly relevant to AI. As training and inference become cheaper, AI is deployed in more applications at greater scale. A 10x improvement in inference efficiency that enables 100x growth in usage does not reduce total energy consumption it increases it. Efficiency gains are necessary but not sufficient; they need to be paired with mechanisms that also constrain total demand.

6.4. *Embodied Carbon:*

Virtually all current estimates of AI's carbon footprint ignore embodied carbon: the emissions generated during the manufacture of GPUs, servers, and data center infrastructure. Fabricating a modern chip at leading-edge semiconductor nodes is an energy-intensive industrial process involving exotic chemicals, ultrapure water, and cleanroom facilities. Estimates for the embodied carbon of a single high-end GPU range from 150 to 500 kg CO₂e. Across a training cluster of

tens of thousands of chips, this represents thousands of tonnes of emissions before the first training step runs and it almost never appears in any published accounting.

6.5. Geopolitical and Infrastructure Inequities:

The benefits of efficient, renewably powered AI are not distributed evenly. Advanced semiconductor manufacturing is concentrated in a handful of countries. Renewable energy infrastructure is unevenly distributed globally. The ability to choose low-carbon data center locations, schedule for renewable availability, or invest in cutting-edge hardware is much greater for well-resourced organizations in developed economies than for researchers and developers in less resource-rich environments. Any serious policy approach to AI sustainability needs to grapple with these distributional questions, or risk creating barriers that disproportionately affect smaller actors while doing little to constrain the large labs whose aggregate footprint dominates.

7. Conclusion

7.1. Synthesis of Key Findings:

The environmental cost of training large AI models is real, measurable, and growing. Training frontier models can produce hundreds to thousands of tonnes of CO₂ equivalent, and the cumulative energy cost of inference over a model's operational lifetime often dwarfs even that figure. This is not a theoretical future problem it is happening now, at scale, largely without public accountability. The measurement tools exist but are underused. The solutions exist but are not yet the default. The barriers are real but not insurmountable.

7.2. Path Forward:

There is genuine reason for optimism. The solutions exist and many are proven. Algorithmic efficiency techniques pruning, quantization, knowledge distillation, Chinchilla-optimal scaling can dramatically reduce compute requirements for both training and inference. Geographic and temporal scheduling can cut emissions by 80 percent or more without touching the model. Renewable energy procurement, hardware improvements, inference optimization, and transparency standards all contribute meaningfully when applied together.

What is missing is not the technology it is the institutional will to treat sustainability as a genuine priority rather than a footnote. Researchers should routinely include energy and emissions estimates in published results. Conferences and journals should require this disclosure. Cloud providers should make green scheduling the default. Policymakers should consider carbon reporting requirements for large-scale AI development and invest in grid decarbonization, which is the single most impactful systemic change available.

Ultimately, this is not really a question of whether AI can be made more sustainable it clearly can. The question is whether the people building these systems will choose to prioritize that before the environmental cost becomes irreversible. The tools are there. The data is there. What remains is the decision to act on it.

8. References

Journal Articles and Conference Papers:

- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). doi:10.18653/v1/P19-1355
- Patterson, D., Gonzalez, J., Le, Q., et al. (2021). Carbon Emissions and Large Neural Network Training. arXiv preprint arXiv:2104.10350.
- Dodge, J., Prewitt, T., et al. (2022). Measuring the Carbon Intensity of AI in Cloud Instances. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. doi:10.1145/3531146.3533234
- Henderson, P., Hu, J., Romoff, J., et al. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research*, 21(248), 1–43.
- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. doi:10.1145/3381831
- Wu, C. J., Raghavendra, R., et al. (2022). Sustainable AI: Environmental Implications, Challenges, and Opportunities. *Proceedings of Machine Learning and Systems (MLSys)*.
- Lacoste, A., Lottick, A., Schwartz, R., et al. (2019). Quantifying the Carbon Emissions of Machine Learning. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.
- Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. arXiv preprint arXiv:2203.15556. (The “Chinchilla” paper.)
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
- Bannour, N., Ghannay, S., Névéal, A., & Ligozat, A. L. (2021). Evaluating the Carbon Footprint of NLP Methods: A Survey and Analysis of Existing Tools. *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*.

Websites:

- Google. (2023). Google 2023 Environmental Report. Retrieved from <https://sustainability.google/reports>
- Microsoft. (2023). Microsoft 2023 Environmental Sustainability Report. Retrieved from <https://www.microsoft.com/en-us/corporate-responsibility/sustainability>
- OpenAI. (2023). GPT-4 Technical Report. Retrieved from <https://openai.com/research/gpt-4>
- ElectricityMap. (2024). Real-time Carbon Intensity of Electricity. Retrieved from <https://electricitymap.org>
- WattTime. (2024). Automated Emissions Reduction (AER). Retrieved from <https://www.watttime.org>