PEER-REVIEWED JOURNAL OF COMPUTER SCIENCE (PRJCS)

ISSN Online : Open Access Scholarly Journal

Sustainable Green Computing and Carbon-Aware Artificial Intelligence

Tintu George

Assistant Professor, Department of BCA AI, Sri Ramakrishna College of Arts & Science, Coimbatore, India.

Article information

Received: 13th April 2026
Received in revised form: 21st May 2026
Accepted: 30th May 2026
Available online: 10th June 2026

Volume: 1
Issue: 6
DOI: https://doi.org/10.5281/zenodo.20554629

Abstract

The energy footprint of large-scale computing has become a pressing concern as artificial intelligence workloads expand. Training a single frontier language model can consume thousands of megawatt-hours, and global data-centre electricity demand is projected to rise sharply over the next decade. This paper surveys the field of sustainable computing with a focus on artificial intelligence: efficiency at the algorithm, model, and system levels; carbon-aware scheduling that exploits regional and temporal variation in grid carbon intensity; hardware efficiency improvements; and reporting frameworks. We discuss the policy landscape, including the EU AI Act and U.S. Executive Order on AI, and the role of standardised carbon accounting. Open challenges in measuring and reducing both operational and embodied emissions of AI systems are highlighted, alongside the trade-offs between efficiency, capability, and equity.

Keywords: - Green Computing, Sustainable AI, Carbon-Aware Computing, Energy Efficiency, MLPerf, Data Centres, Embodied Carbon.

.

I. INTRODUCTION

Computing accounts for roughly 1-2% of global electricity consumption, with data centres alone estimated to consume 1.0-1.3% of world electricity in 2022 [1]. The growth of artificial intelligence workloads has shifted the trajectory: the International Energy Agency projects that data-centre electricity demand could approximately double by 2030 [2], driven primarily by AI training and inference. Strubell et al. drew early attention to the environmental cost of deep-learning research [3], and a series of subsequent studies have refined the picture and introduced disciplined accounting [4], [5].

This paper surveys sustainable green computing as it applies to AI. Section II covers the energy and carbon profile of contemporary AI. Section III discusses algorithmic and model-level efficiency. Section IV examines carbon-aware scheduling. Section V reviews hardware advances. Section VI surveys reporting and policy. Section VII identifies open challenges, and Section VIII concludes.

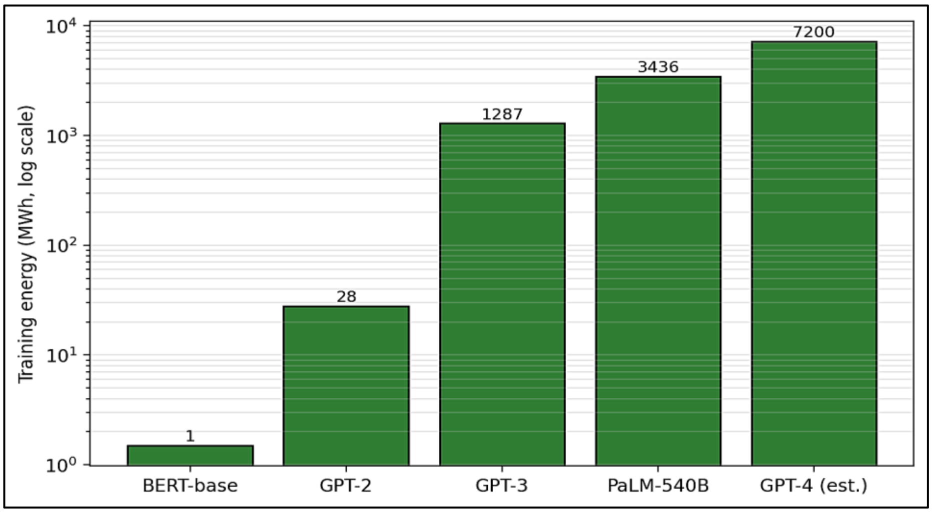
II. THE CARBON FOOTPRINT OF AI

Carbon emissions from computing decompose into operational emissions, arising from electricity consumed in operation, and embodied emissions, arising from manufacturing and disposal of hardware. For AI workloads, operational emissions dominate over the typical lifetime, but embodied emissions of GPUs and accelerators are non-trivial and increasingly significant as hardware refresh cycles shorten [6]. Patterson et al. [4] estimated training emissions for several frontier models, finding that careful choices of model, datacenter, and processor can reduce emissions by up to two orders of magnitude relative to naive baselines.

RESEARCH PUBLICATIONS

Figure 1 plots representative training-energy estimates (in MWh) for a sequence of widely cited language models. The exponential trend is unmistakeable, although recent work suggests that the marginal capability per joule has improved as models become more compute-efficient [7]. Inference now contributes more than training to the operational footprint of widely deployed models, since training is one-time and inference is continuous [8].

Fig. 1. Approximate training energy of representative language models (log scale).



III. EFFICIENCY AT THE ALGORITHM AND MODEL LEVEL

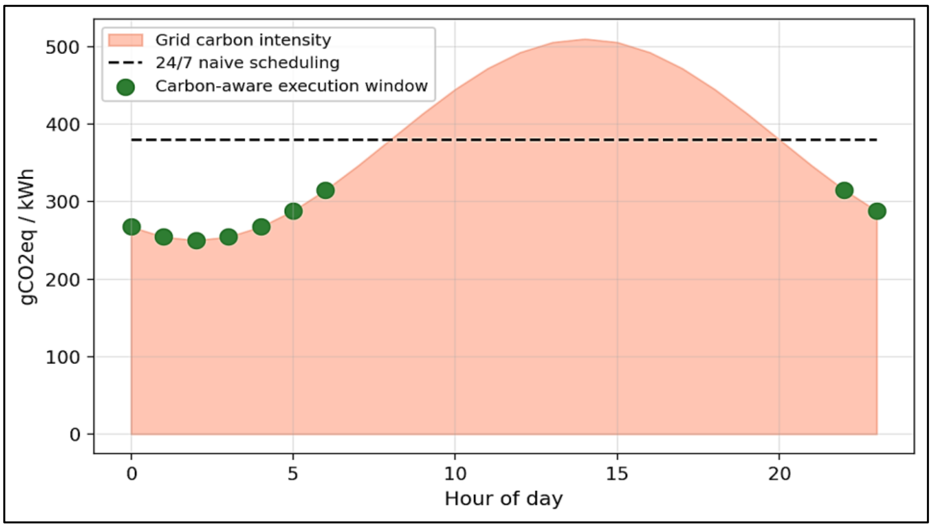
Several algorithmic strategies reduce AI energy use. The Chinchilla scaling analysis [9] showed that for a fixed compute budget, smaller models trained on more tokens achieve lower loss than the prior generation of larger, under-trained models, an immediate efficiency win. Mixture-of-experts architectures activate only a subset of parameters per input, reducing inference compute [10]. Model distillation, pruning, and quantisation, discussed in detail in the small-language-model literature, all reduce per-query energy at deployment time. Schwartz et al. coined the term green AI to advocate that efficiency be reported as a first-class research metric alongside accuracy [11]. Henderson et al. proposed concrete tooling for tracking experiment energy and emissions [12].

IV. CARBON-AWARE COMPUTING

The carbon intensity of electricity varies substantially across regions and across hours of the day, ranging from below 50 gCO2eq/kWh in renewable-rich regions to above 700 gCO2eq/kWh in coal-dominated systems [13]. Carbon-aware computing exploits this variation by deferring or relocating flexible workloads to lower-intensity windows. Microsoft, Google, and Meta have reported reductions of 20-50% in operational carbon emissions for elastic batch workloads using carbon-aware schedulers [14], [15]. Figure 2 illustrates the underlying principle on a synthetic but representative grid-intensity profile.

Carbon-aware computing is not without trade-offs. Workload deferral can increase latency, and geographic shifting may conflict with data-residency regulations. The Green Software Foundation [16] has issued patterns and a Software Carbon Intensity (SCI) specification to guide practitioners.

Fig. 2. Hourly grid carbon intensity (synthetic) and a carbon-aware execution window.



V. HARDWARE EFFICIENCY

Modern accelerators have steadily improved performance per watt. Nvidia's H100 and B200 generations, Google's TPU v4 and v5, and AMD's MI300 series have each delivered roughly 2-4x improvements in tokens-per-joule over their predecessors. Liquid cooling, including direct-to-chip and immersion approaches, reduces datacenter PUE (power usage effectiveness) substantially below the 1.5 industry baseline; hyperscale operators report PUE values close to 1.1 [17]. Embodied-carbon analysis remains less mature than operational analysis; recent work shows that embodied emissions are a non-trivial fraction of life-cycle emissions for GPUs operating in high-renewable regions [6].

VI. REPORTING AND POLICY

Reporting frameworks have proliferated. The MLCO2 calculator [18] provides standardised energy and emissions estimates for ML experiments. CodeCarbon [19] instruments arbitrary Python code. MLPerf has introduced an energy track that measures power during training and inference. On the policy side, the EU AI Act [20] and the U.S. Executive Order on AI [21] both flag the environmental impact of large models, and several jurisdictions have begun mandating disclosure of training-time energy for frontier systems. Voluntary commitments such as the Climate Neutral Data Centre Pact set targets for European datacenter operators.

The reporting picture is uneven across regions and across the corporate sector. Hyperscale operators publish increasingly detailed sustainability reports because shareholder, regulatory, and customer pressure pushes them to do so, while mid-tier and on-premise operators often publish little. The asymmetry distorts the public conversation, since the most visible numbers describe the most efficient facilities while the long tail of less efficient deployments stays invisible. Independent academic studies and watchdog reports have begun to fill this gap, but their methodologies vary and their access to primary data is limited. The value of an audit-grade, jurisdiction-wide measurement infrastructure is hard to overstate, and the EU's mandatory reporting from 2026 onward is the most concrete step in that direction. The expectation among policy specialists is that other jurisdictions will follow within a few years, with reporting becoming the baseline rather than a competitive differentiator.

Table 1. Energy-Efficiency Strategies at Different Layers

Strategy	Layer	Typical reduction
Compute-optimal scaling [9]	Model design	30-70%
Distillation + 4-bit quantisation	Model deployment	5-10x
Carbon-aware scheduling [14]	Job scheduling	20-50%
Datacenter PUE 1.1 [17]	Facility	20-30%
MoE routing [10]	Architecture	2-5x inference
Hardware generation upgrade	Hardware	2-4x per generation

VII. CHALLENGES AND OPEN PROBLEMS

Several open problems persist. Measurement remains inconsistent: different studies report different boundaries (training only, full lifecycle, with or without retries), making comparisons unreliable [22]. Embodied carbon of high-end accelerators is known with low precision and depends on supplier-specific data. Inference at planetary scale, particularly for user-facing chat assistants, may overtake training as the dominant emission source. Rebound effects, in which efficiency improvements drive larger workloads, can offset gains [23]. Equity considerations matter: the largest models concentrate emissions in a small number of organisations, while their benefits and costs are distributed unevenly across the world. Finally, the interaction of AI workloads with grid stability sudden, large, dispatchable loads can both stress and support grids is an emerging research and policy frontier.

VIII. WATER, EMBODIED CARBON, AND THE WIDER FOOTPRINT

The carbon footprint of computing has received the bulk of academic attention, but it is not the only environmental impact that data centres impose. Water consumption is increasingly prominent. A typical hyperscale facility in a hot, arid climate consumes between five and twenty million litres of water per megawatt of capacity per year, primarily through evaporative cooling. Recent analyses estimate that the major US AI training runs of 2022-23 consumed several hundred million litres of fresh water in cooling [24]. The interaction with regional drought, particularly in Arizona, Texas, and parts of the Iberian Peninsula, has made water use a contested local issue. Several operators have committed to water-positive operation by 2030, with closed-loop and air-cooled designs as the principal levers, but the transition is uneven across regions.

Embodied carbon is the second under-counted impact. Manufacturing a high-end GPU emits in the order of one hundred to several hundred kilograms of carbon dioxide equivalent, depending on the foundry node, the supply chain, and accounting boundaries [6]. For a fleet of tens of thousands of accelerators on a three- to five-year refresh cycle, embodied emissions become a non-trivial fraction of life-cycle impact, particularly when operational electricity is dominantly low-carbon. The shift to advanced foundry nodes (3 nm and below) and the rapid generational refresh of AI accelerators tend to push embodied emissions upward. Extending refresh cycles through aggressive software optimisation is one mitigation; secondary-market reuse is another, and several hyperscale operators now run published programmes for reusing previous-generation accelerators in lower-priority workloads.

Standardised reporting remains a significant gap. Different studies use different boundaries: training only, full lifecycle, with or without retries, with or without data preparation, aggregated across organisations or per paper. The MLCO2 calculator [18] and CodeCarbon [19] have improved consistency at the level of individual experiments, but enterprise-wide footprints are still difficult to compare across organisations. The Software Carbon Intensity specification of the Green Software Foundation [16] proposes a common metric, but adoption is voluntary and uneven. Until reporting is standardised and audited, the public discussion will continue to feature large headline numbers whose comparability is questionable.

Policy is starting to fill some of the gap. The European Energy Efficiency Directive (EU 2023/1791) imposes mandatory energy-and-water reporting on data centres above 500 kW from 2024, with public disclosure beginning in 2026. The U.S. Executive Order on AI [21] requires major training runs to report energy use to the National Institute of Standards and Technology. Several Asian jurisdictions are following with similar measures. The combination of standardised reporting and regulatory pressure will probably do more for sustainable AI in the next five years than any single technical innovation. The point of accounting, in the end, is to make trade-offs visible to decision-makers, and decision-makers cannot manage what they cannot see.

It is also worth confronting a tension that the green-AI literature sometimes sidesteps. Improvements in efficiency tend to enable larger workloads rather than smaller energy bills, a pattern called the Jevons effect that has been observed across computing for decades. The doubling of inference efficiency on every new generation of accelerator has so far been more than offset by the growth of inference workloads, with net energy use rising. Carbon-aware scheduling reduces the carbon intensity of a fixed workload but does not, by itself, slow workload growth. Honest accounting must therefore consider both intensity and absolute volume, and the right metric for evaluating sustainability programmes is total emissions over a stated boundary rather than per-query efficiency alone. The implication for policy is that pure efficiency standards may be insufficient and that absolute caps, total-emissions reporting, or carbon-pricing mechanisms may eventually be needed for the largest workloads. The implication for engineers is that efficiency work is necessary but not sufficient; the question of whether a particular workload should run at all, in some boundary cases, becomes a legitimate engineering question. Few organisations are comfortable asking that question today, and few have the governance machinery to answer it, but the conversation is overdue. Sustainability is not only a matter of doing the same things better; it is also a matter of choosing what to do, and that choice belongs as much to engineering as to policy.

IX. CONCLUSION

Sustainable computing is no longer a niche concern. As AI workloads grow, the field must adopt disciplined measurement, prioritise efficiency at all layers from algorithm to facility, and exploit the temporal and geographic flexibility of the grid through carbon-aware computing. Hardware advances and architectural innovations help but cannot substitute for honest reporting and informed engineering choices. The combined research, industry, and policy attention now directed at sustainable AI is encouraging, but the trajectory of total emissions remains upward. Bending that curve will require sustained, coordinated effort across all layers of the computing stack.

A final point worth emphasising is that the conversation about sustainable AI cannot be conducted purely within the AI community. Power generation, water rights, semiconductor manufacturing, regional climate policy, and grid operation are all upstream of the questions discussed here, and the most consequential decisions will often be made by stakeholders who do not consider themselves part of AI at all. Researchers and engineers in the field can contribute by reporting honestly, by designing systems that are amenable to flexibility and partial loads, and by participating in policy discussions when invited. The temptation to wait for someone else to solve the broader problem is real but unhelpful. The pattern that has emerged from the most credible sustainability programmes is one of cross-functional engagement, with engineers, policy specialists, and operations teams sharing accountability for outcomes that no single discipline can deliver alone. The next decade of progress on sustainable AI will depend on whether that pattern generalises across the industry, and whether public regulation provides the framework within which private engineering can be held to account. Both fronts need attention, and neither is a substitute for the other.

REFERENCES

[1] International Energy Agency. “Data Centres and Data Transmission Networks.” *IEA Tracking Report*, 2023.

[2] International Energy Agency. *Electricity 2024: Analysis and Forecast to 2026*. IEA Report, 2024.

[3] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. “Energy and Policy Considerations for Deep Learning in NLP.” In *Proceedings of the ACL*, 2019.

[4] Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, et al. “Carbon Emissions and Large Neural Network Training.” arXiv preprint arXiv:2104.10350, 2021.

[5] Patterson, David, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, et al. “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.” *Computer* 55, no. 7 (2022): 18–28.

[6] Gupta, Udit, Young Geun Kim, Stephen Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, et al. “Chasing Carbon: The Elusive Environmental Footprint of Computing.” In *Proceedings of the IEEE HPCA*, 2021.

[7] Luccioni, Sasha, Alejandro Hernandez-Garcia, and Yacine Jernite. “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” In *Proceedings of the ACM FAccT*, 2024.

[8] Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Narges Ardalani, Kwanghoon Maeng, et al. “Sustainable AI: Environmental Implications, Challenges and Opportunities.” In *Proceedings of the MLSys*, 2022.

[9] Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, et al. “Training Compute-Optimal Large Language Models.” In *Proceedings of NeurIPS*, 2022.

[10] Fedus, William, Barret Zoph, and Noam Shazeer. “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity.” *Journal of Machine Learning Research* 23 (2022).

[11] Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. “Green AI.” *Communications of the ACM* 63, no. 12 (2020): 54–63.

[12] Henderson, Peter, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. “Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning.” *Journal of Machine Learning Research* 21 (2020).

[13] European Environment Agency. “Greenhouse Gas Emission Intensity of Electricity Generation in Europe.” *EEA Indicator*, 2023.

[14] Bashir, A. A., Z. Liu, Bilge Acun, Arun Sundaresan, Kim Hazelwood, Misha Naumov, and B. Saha. “Carbon-Aware Computing for Datacenters.” *Communications of the ACM* 67, no. 7 (2024): 60–70.

[15] Radovanovic, Ana, Roy Koningstein, Igor Schneider, Bo Chen, Andre Duarte, Ben Roy, et al. “Carbon-Aware Computing for Datacenters.” *IEEE Transactions on Power Systems* 38, no. 2 (2023): 1270–1280.

[16] Green Software Foundation. *Software Carbon Intensity (SCI) Specification, Version 1.0*. Standard, 2023.

[17] Uptime Institute. *Annual Global Data Center Survey 2023*. Uptime Institute Report, 2023.

[18] Lacoste, Alexandre, Sasha Luccioni, Victor Schmidt, and Thomas Dandres. “Quantifying the Carbon Emissions of Machine Learning.” In *Proceedings of the NeurIPS Climate Change AI Workshop*, 2019.

[19] Schmidt, Kevin, Victor Schmidt, Kemal Dolapci, and Sasha Luccioni. *CodeCarbon: A Python Package for Tracking the Energy and Emissions of Computing*. GitHub repository, 2020–2024.

[20] European Parliament and Council. *Regulation on Artificial Intelligence (the AI Act)*. Regulation (EU) 2024/1689, 2024.

[21] The White House. *Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Executive Order, October 30, 2023.

[22] Anthony, Laurence F. W., Benjamin Kanding, and Raghavendra Selvan. “CarbonTracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models.” In *Proceedings of the ICML Workshop on Challenges in Deploying and Monitoring ML Systems*, 2020.

[23] Hofmann, Philipp, René Bohnsack, Alexander Geilen, Philipp Hennig, and Bernd Pfitzmann. “Rebound Effects in IT for Sustainability: A Review.” *Journal of Cleaner Production* 286 (2021).

[24] de Vries, Alex. “The Growing Energy Footprint of Artificial Intelligence.” *Joule* 7, no. 10 (2023): 2191–2194.