The American Journal of Engineering and Technology
ISSN 2689-0984 Volume 07 - 2025

From Red AI to Green AI: A Unified Survey of Lifecycle Costs, Efficiency Techniques, and a Comprehensive Reporting Framework

[1]Pinaki Bose
[1]Independent Researcher, USA

Received: 10th Sep 2025 | Received Revised Version: 16th Oct 2025 | Accepted: 19th Nov 2025 | Published: 30th Nov 2025

Volume 07 Issue 11 2025 | Crossref DOI: 10.37547/tajet/v7i11-310

Abstract

The exponential growth of large-scale Artificial Intelligence (AI) models, or "Red AI" , has led to a 300,000-fold increase in computational demand since 2012 , raising significant environmental and sustainability concerns. While the high carbon cost of model training (e.g., GPT-3's estimated 550 metric tons of CO2e) is well-documented, this focus obscures the dominant environmental burden: model inference, which can account for up to 90% of a model's total lifecycle energy consumption. A critical research gap exists in the unified analysis of carbon cost versus performance metrics across this entire AI lifecycle. Furthermore, the field lacks a standardized, comprehensive framework for Green AI reporting, hampering transparent and verifiable comparisons. This paper addresses this gap through a systematic review and quantitative synthesis of Green AI. We systematically categorize and evaluate three pillars of technical optimization: (1) model compression, (2) hardware-aware AI, and (3) low-power inference techniques. This analysis reveals that high-level architectural choices—such as using general-purpose generative models for discriminative tasks—are orders of magnitude (e.g., 14.6x to 30x) less efficient than task-specific models. We also highlight a "measurement crisis," where common reporting tools like CodeCarbon underestimate true energy consumption by 20-40% compared to ground-truth measurements. We conclude by proposing a comprehensive, lifecycle-based Green AI reporting framework, designed to integrate with existing GHG and ISO standards. This framework mandates unified cost-performance metrics (e.g., CO2e/ inference / performance-unit) to enable transparent, verifiable, and-informed decision-making for sustainable AI development.

Keywords: Green AI, Sustainable AI, Energy-Efficient AI, Model Compression, Hardware-Aware AI, Carbon Footprint, AI Lifecycle Assessment, LLM.

Cite This Article: Bose, P. (2025). From red AI to green AI: A unified survey of lifecycle costs, efficiency techniques, and a comprehensive reporting framework. The American Journal of Engineering and Technology, 7(11), 204–212. https://doi.org/10.37547/tajet/v7i11-310.

1. Introduction

The field of Artificial Intelligence (AI) is undergoing a paradigm shift characterized by an exponential increase in model scale and capability. This trend, however, has resulted in a parallel explosion in computational and energy demands. The computations required for deep learning research, for instance, saw an estimated 300,000-fold increase between 2012 and 2018, a growth rate far outpacing Moore's Law. This resource-intensive, accuracy-first paradigm, where performance is pursued at any environmental cost, has been termed "Red AI".

The environmental impact of Red AI is substantial. The training of a single large model, such as GPT-3, is

estimated to require 1,287 MWh of electricity and generate over 550 metric tons of CO_2 equivalent (CO_2e), a footprint comparable to the lifetime emissions of several cars. This has understandably drawn significant public and academic attention to the environmental cost of *training*.

However, this focus on the one-time training cost obscures a far larger and more persistent environmental challenge: the cost of *inference*. As AI models move from development to global deployment, they are executed billions or trillions of times. Recent analyses suggest that this operational inference phase, not training, constitutes the dominant portion of a model's lifecycle energy use, potentially accounting for up to 90% of the total [1]. A single generative AI query, for example, can consume 10 times the electricity of a traditional web search. Scaling a single, highly efficient GPT-4o query (0.42 Wh) to 700 million queries per day—a realistic scale for a global service—would result in an annual electricity demand equivalent to 35,000 U.S. homes. The primary challenge for sustainable, or "Green AI," therefore lies in mitigating the cumulative cost of inference.

Despite the urgency of this challenge, a significant research gap persists. While many surveys have cataloged Green AI techniques [5] , they often focus on a single optimization vector (e.g., compression) or a single lifecycle phase (e.g., training). A *unified* analysis that quantitatively links specific techniques (e.g., compression, hardware choices) to their cost-performance trade-offs across the *full* AI lifecycle—from data acquisition to training, deployment, and inference—is critically missing.

This gap is dangerously compounded by the lack of a standardized, comprehensive framework for reporting these metrics. Inconsistent methodologies and opaque reporting make "apples-to-apples comparisons... nearly impossible" [6]. Without a standard, it is impossible to verify "green" claims or make informed, sustainable choices.

This paper provides a three-fold contribution to address this multifaceted gap. First, in Section II.A, we provide a systematic categorization and evaluation of Green AI optimization strategies at the software, hardware, and system levels. Second, in Section II.B, we deliver the core contribution: a unified quantitative analysis of carbon cost versus performance across the AI lifecycle, evaluating architectural and technical trade-offs based on recent benchmarks. Third, in Section II.C, we leverage this analysis to propose a comprehensive, lifecycle-based Green AI reporting framework designed to standardize and validate emissions and performance disclosure. Section III concludes the paper.

II.A Systematic Categorization of Green AI Optimization Strategies

Achieving Green AI requires a full-stack approach, with optimizations at every level of abstraction. This section systematically categorizes the primary technical levers for improving AI efficiency.

1. Software-Level Optimization: Model Compression

Model compression techniques aim to reduce the computational and memory footprint of a model, thereby decreasing its energy consumption, particularly during inference. The primary methods include:

- **Network Pruning:** This technique removes redundant parameters (weights) from a trained network. **Pruning Techniques for Network Compression -**Network pruning removes redundant parameters from trained neural networks, primarily via two types - **Unstructured Pruning:** Removes individual weights, achieving high sparsity, but usually requires specialized hardware for speedup. **Structured Pruning:** Removes coherent blocks (e.g., channels, filters). It yields lower compression but provides immediate speedup on standard hardware.
- **Quantization:** This technique reduces the numerical precision of a model's weights and activations, typically from 32-bit floating-point (FP32) to lower-precision formats like 16-bit (FP16) or 8-bit integer (INT8). Quantization reduces the model's memory footprint and bandwidth requirements. More importantly, it allows computation to be performed using lower-power integer arithmetic units, which are significantly more energy-efficient than floating-point units.
- **Knowledge Distillation (KD):** In KD, knowledge is transferred from a large, complex "teacher" model to a smaller, more efficient "student" model. The student model is trained to mimic the teacher's outputs (e.g., soft probabilities) or intermediate representations, thereby learning a complex function without requiring a large architecture.

These techniques are often applied in combination. For example, a common pipeline involves using KD to create a smaller student model, which is then pruned and fine-tuned. More advanced joint optimization methods seek to

apply pruning, quantization, and distillation simultaneously within a single training pipeline [8]. A critical, and often overlooked, dependency exists between software compression and hardware. As noted in , unstructured pruning *alone* yields no actual run-time speedups on a standard device. The model may be sparse, but a standard GPU, which is a dense-math accelerator, will still execute the same number of operations, simply multiplying by zero. The energy savings are only realized when this technique is paired with specialized sparse-math libraries or hardware (e.g., FPGAs or ASICs) that can perform "zero-skipping" to bypass computation on zero-valued weights [9].

Table 1- Comparative Analysis of Model Compression Techniques

Technique	Sub-Type	Typical Compression Ratio	Typical Accuracy Drop	Hardware Dependency
Pruning	Magnitude Pruning (Unstructured)	5-10x	1-3%	Requires sparse libraries/hardware
	Structured Pruning	2-5x	2-5%	Direct (runs on standard hardware)
Quantization	INT8 Quantization	4x	0.5-2%	Widespread support (e.g., $INT8$ units)
	Binary Networks	32x	10-15%	Limited (requires custom hardware)
Knowledge	Response Distillation	3-10x	2-5%	Direct (student model is standard)
Distillation	Feature Distillation	3-8x	1-3%	Direct (student model is standard)

2. Hardware-Level Optimization: Hardware-Aware AI

The hardware platform on which an AI model is executed is a fundamental determinant of its energy efficiency. The choice of hardware represents a trade-off on a spectrum from generality to efficiency:

- **CPUs and GPUs:** CPUs offer maximum flexibility but have low parallelism, making them highly inefficient for deep learning workloads. GPUs, with their Single Instruction, Multiple Threads (SIMT) architecture, provide the massive parallelism necessary for training and are the workhorse of Red AI, but they are power-intensive, often consuming $>100W$.
- **ASICs (e.g., TPUs):** Application-Specific Integrated Circuits (ASICs), such as Google's Tensor Processing Unit (TPU), are custom-designed for a specific task—in this case, matrix computations. This specialization makes them highly efficient for AI inference, offering superior throughput and energy efficiency (Giga-Operations per Second per Watt, or GOPS/W) compared to GPUs, but at the cost of programmability.

- **FPGAs (Field-Programmable Gate Arrays):** FPGAs offer a compromise. They consist of programmable logic blocks and can be reconfigured to create a custom datapath for a specific AI model. This allows them to exploit model-specific optimizations like sparsity (see Section II.A.1) and achieve significantly lower power consumption (often $<10W$) [10]. Studies have shown FPGAs can achieve substantially higher energy efficiency (e.g., 6.54x) than GPUs for the same workload.
- **Neuromorphic Computing:** This represents a complete paradigm shift. Neuromorphic chips are brain-inspired, *non-Von Neumann* architectures [11]. Unlike traditional systems that shuttle data between separate CPU and memory units (the "Von Neumann bottleneck"), neuromorphic systems are event-driven (asynchronous) and compute using "spikes". They consume power only when a neuron "fires," offering a path toward ultra-low-power, on-device, and continuous learning.

Table 2- Efficiency-Performance of AI Hardware Accelerators

Platform	Architecture	Programmability	Typical Power (W)	Efficiency (GOPS/W)
CPU	Von Neumann	High	10^2	1-10
GPU	Von Neumann	High	>10^2	10^3
FPGA	Von Neumann	Medium	<10	10-100
ASIC (TPU)	Von Neumann	Low	10^2	>10^3
Neuromorphic	Non-Von Neumann	Low (Specialized)	Ultra-Low (mW)	N/A (Event-based)

3. System-Level Optimization: Efficient Inference

The inference phase, as the dominant lifecycle cost, requires a dedicated set of system-level optimizations. The key metrics for evaluating inference are distinct from training; they include *throughput* (examples/sec), *latency* (time/example), *model size* (memory), and *energy use* (Joules/example).

A primary strategy is "Green AI by design," which involves selecting an efficient model architecture from the outset, rather than attempting to compress a large "Red AI" model. Lightweight architectures such as MobileNet, SqueezeNet, EfficientNet, and ShuffleNet are designed specifically for high efficiency and are foundational for on-device and edge AI [12].

Deploying these models in resource-constrained environments (e.g., IoT, mobile) requires a holistic "Edge AI" approach. This involves a co-design of the compressed model, the compiler, and the hardware [14]. An often-overlooked factor is the AI runtime itself. Studies show that the choice of software framework (e.g., PyTorch vs. TensorFlow) and interchange format (e.g., ONNX) can have a significant and unpredictable impact on energy consumption, even when running the *same* model on the *same* hardware [15]. This highlights that the entire software stack, not just the model, must be optimized for a true Green AI solution.

II.B. A Unified Analysis of Carbon Cost vs. Performance Across the AI Lifecycle

This section provides the core quantitative synthesis of the paper, linking the optimization techniques from Section II.A to their measured impact on lifecycle costs and performance.

1. Disambiguating Lifecycle Costs: Training vs. Inference

The "Red AI" paradigm has been associated with a "pro-computation" stance, where proponents of foundation models claim the enormous, one-time training cost is amortized by the model's broad applicability and low-cost reuse [16]. This analysis, however, identifies this as the "Amortization Fallacy."

The inference cost, while small per-query, is not negligible when scaled. A recent infrastructure-aware benchmark of 30 state-of-the-art LLMs provides stark quantitative evidence.

- A single *short* query to GPT-4o consumes 0.42 Wh.
- A *long* query (e.g., 10k input tokens) to a model like DeepSeek-R1 consumes over 33 Wh.
- Scaling the "small" 0.42 Wh query to 700 million queries per day—a reasonable estimate for a global service—results in a projected annual environmental impact of:
  - **Electricity:** Equivalent to 35,000 U.S. homes.
  - **Water:** Matches the annual drinking needs of 1.2 million people.
  - **Carbon:** Requires a Chicago-sized forest to offset.

This data irrefutably establishes that the cumulative operational cost of inference, not the one-time training cost, is the paramount challenge for sustainable AI.

2. Architectural Cost-Performance Trade-offs

The single most significant factor in an AI system's carbon footprint is the high-level *architectural choice*—the "right tool for the job." Using a large, general-purpose model for a task that a smaller, task-specific model can perform is the quintessential "Red AI" anti-pattern.

- **Generative vs. Task-Specific Models:** Using multi-purpose generative models for simple discriminative tasks is "orders of magnitude more expensive" than using task-specific models [17]. One study found that general-purpose AI can use 20 to 30 times more energy. A direct quantitative comparison for text classification showed:
  - **Task-Specific:** bert-base-multilingual-uncased-sentiment emits **0.32g** CO_2e per 1,000 queries.
  - General-Purpose: BLOOMz-7B (a large generative model) used for the same task emits 4.67g CO_2e per 1,000 queries. This represents a 14.6-fold increase in the carbon footprint for no discernible performance benefit on that task.
- **Diffusion Model Architectures:** This trend holds for image generation. An empirical study of 17 state-of-the-art image generation models found that energy consumption varies "drastically," with up to a **46-fold** difference between models [18]. A key finding from this study is that models using **U-Net-based backbones tend to consume less energy** than those using newer, Transformer-based backbones. This suggests that newer, more complex architectures are not axiomatically more efficient.

3. Quantifying the Impact of Green AI Techniques: A Lifecycle Synthesis

When the *correct* architecture is chosen, the Green AI techniques from Section II.A can offer substantial, measurable savings.

- **Case Studies in Compression:**
  - **Transformers:** Applying joint pruning and knowledge distillation to a BERT model reduced energy consumption by **32.097%** while maintaining 95.9% accuracy [19]. In a separate study, static quantization applied to other transformer models achieved a **29.14%** energy saving.
  - **Classical Models:** The impact is even more dramatic on non-neural models. One study on tree-based ensembles (e.g., Random Forest) found that aggressive pruning achieved a **97.6% reduction in carbon emissions** while maintaining 94.5% of baseline accuracy [20]. This highlights that for many tasks, avoiding deep learning entirely is the "greenest" choice.
- **Synthesis:** Even *within* a model class, the choice of a Green AI-focused model is critical. The same 2025 inference benchmark revealed a **greater than 70-fold** energy-per-query difference between the most energy-intensive LLM (DeepSeek-R1, ~33 Wh) and the most energy-efficient (GPT-4.1 nano, ~0.45 Wh).

Holistic approaches that combine optimizations across the entire pipeline—from data selection to model architecture, training, and inference—can achieve compounding returns, with studies demonstrating energy reductions of up to **94.6%** [21]. Table III provides a unified synthesis of these cost-performance trade-offs.

Table 3- Unified Lifecycle Cost vs. Performance Benchmarks

Model / Architecture	Lifecycle Stage	Performance Metric	Energy Cost (MWh or Wh/query)	Carbon Cost (tons CO2e or g CO2 e/query)
GPT-3	Training	N/A (Foundation)	1,287 MWh	550 tons CO_2e
DeepSeek-R1	Inference (Long)	N/A	33.634 Wh / query	>14 g CO_2e / query
GPT-4.1 nano	Inference (Long)	N/A	0.454 Wh / query	<0.3 g CO_2e / query
Finetuned BERT	Inference (1k queries)	Task-Specific Accuracy	N/A	0.32 g CO_2e / 1k queries
BLOOMz-7B	Inference (1k queries)	Task-Specific Accuracy	N/A	4.67 g CO_2e / 1k queries
Pruned Random Forest	Inference	94.5% Baseline Accuracy	N/A	97.6% carbon reduction vs. baseline
BERT + Pruning/KD	Inference	95.9% Baseline Accuracy	32.097% energy reduction vs. baseline	N/A

C. Toward a Comprehensive Framework for Green AI Reporting

The analysis in Section II.B is possible only by synthesizing disparate, recent benchmarks. It is not the standard. The field of Green AI is currently hampered by opaque, inconsistent, and non-standardized reporting [7]. This lack of transparency, where companies "report whatever they choose" , makes validated comparisons impossible and creates a significant risk of "greenwashing."

1. Limitations of Current Metrics and Tools

A "measurement crisis" exists at the foundation of Green AI reporting. While practitioner-focused tools like *CodeCarbon* and *ML CO2 Impact* [22] are commendable for raising awareness, ground-truth validation studies have revealed their significant inaccuracies.

- **The "Tools" Gap:** A 2025 systematic evaluation [3] compared these software-based tools against external, ground-truth power meters. The results were alarming:
  - The tools produced "errors of **up to 40%**".
  - *CodeCarbon*, which uses dynamic profiling, *consistently* **underestimates** true energy

consumption by **20-30%**.
    - The reason for this discrepancy is that these tools "neglect hardware components" that they cannot profile from software, such as the power supply unit (PSU), cooling systems, and peripherals. This cooling and system overhead can account for over 10% of the total AI power consumption. This data indicates that most current Green AI benchmarks are likely under-reporting the true environmental cost.
- **The "Embodied Carbon" Gap:** Current reporting almost exclusively focuses on *operational* carbon (the electricity used, or Scope 2 emissions) [23]. This ignores two other critical factors:
    1. **Embodied Carbon:** The massive carbon footprint generated during the *manufacturing* of the specialized hardware (e.g., GPUs, TPUs) [24].
    2. **E-Waste:** The environmental cost of hardware disposal, as rapid innovation cycles lead to tons of servers and AI chips ending up in landfills.
- **The "Hidden Costs" Gap:** Reporting also omits other environmental impacts, most notably the *water footprint*. Training GPT-3 was estimated to consume 700 kiloliters of water for cooling , and global AI-related water withdrawals are a growing concern.

2. A Proposed Lifecycle Reporting Framework

To address these gaps, we propose a multi-dimensional, lifecycle-based framework for transparent and comparable Green AI reporting. This framework is built on three pillars:

- **Pillar 1: Full Lifecycle Scope.** Reporting must be comprehensive, breaking down costs across the entire AI lifecycle, as defined by: (1) Data Collection and Storage, (2) Model Training, (3) Model Inference (Operational Cost), (4) Hardware (Embodied Carbon), and (5) Disposal (E-Waste). The assessment must cover the *full system* power, including host CPU, memory, and cooling, not just the AI accelerator.
- **Pillar 2: Integration with Global Standards.** To ensure corporate and regulatory compatibility, Green AI metrics must integrate with established environmental accounting standards. This framework does not replace, but rather provides a domain-specific application of:
    - **GHG Protocol (Scopes 1, 2, 3):**
    - **ISO 14064:** Greenhouse gas quantification and reporting.
    - **ISO 14068-1:** Climate change management and carbon neutrality.
- **Pillar 3: Unified Cost-Performance Metrics.** This is the framework's core mechanism for enabling true cost-benefit analysis. We propose the mandatory reporting of *unified metrics* that bind environmental cost to model performance, moving the field toward "eco-efficiency" ranking [4]. This forces a transparent answer to the question: "What is the environmental cost of this model's accuracy?"

Proposed metrics include:

- **For Training:**
    - Total Energy (MWh) / Model Performance (e.g., MMLU Score)
    - Total Carbon (tons CO_2e) / Model Performance
    - Total Water (kL) / Model Performance
- **For Inference:**
    - Energy per Unit (Wh) / 1,000 Inferences
    - Carbon per Unit (g CO_2e) / 1,000 Inferences
    - Water per Unit (mL) / 1,000 Inferences
    - Compute Carbon Intensity (CCI)

By standardizing on such metrics, this framework moves Green AI from a vague ideal to a quantitative engineering discipline, enabling true, data-driven trade-offs between performance and sustainability.

3. CONCLUSION

The environmental footprint of AI, once a niche concern, has become a central challenge to the field's sustainable growth. This paper has provided a three-part contribution to the field of Green AI.

First, we systematically categorized the technical levers for Green AI, covering Software-level optimizations (pruning, quantization, KD) , Hardware-level choices (GPU, TPU, FPGA, Neuromorphic) , and System-level strategies (lightweight architectures, efficient runtimes) [13].

Second, we presented a novel quantitative synthesis (Table III) that fills a critical research gap. This analysis provided two key findings:

1. **Inference Dominance:** The dominant lifecycle cost is not training, but the cumulative, operational cost of inference, which can account for 90% of energy use and, at scale, has an environmental impact

equivalent to entire cities.

2. **Architectural Impact:** The most critical factor for sustainability is the high-level *architectural choice*. We quantified that using general-purpose generative models for task-specific work is orders of magnitude (14.6x to 46x) less efficient [9] and that efficiency varies by over 70x even among LLMs.

Third, we identified a "measurement crisis," demonstrating that common reporting tools *underestimate* the true energy cost of AI by 20-40% [2]. To address this, we proposed a comprehensive, lifecycle-based reporting framework. This framework mandates integration with global GHG/ISO standards and, most importantly, the adoption of unified cost-performance metrics [16] to enable transparent, quantitative trade-offs.

Future work must proceed with urgency in four key areas:

1. **Validating Tools:** Researchers must focus on ground-truth validation and improvement of open-source measurement tools, enhancing them to account for full-system power, including cooling and peripherals [3].
2. **Quantifying Hidden Costs:** The field must move beyond operational carbon to systematically measure and report on the "hidden" environmental costs of AI, particularly the embodied carbon of hardware [24], water consumption , and e-waste.
3. **Investigating Anomalies:** Counter-intuitive findings, such as the *deterioration* of energy efficiency from quantization in some generative models, must be investigated thoroughly.
4. **Mandating Transparency:** We call for standards bodies (e.g., IEEE) and leading conference organizers (e.g., NeurIPS, ICML) to adopt this paper's proposed framework, making transparent, multi-dimensional environmental reporting a mandatory component of all AI research and deployment.

References

1. General Purpose AI Uses 20 to 30 Times More Energy than Task-Specific AI - Proof News, https://www.proofnews.org/general-purpose-ai-uses-20-to-30-times-more-energy-than-task-specific-ai/
2. How AI Uses Energy - Third Way, https://www.thirdway.org/memo/how-ai-uses-energy
3. How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference - arXiv, https://arxiv.org/pdf/2505.09598
4. Smarter sustainability: How technology can transform climate metrics and disclosure, https://ccli.ubc.ca/smarter-sustainability-how-technology-can-transform-climate-metrics-and-disclosure/
5. Measuring AI's Energy/Environmental Footprint to Access Impacts, https://fas.org/publication/measuring-and-standardizing-ais-energy-footprint/
6. Lower Numerical Precision Deep Learning Inference and Training - Intel, https://www.intel.com/content/dam/develop/external/us/en/documents/lower-numerical-precision-deep-learning-jan2018-754765.pdf
7. PQK: Model Compression via Pruning, Quantization, and Knowledge Distillation - arXiv, https://arxiv.org/abs/2106.14681
8. A Survey on Neural Network Hardware Accelerators - IEEE Computer Society, https://www.computer.org/csdl/journal/ai/2024/08/10472723/1ViYSMvUFI4
9. Inference, Low-Cost Models, and Compression - CS@Cornell, https://www.cs.cornell.edu/courses/cs6787/2018fa/Lecture11.pdf
10. Edge Intelligence: A Review of Deep Neural Network Inference in ..., https://www.mdpi.com/2079-9292/14/12/2495
11. Towards a Methodology and Framework for AI Sustainability Metrics - HotCarbon, https://hotcarbon.org/assets/2023/pdf/a13-eilam.pdf
12. he Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation, https://www.researchgate.net/publication/392918101_The_Hidden_Cost_of_an_Image_Quantifying_the_Energy_Consumption_of_AI_Image_Generation
13. [2506.17016] The Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation - arXiv, https://arxiv.org/abs/2506.17016
14. The Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation, https://arxiv.org/html/2506.17016v1
15. Comparative analysis of model compression techniques for ..., https://pubmed.ncbi.nlm.nih.gov/40604122/
16. Energy-Efficient Transformer Inference: Optimization Strategies for Time Series Classification - arXiv, https://arxiv.org/html/2502.16627v4
17. Energy-Efficient Transformer Inference: Optimization Strategies for Time Series Classification - arXiv, https://arxiv.org/pdf/2502.16627

18. Reducing Carbon Footprint of Machine Learning Through Model ..., https://www.ijisrt.com/assets/upload/files/IJISRT25AUG970.pdf
19. mlco2/codecarbon: Track emissions from Compute and recommend ways to reduce their impact on the environment. - GitHub, https://github.com/mlco2/codecarbon
20. How to estimate and reduce the carbon footprint of machine learning models, https://towardsdatascience.com/how-to-estimate-and-reduce-the-carbon-footprint-of-machine-learning-models-49f24510880/
21. Ground-Truthing AI Energy Consumption: Validating CodeCarbon Against External Measurements - arXiv, https://arxiv.org/pdf/2509.22092
22. Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends - arXiv, https://arxiv.org/html/2502.01671v1
23. Sustain AI: A Multi-Modal Deep Learning Framework for Carbon Footprint Reduction in Industrial Manufacturing - MDPI, https://www.mdpi.com/2071-1050/17/9/4134
24. Criteria for Credible AI-assisted Carbon Footprinting Systems: The Cases of Mapping and Lifecycle Modeling - arXiv, https://arxiv.org/html/2509.00240v1