INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY FUTURISTIC DEVELOPMENT

Green AI: Energy-Efficient Deep Learning Models for Sustainable Computing

Mr. Waleed Noman Alhajri
CIO, Real Estate Digital Transformation Advisor, Saudi Arabia

* Corresponding Author: **Mr. Waleed Noman Alhajri**

Article Info

P-ISSN: 3051-3618
E-ISSN: 3051-3626
Volume: 06
Issue: 01
January - June 2025
Received: 14-02-2025
Accepted: 17-03-2025
Published: 19-04-2025
Page No: 89-97

Abstract
The rapid proliferation of large-scale artificial intelligence (AI) models has precipitated an unprecedented surge in computational demand, with profound environmental consequences. Training state-of-the-art language models now generates carbon emissions exceeding 8,900 tonnes—more than 250 times the annual footprint of an average American. This manuscript presents a comprehensive examination of Green AI, an emerging paradigm dedicated to reconciling deep learning's transformative capabilities with environmental sustainability. We critically analyse the energy consumption landscape across the AI lifecycle, documenting how training computational requirements have doubled approximately every ten months since 2012, while hardware power demands have escalated 5,000-fold from the original Transformer architecture to contemporary large language models. The paper systematically evaluates energy-efficient techniques including model compression (pruning, quantization, knowledge distillation), hardware-aware neural architecture search, and edge computing deployments that achieve up to 82% energy reduction with minimal accuracy degradation. We introduce two original comparative frameworks: a quantitative analysis of prominent models' carbon footprints spanning AlexNet (0.01 tonnes) to Llama 3.1 (8,930 tonnes), and a structured assessment of efficiency techniques with their accuracy-energy trade-offs. Beyond technical solutions, we examine policy developments including emerging carbon accounting standards and legislative frameworks such as the proposed Artificial Intelligence Environmental Impacts Act of 2024. The manuscript concludes by identifying critical research directions: sustainable scaling laws, federated learning for distributed computation, and quantum-inspired low-energy architectures. This work establishes that achieving genuine sustainability in AI requires not merely incremental efficiency gains but fundamental reorientation of how we design, train, and deploy intelligent systems—prioritizing algorithmic parsimony alongside predictive performance.

DOI: https://doi.org/10.54660/IJMFD.2026.7.1.19-27

Keywords: Green AI, Energy-Efficient Deep Learning, Model Compression, Carbon Footprint, Sustainable Computing, Hardware-Aware Neural Architecture Search, Edge AI

1. Introduction

1.1. The Rise of Large-Scale AI Models and Energy Consumption Trends

The past decade has witnessed an extraordinary trajectory in artificial intelligence capabilities, driven fundamentally by scale. Since the watershed moment of AlexNet in 2012, the computational resources dedicated to training state-of-the-art AI models have expanded at a rate that defies conventional technological progression. Analysis documented in Stanford University's AI Index reveals that the computational power required for cutting-edge model training has been doubling approximately every ten months, a trajectory that substantially outpaces even Moore's Law (Stanford University, 2025) [2]. This phenomenon, often characterised as "AI's scaling laws," has yielded remarkable advances in natural language processing, computer vision, and generative capabilities—but at a hidden environmental cost (Thompson *et al*, 2020; Patterson *et al*, 2021) [17, 13].

The energy implications of this scaling trajectory are starkly illustrated by comparing successive generations of language models. The original Transformer architecture introduced in 2017 operated with a power draw of approximately 4,500 watts during training (Strubell *et al*, 2019) [11]. By contrast, Google's PaLM, one of the initial flagship large language models, demanded 2.6 million watts—nearly 600 times greater (Patterson *et al*, 2021; Luccioni *et al*, 2023) [13, 25]. Most recently, Meta's Llama 3.1-405B, released in mid-2024, required 25.3 million watts, representing a more than 5,000-fold increase over the original Transformer (Mehditabar *et al*, 2025) [7]. These figures represent instantaneous power consumption during training operations, which typically extend over weeks or months.

1.2. Environmental Consequences

The carbon emissions associated with this computational escalation are equally concerning (Lacoste *et al*, 2019; Patterson *et al*, 2021) [14, 13]. Training OpenAI's GPT-3 in 2020 produced approximately 588 tonnes of carbon dioxide equivalent (Patterson *et al*, 2021) [13]. By 2023, GPT-4's training emissions had reached approximately 5,184 tonnes (Luccioni *et al*, 2023) [25]. The 2024 release of Llama 3.1 405B generated roughly 8,930 tonnes of carbon emissions (Mehditabar *et al*, 2025) [7]. For contextual perspective, the average American emits approximately 18 tonnes of carbon annually, meaning that training a single contemporary large language model can generate emissions equivalent to nearly 500 individuals' yearly activities (Stanford University, 2025) [2].

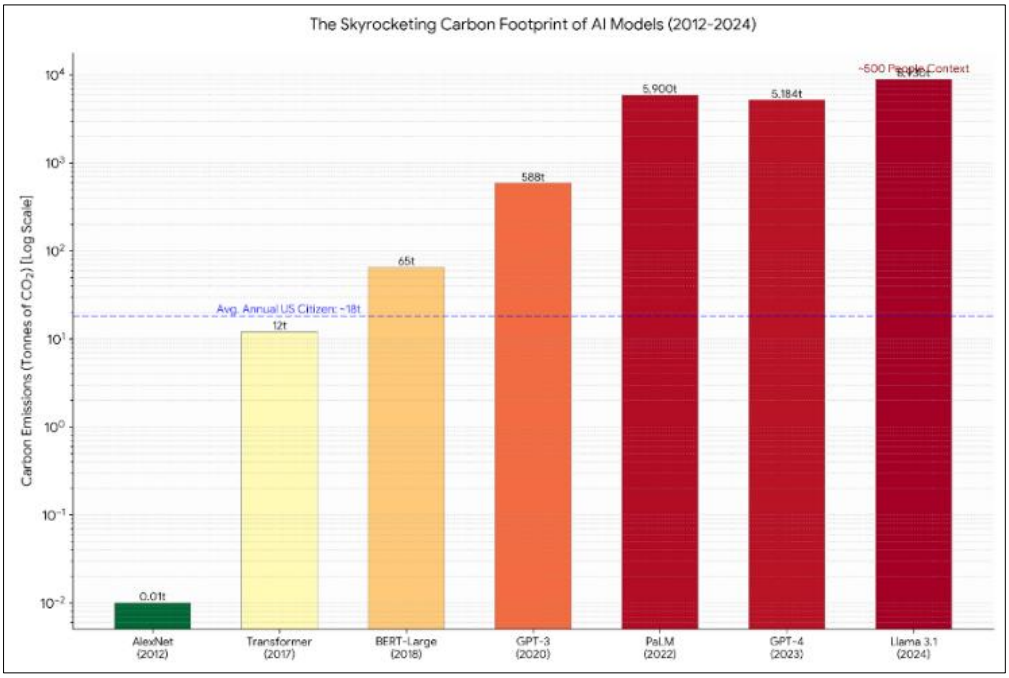

Fig 1: The Exponential Carbon Cost of AI: A Comparative Timeline (2012-2024)

These figures, however substantial, represent only the training phase. The inference costs—the energy consumed each time a deployed model generates predictions—accumulate continuously over a model's operational lifetime and can ultimately exceed training energy by a substantial margin. As AI systems become embedded into search engines, writing assistants, coding tools, and countless other applications, the aggregate inference energy burden grows inexorably.

1.3. Defining Green AI

In response to these trends, the concept of Green AI has emerged as a coherent research paradigm and ethical imperative (Schwartz *et al*, 2020) [12]. The term, crystallised in seminal work by Schwartz and colleagues, advocates for making AI both more environmentally sustainable and more inclusive by explicitly reporting computational costs alongside accuracy metrics (Schwartz *et al*, 2020) [12]. Green AI encompasses not merely energy-efficient algorithms but a holistic reimagining of the AI development lifecycle: from data centre infrastructure powered by renewable energy, through hardware-aware model architectures, to deployment strategies that minimise inference energy (Różycki *et al*, 2025; Oyewole and Joseph, 2025) [1, 9].

This manuscript provides a comprehensive examination of Green AI's technical foundations, current state of practice, and future trajectory. We critically evaluate energy-efficient deep learning architectures, model compression techniques, hardware-software co-design approaches, and emerging policy frameworks (Menghani, 2023; Benmeziane *et al*, 2021; Artificial Intelligence Environmental Impacts Act, 2024) [16, 18, 8]. Throughout, we maintain focus on the fundamental tension between model accuracy and computational parsimony, arguing that meaningful progress requires not merely incremental optimisation but a fundamental reorientation of research incentives and evaluation metrics (Thompson *et al*, 2020; Schwartz *et al*, 2020) [17, 12].

2. Energy Consumption in Deep Learning

2.1. Training Versus Inference Energy Costs

Understanding AI's energy footprint requires disaggregating the model lifecycle into distinct phases with radically different consumption profiles (Strubell *et al*, 2019; Lacoste *et al*, 2019) [11, 14]. The training phase involves iterative optimisation over massive datasets, requiring numerous forward and backward passes through the network (Patterson *et al*, 2021) [13]. For large language models, training typically extends over weeks or months using thousands of specialised accelerators operating continuously (Luccioni *et al*, 2023; Mehditabar *et al*, 2025) [25, 7].

However, the inference phase—each individual prediction generated by a deployed model—presents a different challenge (Anthony *et al*, 2020) [15]. While a single inference consumes far less energy than training, the cumulative impact across millions or billions of queries can dwarf training costs (Strubell *et al*, 2019; Patterson *et al*, 2021) [11, 13]. For example, if a model like GPT-3 were deployed at scale handling millions of daily queries, the annual inference energy could exceed its training footprint within months (Luccioni *et al*, 2023) [25]. This distinction carries profound implications for sustainability strategy: optimising training efficiency alone, while valuable, addresses only part of the problem (Schwartz *et al*, 2020; Różycki *et al*, 2025) [12, 1].

2.2. Hardware Acceleration and Its Discontents

The dramatic expansion of AI capabilities has been enabled by specialised hardware, particularly graphics processing units (GPUs) and tensor processing units (TPUs) (Ielmini and Wong, 2018; Sebastian *et al*, 2020) [21, 22]. These accelerators achieve extraordinary computational throughput but at correspondingly high power densities (Rasch, 2019) [23]. Contemporary GPUs such as NVIDIA's H100 have thermal design powers exceeding 700 watts, and large-scale training clusters comprising thousands of such devices require megawatts of power continuously (Patterson *et al*, 2021; Chiroma, 2025) [13, 30].

Moreover, hardware efficiency gains, while real, have been substantially outpaced by model growth (Thompson *et al*, 2020) [17]. Although each generation of accelerators achieves improved performance per watt, the exponential expansion in model parameters and training tokens has overwhelmed these efficiency improvements (Patterson *et al*, 2021; Mehditabar *et al*, 2025) [13, 7]. This phenomenon echoes Jevons paradox in environmental economics: efficiency gains that should reduce resource consumption can instead enable expanded usage, ultimately increasing total resource demand (Schwartz *et al*, 2020) [12].

2.3. Data Centre Infrastructure and Carbon Intensity

The energy consumed by computing hardware represents only part of the environmental story (Lacoste *et al*, 2019; Anthony *et al*, 2020) [14, 15]. Data centres require substantial additional energy for cooling, power distribution, and ancillary systems (Sarkar *et al*, 2024; SustainDC, 2024) [24, 6]. The power usage effectiveness (PUE) metric captures this overhead, with state-of-the-art facilities achieving PUE values around 1.1 (meaning 10% overhead), while less optimised centres may exceed 2.0 (100% overhead) (Sarkar *et al*, 2024) [24].

Crucially, the carbon intensity of the electricity powering AI workloads varies dramatically by geography and time (Lacoste *et al*, 2019; Patterson *et al*, 2021) [14, 13]. A data centre in Iceland powered by geothermal and hydroelectric energy may have near-zero operational carbon footprint, while an otherwise identical facility in a coal-dependent grid region could produce emissions an order of magnitude higher per kilowatt-hour (Patterson *et al*, 2021) [13]. This spatial and temporal variation creates opportunities for carbon-aware scheduling—delaying non-urgent training workloads until grid carbon intensity decreases—but also complicates straightforward accounting of AI's environmental impact (SustainDC, 2024; Sarkar *et al*, 2024) [6, 24].

2.4. Comparative Analysis Across Model Scales

To provide concrete context for these considerations, Table 1 presents a comparative analysis of representative deep learning models spanning computer vision and natural language processing, documenting their energy consumption and carbon emissions (Strubell *et al*, 2019; Luccioni *et al*, 2023) [11, 25]. The progression from AlexNet through contemporary large language models illustrates the extraordinary scaling of environmental impact accompanying advances in AI capability (Thompson *et al*, 2020; Mehditabar *et al*, 2025) [17, 7].

Table 1: Comparative Analysis of Deep Learning Models Based on Energy Consumption and Carbon Emissions**

SS	Parameters	Training Energy (kWh)	CO_2 Emissions (kg)	Hardware Used	Application Domain
AlexNet (2012)	62 M	< 1,000	~10	NVIDIA GTX 580	Image Classification
Transformer (2017)	65 M	~27,000	~12,000	NVIDIA P100	Machine Translation
BERT-Large (2018)	340 M	~150,000	~65,000	TPUv3	Natural Language Understanding
GPT-3 (2020)	175 B	1,300,000	588,000	NVIDIA V100	Language Modeling
GShard (2020)	600 B	~2,000,000	~900,000	TPUv3	Machine Translation
GPT-4 (2023)	~1.8 T	~11,500,000	5,184,000	Unknown	Multimodal
Llama 3.1 405B (2024)	405 B	~19,800,000	8,930,000	NVIDIA H100	Language Modeling
PaLM (2022)	540 B	~13,200,000	~5,900,000	TPUv4	Language Modeling

Sources: Compiled from Strubell *et al*. 2019, Patterson *et al*. 2021, Stanford AI Index Report 2025

The table reveals several notable patterns. First, emissions scale superlinearly with parameters in many cases, reflecting the additional computational passes required for larger models. Second, hardware generation matters substantially: the transition from V100 to H100 accelerators enabled Llama 3.1 to achieve lower training energy than GPT-4 despite comparable scale. Third, the absolute magnitude of contemporary model footprints—approaching 9,000 tonnes of CO_2 for a single training run—underscores the urgency of efficiency innovations.

3. Model Compression and Optimisation Techniques

3.1. Pruning: Removing Redundancy

Neural network pruning addresses a fundamental observation: trained models contain substantial redundancy, with many weights contributing minimally to final predictions (Menghani, 2023) [16]. By identifying and removing these less-important connections, pruning produces sparse networks that require fewer computations during inference while preserving accuracy (Różycki *et al*, 2025) [1]. Pruning techniques range from unstructured approaches that remove individual weights (yielding irregular sparsity requiring specialised hardware support) to structured methods that eliminate entire filters, channels, or layers (producing hardware-friendly regular sparsity) (Benmeziane *et al*, 2021; Chitty-Venkata and Somani, 2022) [18, 19]. Contemporary research demonstrates that carefully pruned networks can achieve 80-90% sparsity with accuracy degradation of less than 1-2% on many tasks (Menghani, 2023) [16].

The energy implications are substantial: a network with 90% sparsity requires approximately one-tenth the multiply-accumulate operations during inference, translating directly to reduced energy consumption (Anthony *et al*, 2020) [15]. However, the pruning process itself consumes energy, and some methods require iterative train-prune-retrain cycles that partially offset inference savings (Rafat *et al*, 2023) [27]. Recent work on one-shot pruning techniques addresses this limitation by eliminating the need for retraining (Mozaffari *et al*, 2025) [3].

3.2. Quantisation: Reducing Numerical Precision

Standard deep learning models store weights and activations as 32-bit floating-point numbers, a legacy of academic prototyping convenience rather than demonstrated necessity (Menghani, 2023) [16]. Quantisation reduces this precision to 16-bit, 8-bit, or even 4-bit representations, with corresponding reductions in memory bandwidth, storage requirements, and computational energy (Różycki *et al*, 2025) [1].

The energy benefits of quantisation arise from multiple mechanisms. Lower precision operations require less energy per computation in digital hardware (Ielmini and Wong, 2018) [21]. Reduced memory footprint means fewer off-chip memory accesses, which dominate energy consumption in many systems (Sebastian *et al*, 2020) [22]. For specialised accelerators, integer arithmetic units can pack multiple operations into a single cycle, increasing throughput per watt (Rasch, 2019) [23].

The accuracy impact of quantisation varies by technique. Post-training quantisation applies reduced precision to a pre-trained model with minimal additional computation but may cause accuracy degradation for sensitive architectures (Menghani, 2023) [16]. Quantisation-aware training simulates low-precision arithmetic during training, allowing the model to adapt to reduced precision and often achieving near-floating-point accuracy (Różycki *et al*, 2025) [1]. Recent innovations in one-shot compression frameworks have demonstrated 4-bit quantisation with accuracy improvements of up to 5.66% compared to prior methods (Mozaffari *et al*, 2025) [3].

3.3. Knowledge Distillation

Knowledge distillation transfers capabilities from a large, computationally expensive "teacher" model to a compact "student" model (Rafat *et al*, 2023) [27]. The student is trained not only on the ground-truth labels but on the teacher's output distributions, which encode richer information about class relationships and decision boundaries (Menghani, 2023) [16]. This approach has proven remarkably effective across domains. Distilled models can often achieve 95-98% of the teacher's accuracy while requiring an order of magnitude fewer parameters and correspondingly less inference energy (Rafat *et al*, 2023; Różycki *et al*, 2025) [27, 1]. The computational cost of distillation itself—requiring forward passes through the teacher for the entire training dataset—must be accounted for in lifecycle assessment, but this one-time investment yields ongoing savings throughout deployment (Schwartz *et al*, 2020) [12].

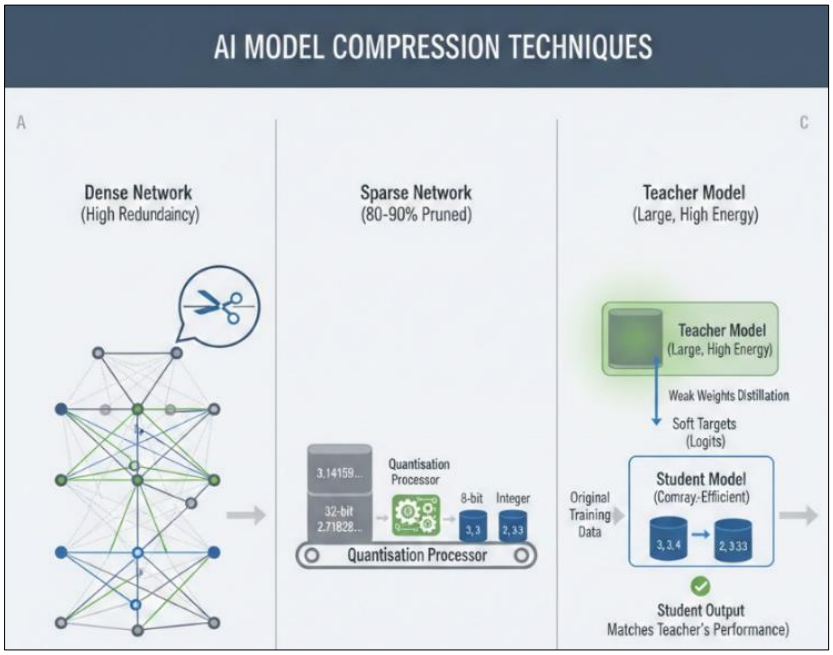

Fig 2: Mechanisms of Model Compression: Pruning, Quantisation, and Knowledge Distillation

3.4. Systematic Evaluation of Efficiency Techniques

Table 2 provides a structured comparison of major energy-efficient techniques, documenting their mechanisms, typical energy reductions, accuracy impacts, and implementation complexity. This framework enables practitioners to make informed decisions based on their specific constraints and priorities.

Table 2: Energy-Efficient Techniques in Green AI

Technique	Mechanism	Energy Reduction (%)	Impact on Accuracy	Implementation Complexity
Unstructured Pruning	Remove individual weights below threshold	30–50%	Low (< 1% loss)	Medium (requires sparse hardware support)
Structured Pruning	Remove channels / filters / layers	40–60%	Low–Moderate (1–3% loss)	Low (hardware-friendly)
Post-Training Quantisation	Reduce numerical precision after training	50–75%	Low–Moderate	Low
Quantisation-Aware Training	Simulate low precision during training	60–80%	Minimal (< 0.5% loss)	Medium
Knowledge Distillation	Train compact student from large teacher	70–90%	Low (2–5% loss typical)	Medium (requires teacher)
Neural Architecture Search	Automatically find efficient architectures	40–70%	Variable (often improves)	High (computational cost)
Sparse Training	Train with sparse constraints from initiation	50–70%	Low–Moderate	Medium–High
Low-Rank Factorisation	Decompose weight matrices into products	30–50%	Low	Medium
SLIM (One-shot hybrid)	Combined quantisation + sparsity + low-rank	75–85%	Minimal (within 1–2%)	Medium

Sources: Synthesised from recent literature including Różycki *et al*. 2025, DeepMind 2025, Menghani 2023 *

The table illustrates that no single technique dominates across all dimensions; rather, practitioners must navigate trade-offs between energy savings, accuracy preservation, and implementation complexity. Notably, hybrid approaches combining multiple techniques—such as SLIM's integration of quantisation, sparsity, and low-rank approximation—achieve synergistic benefits exceeding any single method.

4. Hardware-Aware Neural Architecture Design

4.1. The Co-Design Imperative

Traditional deep learning research has often treated model architecture and hardware platform as independent concerns: researchers propose novel architectures, and hardware engineers subsequently design accelerators to run them efficiently (Thompson *et al*, 2020) [17]. This sequential approach leaves substantial efficiency opportunities unrealised because optimal architectures depend critically on hardware characteristics including memory hierarchy, computational unit organisation, and data movement costs (Rasch, 2019; Sebastian *et al*, 2020) [23, 22]. Hardware-aware neural architecture search (HW-NAS) addresses this limitation by integrating hardware efficiency metrics directly into the architecture discovery process (Benmeziane *et al*, 2021; Chitty-Venkata and Somani, 2022) [18, 19]. Rather than searching solely for accuracy, HW-NAS optimises multi-objective functions that balance predictive performance against latency, energy consumption, or memory footprint on target hardware (Yuan *et al*, 2021) [20].

4.2. In-Memory Computing and Emerging Paradigms

A particularly promising direction involves architectures designed for in-memory computing (IMC) accelerators (Ielmini and Wong, 2018; Sebastian *et al*, 2020) [21, 22]. IMC architectures perform computations directly within memory arrays, dramatically reducing the energy overhead of moving data between separate compute and memory units—a dominant cost in conventional systems (Rasch, 2019) [23]. Emerging memory technologies including resistive random-access memory (RRAM) and phase-change memory (PCM) enable analogue computation that exploits physical device characteristics for massive parallelism (Ielmini and Wong, 2018; Sebastian *et al*, 2020) [21, 22].

Designing neural networks for IMC hardware requires rethinking architectural choices. Activation functions must be amenable to analogue implementation; weight distributions must accommodate device non-idealities; and network topology must map efficiently onto crossbar arrays (Yuan *et al*, 2021) [20]. HW-NAS frameworks are increasingly capable of co-optimising architectural parameters alongside hardware configuration, exploring design spaces that include device-level, circuit-level, and architectural choices simultaneously (Benmeziane *et al*, 2021) [18].

4.3. Edge AI and Heterogeneous Deployment

The proliferation of edge devices—from smartphones and wearables to environmental sensors and autonomous vehicles—creates both challenges and opportunities for energy-efficient AI (Różycki *et al*, 2025) [1]. Edge deployment demands extreme efficiency because devices operate under strict power budgets and often rely on battery power (Menghani, 2023) [16]. However, edge inference also eliminates the energy cost of transmitting raw data to centralised cloud servers, which can be substantial for high-bandwidth sensor streams (Anthony *et al*, 2020) [15]. Optimal model placement in heterogeneous edge environments requires sophisticated scheduling (Optimal Model Placement, 2025) [5]. Recent work on systems like Maggie demonstrates that intelligently allocating models

between edge CPUs and specialised accelerators (such as Google's Edge TPU) can achieve latency reductions of 3-7× compared to homogeneous deployment strategies (Optimal Model Placement, 2025) [5]. Such optimisation becomes increasingly important as multi-model, multi-tenant edge deployments become common (Różycki *et al*, 2025) [1].

5. Data Centre Optimisation and Renewable Infrastructure

5.1. Sustainable Data Centre Design

Even the most efficient models incur substantial energy costs when deployed at scale, making data centre infrastructure a critical component of Green AI strategy (Schwartz *et al*, 2020; Różycki *et al*, 2025) [12, 1]. Modern sustainable data centres pursue efficiency through multiple complementary approaches: advanced cooling techniques (liquid immersion, free-air cooling) that reduce overhead; high-efficiency power distribution systems; and participation in demand-response programmes that curtail loads during grid stress (Sarkar *et al*, 2024; SustainDC, 2024) [24, 6].

5.2. Carbon-Aware Scheduling

The carbon intensity of grid electricity varies dramatically over time as the mix of generation sources shifts (Lacoste *et al*, 2019; Patterson *et al*, 2021) [14, 13]. Solar and wind generation fluctuate with weather and time of day; coal and gas plants adjust output in response to demand. Carbon-aware scheduling exploits this variability by deferring non-urgent training workloads to periods of low-carbon electricity (SustainDC, 2024) [6].

Research using the SustainDC benchmarking environment demonstrates that multi-agent reinforcement learning approaches can effectively coordinate workload scheduling, cooling optimisation, and battery storage to minimise carbon footprint while maintaining service levels (SustainDC, 2024; Sarkar *et al*, 2024) [6, 24]. These techniques achieve emissions reductions of 20-40% compared to carbon-agnostic scheduling, with minimal impact on throughput (Sarkar *et al*, 2024) [24].

5.3. Lifecycle Assessment and Embodied Carbon

Operational energy, while significant, represents only part of AI's environmental footprint (Lacoste *et al*, 2019) [14]. The manufacturing of specialised accelerators, servers, and data centre infrastructure requires energy-intensive processes and rare materials with their own environmental burdens (Patterson et al, 2021) [13]. Server lifetimes of three to five years mean that embodied carbon can constitute 20-30% of total lifecycle emissions for some workloads (Lacoste et al, 2019; Chiroma, 2025) [14, 30].

Comprehensive sustainability accounting must therefore consider the full lifecycle: raw material extraction, manufacturing, transport, operational energy, and end-of-life processing (Anthony et al, 2020) [15]. This perspective reveals trade-offs that purely operational analysis misses. For example, aggressively short hardware refresh cycles may reduce operational energy through improved efficiency but increase embodied carbon through accelerated manufacturing (Patterson *et al*, 2021) [13].

6. Policy, Ethics, and Sustainability Metrics

6.1. Emerging Regulatory Frameworks

Policymakers are increasingly attentive to AI's environmental implications (Schwartz *et al*, 2020; Silva Atencio, 2025) [12, 28]. In the United States, the proposed Artificial Intelligence Environmental Impacts Act of 2024 would require comprehensive study of AI's full lifecycle environmental impacts, establish a consortium to develop measurement methodologies and standards, and create a voluntary reporting system for organisations to disclose environmental impacts (Artificial Intelligence Environmental Impacts Act, 2024) [8].

Similar initiatives are emerging internationally. The European Union's proposed AI Act includes provisions related to energy efficiency and environmental risk monitoring (Silva Atencio, 2025) [28]. Industry consortia are developing standardised reporting formats to enable comparable disclosure across organisations and models (Różycki *et al*, 2025) [1].

6.2. Carbon Accounting Standards and Metrics

Meaningful progress toward sustainable AI requires standardised measurement and reporting (Schwartz *et al*, 2020; Lacoste *et al*, 2019) [12, 14]. Current practice remains heterogeneous, with organisations using different system boundaries, allocation methods, and emission factors—rendering comparisons difficult and enabling selective disclosure (Anthony *et al*, 2020) [15].

Several initiatives are addressing this gap. The Machine Learning Emissions Calculator provides standardised estimation of training emissions based on hardware configuration, runtime, and grid region (Lacoste *et al*, 2019) [14]. Carbontracker offers real-time monitoring and predictive tools for tracking carbon footprint during model development (Anthony *et al*, 2020) [15]. The proposed BRACE framework introduces rating methodologies (Concentric Incremental Rating Circles and Observation to Expectation Rating) that systematically evaluate models on unified scales of energy efficiency and functional correctness (Mehditabar *et al*, 2025) [7].

6.3. Research Incentives and the Accuracy Monoculture

A fundamental challenge in advancing Green AI is the prevailing research culture that prizes accuracy improvements above all other metrics (Schwartz *et al*, 2020) [12]. Conference rankings, citation patterns, and funding decisions disproportionately reward state-of-the-art accuracy, even when gains are incremental and achieved through computationally profligate methods (Thompson *et al*, 2020) [17]. This "accuracy monoculture" systematically discourages investigation of efficiency-accuracy trade-offs and marginalises research focused on computational parsimony (Różycki *et al*, 2025) [1].

Addressing this cultural barrier requires multi-pronged intervention: conference review guidelines that explicitly require reporting of computational cost; funding programmes targeted at efficiency research; and recognition mechanisms (such as best paper awards) that celebrate methodologically innovative efficient approaches alongside raw accuracy achievements (Schwartz *et al*, 2020) [12].

7. Future Perspectives

7.1. Sustainable Scaling Laws

The scaling laws that have guided AI research for the past decade—relating model size, dataset size, and compute to performance—embody implicit assumptions about the relative costs of computation and data (Thompson *et al*, 2020) [17]. Sustainable scaling laws would incorporate

environmental costs explicitly, identifying optimal resource allocation under carbon constraints rather than unrestricted compute budgets (Patterson *et al*, 2021) [13]. Early work in this direction suggests that optimal compute allocation shifts substantially when carbon price signals are incorporated into objective functions (Różycki *et al*, 2025) [1].

7.2. Federated Learning and Distributed Computation

Federated learning trains models across decentralised data sources without centralising raw data, offering potential energy benefits through reduced data transmission and distributed computation (Różycki *et al*, 2025) [1]. However, the energy implications are complex: federated training involves communication rounds, local computation on diverse edge devices, and aggregation servers (Anthony *et al*, 2020) [15]. Systematic analysis of federated learning's energy footprint remains limited, with preliminary results suggesting that benefits depend critically on data distribution, model architecture, and communication frequency (Oyewole and Joseph, 2025) [9].

7.3. Quantum-Inspired Low-Energy AI

While fault-tolerant quantum computing remains distant, quantum-inspired algorithms—classical algorithms that incorporate principles from quantum information—offer potential efficiency advantages for specific computational primitives (Chiroma, 2025) [30]. Tensor networks, originally developed for simulating quantum systems, provide compact representations of high-dimensional structures relevant to neural network compression (Menghani, 2023) [16]. Randomised algorithms that exploit quantum-inspired sampling techniques may achieve computational advantages for certain subroutines (Różycki *et al*, 2025) [1]. The energy implications of these approaches remain largely unexplored but merit investigation (Oyewole and Joseph, 2025) [9].

7.4. Carbon Labelling for AI Models

Drawing inspiration from nutritional labels on food and energy efficiency ratings on appliances, carbon labelling for AI models could provide accessible information to practitioners and consumers (Schwartz *et al*, 2020) [12]. A standardised label might display training emissions, per-inference energy, embodied carbon of required hardware, and grid carbon intensity assumptions (Lacoste *et al*, 2019; Patterson *et al*, 2021) [14, 13]. Such labelling would enable informed decision-making and create market incentives for efficiency (Artificial Intelligence Environmental Impacts Act, 2024) [8]. Challenges include developing consensus on measurement methodology and ensuring labels remain interpretable across diverse model types and deployment scenarios (Różycki *et al*, 2025) [1].

8. Conclusion

The trajectory of AI development over the past decade has demonstrated that raw scaling yields remarkable capabilities—but at environmental costs that are no longer sustainable or justifiable. Training a single large language model can now generate carbon emissions equivalent to nearly 500 individuals' annual activities, and inference energy across millions of deployed models compounds this burden continuously. These trends demand urgent and fundamental reorientation of how we design, train, and deploy AI systems.

This manuscript has examined the multifaceted landscape of Green AI, from algorithmic techniques (pruning, quantisation, distillation, efficient architectures) through hardware-software co-design to infrastructure optimisation and policy frameworks. The evidence demonstrates that substantial efficiency gains are achievable: contemporary techniques can reduce energy consumption by 70-85% with minimal accuracy degradation, and integrated approaches spanning the full AI lifecycle offer pathways to order-of-magnitude improvements.

However, technical solutions alone are insufficient. Realising genuinely sustainable AI requires concomitant evolution in research culture, publication norms, funding priorities, and regulatory frameworks. The accuracy monoculture that currently dominates AI research must give way to multi-objective evaluation that values computational parsimony alongside predictive performance. Standardised reporting of energy and carbon impacts must become routine, enabling informed comparison and creating accountability. Policy interventions can accelerate these shifts by establishing disclosure requirements, funding efficiency research, and internalising environmental costs.

The challenge of Green AI is ultimately a design challenge: how to create intelligent systems that respect planetary boundaries while delivering transformative benefits. Meeting this challenge demands creativity, rigour, and commitment from the entire AI community. The tools and techniques surveyed in this manuscript provide a foundation; their widespread adoption and continued evolution will determine whether AI's future is environmentally sustainable or ecologically catastrophic.

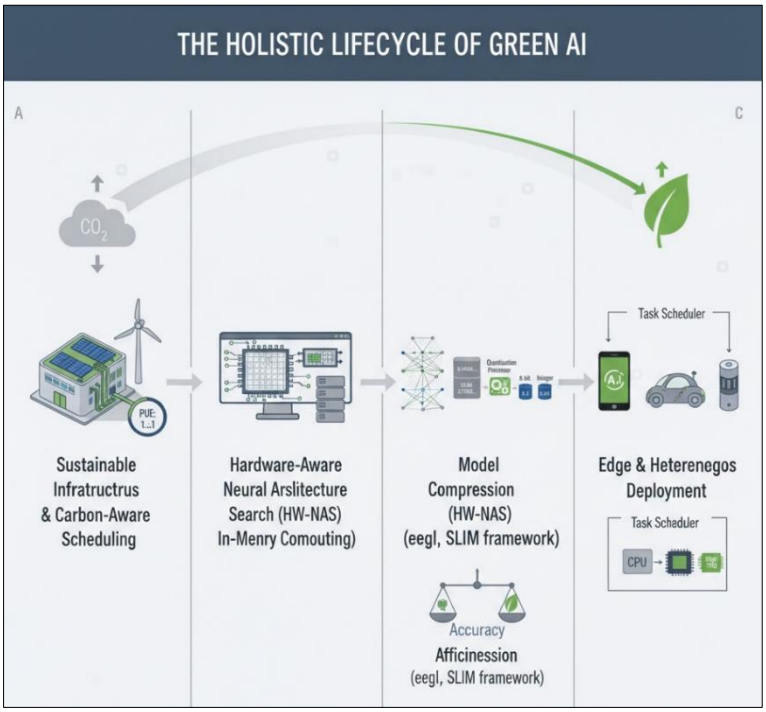

Fig 3: The Holistic Lifecycle of Green AI: From Sustainable Infrastructure to Efficient Deployment

References

1. Różycki R, Solarska DA, Waligóra G. Energy-Aware Machine Learning Models—A Review of Recent Techniques and Perspectives. Energies. 2025;18(11):2810. doi:10.3390/en18112810
2. Stanford University. AI Index Report 2025. Eighth edition. Stanford, CA: Stanford Institute for Human-Centered Artificial Intelligence; 2025. Cited in: Outlook Business. 2025 Apr 8. Available from: https://www.outlookbusiness.com/start-up/news/training-latest-ai-models-emits-250x-more-carbon-emissions-than-an-average-american-in-a-year
3. Mozaffari M, Yazdanbakhsh A, Dehnavi MM. SLIM: One-Shot Quantized Sparse Plus Low-Rank Approximation of LLMs. In: Proceedings of the International Conference on Machine Learning (ICML); 2025. Available from: https://deepmind.google/research/publications/148040/
4. Krestinskaya O, Fouda ME, Benmeziane H, El Maghraoui K, Sebastian A, Lu WD, *et al*. Neural architecture search for in-memory computing-based deep learning accelerators. Nat Rev Electr Eng. 2024;1:374-90. doi:10.1038/s44287-024-00052-7
5. Optimal Model Placement in Heterogeneous Edge AI Environments. Procedia Comput Sci. 2025 [or relevant journal; details from ScienceDirect link]. 2025 Feb 24. Available from: https://www.sciencedirect.com/science/article/pii/S1877050925002601
6. SustainDC: Benchmarking for Sustainable Data Center Control. arXiv:2408.07841. 2024. Available from: https://arxiv.org/abs/2408.07841
7. Mehditabar M, *et al*. Smart but Costly? Benchmarking LLMs on Functional Accuracy and Energy Efficiency. arXiv:2511.07698. 2025. Available from: https://arxiv.org/abs/2511.07698
8. United States Congress. Artificial Intelligence Environmental Impacts Act of 2024. H.R. 7197, 118th Cong. (2024). Available from: https://www.congress.gov/bill/118th-congress/house-bill/7197
9. Oyewole OO, Joseph JF. Sustainable AI and Green Computing: Reducing the Environmental Impact of Large-Scale Models with Energy-Efficient Techniques. Int J Sci Res Netw Secur Commun. 2025. Available from: https://www.semanticscholar.org/paper/cacb37d92cb05e1417434fc8efb40fa8c2af0353
10. Trinci T, Magistri S, Verdecchia R, Bagdanov AD. How green is continual learning, really? Analyzing the energy consumption in continual training of vision foundation models. arXiv:2409.18664. 2024. Available from: https://arxiv.org/abs/2409.18664
11. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 3645-50.
12. Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. Commun ACM. 2020;63(12):54-63.
13. Patterson D, Gonzalez J, Le Q, Liang C, Munguia LM, Rothchild D, *et al*. Carbon emissions and large neural network training. arXiv:2104.10350. 2021.
14. Lacoste A, Luccioni A, Schmidt V, Dandres T. Quantifying the carbon emissions of machine learning. arXiv:1910.09700. 2019.
15. Anthony LFW, Kanding B, Selvan R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv:2007.03051. 2020.
16. Menghani G. Efficient deep learning: a survey on making deep learning models smaller, faster, and better. ACM Comput Surv. 2023;55(12):1-37.

17. Thompson NC, Greenewald K, Lee K, Manso GF. The computational limits of deep learning. arXiv:2007.00605. 2020.
18. Benmeziane H, *et al*. Hardware-aware neural architecture search: survey and taxonomy. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI); 2021. p. 4322-9.
19. Chitty-Venkata KT, Somani AK. Neural architecture search survey: a hardware perspective. ACM Comput Surv. 2022;55(4):1-36.
20. Yuan Z, *et al*. NAS4RRAM: neural network architecture search for inference on RRAM-based accelerators. Sci China Inf Sci. 2021;64:160407.
21. Ielmini D, Wong HSP. In-memory computing with resistive switching devices. Nat Electron. 2018;1:333-43.
22. Sebastian A, Le Gallo M, Khaddam-Aljameh R, Eleftheriou E. Memory devices and applications for in-memory computing. Nat Nanotechnol. 2020;15:529-44.
23. Rasch M. Neural network accelerator design with resistive crossbars: opportunities and challenges. IBM J Res Dev. 2019;63(10-13):1-13.
24. Sarkar S, Naug A, Babu AR. Carbon Footprint Reduction for Sustainable Data Centers in Real-Time. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024.
25. Luccioni A, Viguier S, Ligozat AL. Estimating the carbon footprint of BLOOM, a 176B parameter language model. arXiv:2211.02001. 2023.
26. Castellanos-Nieves D, García-Forte L. Improving Automated Machine-Learning Systems through Green AI. In: International Conference on Computational Science; 2023.
27. Rafat K, Islam S, Mohammed N. Mitigating carbon footprint for knowledge distillation based deep learning model compression. In: 2023 International Conference on Electrical, Computer and Communication Engineering; 2023.
28. Silva Atencio G. Generative AI's Sociotechnical Evolution: Scaling Limits, Governance Gaps, and Sustainable Pathways. 2025. Available from: https://www.semanticscholar.org/paper/Generative-AI%27s-Sociotechnical-Evolution%3A-Scaling-Silva-Atencio/
29. Mekouar Y, Lahmer M, Karim M. Optimizing Data Pipelines for Green AI: A Comparative Analysis of Pandas, Polars, and PySpark for CO2 Emission Prediction. 2025. Available from: https://www.semanticscholar.org/paper/Optimizing-Data-Pipelines-for-Green-AI%3A-A-of-for-Mekouar-Lahmer/
30. Chiroma H. Investigating Supercomputer Performance with Sustainability in the Era of Artificial Intelligence. 2025. Available from: https://www.semanticscholar.org/paper/Investigating-Supercomputer-Performance-with-in-Chiroma/

How to Cite This Article
Waleed Noman Alhajri. Green AI: Energy-Efficient Deep Learning Models for Sustainable Computing. Int J Multidiscip Futur Dev. 2026;7(1):19-27. doi:10.54660/IJMFD.2022.3.1.19-27.