Green AI: A Systematic Review of Energy-Efficient Training and Inference Paradigms for Frontier Models

Dr.Amol Kumbhare[1], Dr. Mahesh Singh Gautam, Mrs Preeti Rajput[2], Ms Priyanka Tandilkar[3]
Dr. A. P. J. Abdul Kalam University, Indore
Corresponding Author dramol@aku.ac.in

Abstract
The rapid proliferation of frontier AI models, particularly Large Language Models (LLMs) and Multimodal Foundation Models, has catalysed unprecedented advancements in artificial intelligence. However, this progress is heavily reliant on resource-intensive scaling laws, leading to unsustainable energy consumption and carbon emissions. Generative AI systems currently consume tens of terawatt-hours (TWh) annually, rivalling the energy demands of small nations. This review paper synthesizes the state-of-the-art methodologies driving the transition from *"Red AI"* (accuracy-prioritized) to *"Green AI"* (efficiency-prioritized). We critically examine the carbon footprint of frontier models (e.g., Llama-3, Qwen2.5, Deep Seek-R1) and evaluate algorithmic interventions across the model lifecycle, focusing on sparse architectures, quantization, pruning, knowledge distillation, and hardware-software co-design. Finally, we highlight open challenges in standardized emissions tracking and the integration of neuromorphic computing.

Keywords: *Green AI; Large Language Models (LLMs); Energy-Efficient Deep Learning; Carbon Footprint Mitigation; Model Quantization; Parameter-Efficient Fine-Tuning (PEFT); Sparse Architectures; Sustainable AI.*

I. INTRODUCTION

For the past decade, AI research has been dominated by empirical scaling laws, which posit that model performance improves predictably as a power-law function of parameter count, dataset size, and compute budget [1]. This paradigm has driven the development of massive models exceeding hundreds of billions of parameters. While these models exhibit emergent reasoning capabilities, their computational overhead has triggered a severe environmental crisis [2].

The concept of Green AI advocates for energy-efficient model architectures, environmentally conscious training paradigms, and optimized inference pipelines. Unlike traditional optimization, which strictly

targets latency or throughput, Green AI explicitly targets the reduction of floating-point operations (FLOPs), memory bandwidth bottlenecks, and overall carbon equivalent emissions (CO_2eq) [1, 3]. This paper provides a comprehensive taxonomy of energy-efficiency techniques actively deployed in 2025–2026 to mitigate the environmental impact of frontier models.

II. THE CARBON FOOTPRINT OF FRONTIER MODELS

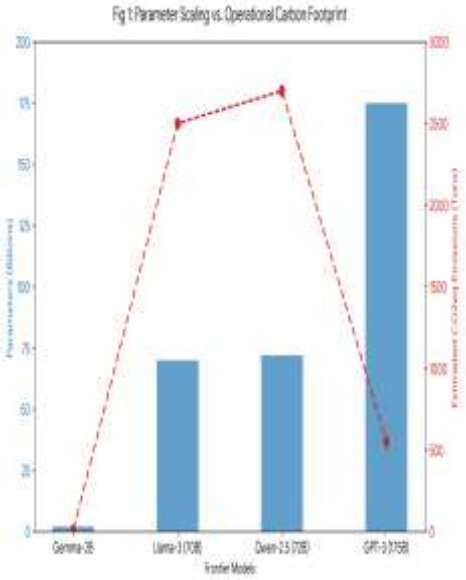
The environmental impact of deep learning is bifurcated into **embodied carbon** (emissions from manufacturing hardware) and **operational carbon** (emissions from electricity consumed during training and inference) [4, 14]. Operational carbon Cop is generally modeled as:

$$Cop = E \times I \times PUE$$

Where E is the energy consumed (in kWh), I is the carbon intensity of the local power grid ($kgCO_2eq/kWh$), and PUE is the Power Usage Effectiveness of the data centre [13].

Recent empirical studies highlight that while inference accounts for the majority of a model's lifetime energy consumption due to continuous deployment, the training phase remains a massive, concentrated emissions event [4]. Furthermore, newer reasoning models such as DeepSeek-R1 generate significantly more internal tokens before

producing an answer, drastically increasing inference energy per query compared to standard text generation [17].



III. .”

Figure 1: A comparative analysis of frontier models depicting parameter size (in billions) against their estimated operational carbon equivalent emissions (CO2eq in tons) during the training phase. While legacy models like GPT-3 exhibit lower emissions relative to their parameter count due to smaller training token budgets, modern dense models (Llama-3, Qwen-2.5) demonstrate a non-linear surge in carbon footprint driven by extreme data scaling.

Table 1: Environmental Cost Comparison of Selected Frontier Models

Model	**Parameters**	**Training Energy (MWh)**	**Estimated CO_2eq (Tons)**
GPT-3 (Baseline)	175B	1,287	552
Gemma-2B-it	2B	< 50	< 20
Llama-3 (Meta) [16]	70B	~6,000	~2,500
Qwen-2.5 [16]	72B	~6,500	~2,700

Key Insight: Global generative AI workloads are now estimated to consume approximately 29.3 TWh annually roughly equivalent to the total electricity consumption of Ireland [14].

III. Energy-Efficient Training Paradigms

Training frontier models requires massive GPU clusters operating for months. Green AI interventions at this stage focus on reducing redundant computations and accelerating convergence [2, 4].

A. 3.1 Sparse Attention and Alternative Architectures

Standard Transformer self-attention scales quadratically with sequence length, exhibiting a time and memory complexity of $O(N^2)$, where N is the context length [6]. To mitigate this, state-of-the-art models are shifting toward:

- Sparse Transformers: Utilizing localized or strided attention masks to reduce complexity to O (N log N).
- State Space Models (SSMs): Architectures like Mamba [5] and RWKV [7] bypass the attention mechanism entirely, offering linear time complexity O(N) and a bounded memory footprint during training, leading to massive energy reductions for long-context tasks.

Flash Attention [6] further reduces memory overhead through IO-aware computation, providing a hardware-level complement to architectural sparsity.

B. 3.2 Parameter-Efficient Fine-Tuning (PEFT)

Training models from scratch is environmentally prohibitive for most institutions. PEFT techniques ensure that only a minuscule fraction of parameters are updated for downstream tasks. Low-Rank Adaptation (LoRA) [8] decomposes weight updates into low-rank matrices, reducing the trainable parameter count by over 98% and decreasing GPU memory and energy consumption

by up to 65% during fine-tuning. QLoRA [9] further extends this by enabling efficient fine-tuning of already-quantized models, compounding savings in both memory and compute.

IV. Inference Optimization Techniques

Because models are deployed globally and queried billions of times daily, inference optimization is the most critical frontier for carbon mitigation [14].

C. 4.1 Post-Training Quantization (PTQ)

Quantization reduces the precision of a model's weights and activations from 16-bit floating-point (FP16) to lower-bit integer formats (e.g., INT8, INT4) [10, 11].

Mechanism: By compressing the mathematical representation of weights, quantization reduces memory access costs the primary energy sink in modern GPUs and allows the use of highly efficient integer arithmetic logic units (ALUs).

Impact: Recent studies demonstrate that applying INT4 quantization to a 70B parameter model reduces inference energy consumption and latency by up to 55%, with negligible degradation in benchmark accuracy (typically <1% drop in F1 scores) [10].

Activation-aware quantization methods such as AWQ [11] further improve robustness by identifying and protecting salient weight channels during compression.

D. 4.2 Network Pruning

Pruning identifies and removes redundant or "unimportant" weights within a neural network [12].

- Unstructured Pruning: Zeros out individual weights, requiring specialized sparse matrix multipliers to realize energy savings.
- Structured Pruning: Removes entire channels, attention heads, or layers, allowing the compressed model to run efficiently on standard hardware accelerators. Sheared LLaMA [12] demonstrates that structured pruning of a large pre-trained model can yield smaller, high-performing variants without full retraining, significantly reducing the carbon cost of producing new model sizes.

E. 4.3 Knowledge Distillation

This technique involves a large, resource-intensive "Teacher" model transferring its generalized representations to a much smaller, energy-efficient "Student" model [3]. The student learns to mimic the teacher's output distribution, achieving near-parity in performance while requiring a fraction of the inference energy. This approach has been central to the development of efficient open-source models such as the Gemma family.

V Hardware-Software Co-Design and Infrastructure

Green AI is not solely an algorithmic pursuit; the underlying silicon and energy grids play a deterministic role in total emissions [13, 14].

F. 5.1 Carbon-Aware Scheduling

Cloud providers are increasingly implementing spatial and temporal workload shifting. By pausing training runs or shifting them to data centres located in regions currently powered by high yields of renewable energy (e.g., solar peaks or high wind periods), operational carbon can be reduced by up to 30% without any change to the model architecture [13].

G. 5.2 Neuromorphic and Edge Computing

To bypass the von Neumann bottleneck the energy cost of constantly moving data between memory and processors the industry is exploring neuromorphic chips such as Intel Loihi [15]. These brain-inspired architectures utilize Spiking Neural Networks (SNNs), which only consume power when a "spike" (activation) occurs, offering up to a

100× reduction in energy consumption for specific sensor and edge-AI workloads.

VI. Open Challenges and Future Directions

Despite significant advancements, several systemic challenges impede the universal adoption of Green AI:

The Jevons Paradox in AI

As inference becomes more energy-efficient and cheaper, the demand and utilization of LLMs increase exponentially. This rebound effect often results in higher total energy consumption globally, canceling out efficiency gains [1, 14].

Lack of Standardized Telemetry

Current carbon reporting relies heavily on retrospective estimation tools (e.g., CodeCarbon, MLCO2) [13]. The field urgently requires standardized, hardware-level, real-time energy telemetry that is mandated as part of model release cards and reproducibility checklists.

The Cost of "Thinking"

Next-generation models utilizing reinforcement learning for advanced reasoning such as DeepSeek-R1 [17] generate extended hidden chain-of-thought token sequences. While this improves accuracy for complex mathematical and coding tasks, it dramatically spikes the energy cost per query, creating a fundamental trade-off between reasoning depth and carbon efficiency.

VII. Conclusion

The trajectory of frontier AI models is environmentally unsustainable under the current compute-intensive paradigm. While interventions like INT4 quantization [10, 11], LoRA [8], and carbon-aware scheduling [13] offer immediate relief, achieving true sustainability requires a paradigm shift. The AI community must move beyond evaluating models solely on capability benchmarks (e.g., MMLU, HumanEval) and universally adopt accuracy-energy Pareto optimality as the primary metric for state-of-the-art status. Building smarter AI must unequivocally mean building sustainable AI

Acknowledgements

The “Acknowledgement section” is the general term for the list of contributions, credits, and other information included at the end of the text of a manuscript but before the references. Conflicts of interest and financial disclosures must be listed in this section. Authors should obtain written permission to include the names of individuals in the Acknowledgement section.

References

[1] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

[2] 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), 3645–3650.

[3] Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2022). Estimating the carbon footprint of BLOOM, a 176B parameter language model. Journal of Machine Learning Research, 24(253), 1–15.

[4] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

[5] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

[6] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with

IO-awareness. Advances in Neural Information Processing Systems (NeurIPS), 35, 32087–32102.

[7] Peng, B., Alcaide, E., Anthony, Q., et al. (2023). RWKV: Reinventing RNNs for the Transformer Era. Proceedings of EMNLP 2023.

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR).

[9] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems (NeurIPS).

[10] Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. International Conference on Learning Representations (ICLR).

[11] Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., & Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. Proceedings of Machine Learning and Systems (MLSys).

[12] Xia, M., Gao, Z., Zeng, Z., & Chen, D. (2023). Sheared LLaMA: Accelerating language model pre-training via structured pruning. International Conference on Learning Representations (ICLR).

[13] Dodge, J., Prewitt, T., des Combes, R. T., Odmark, E., Schwartz, R., ... & Strubell, E. (2022). Measuring the carbon intensity of AI in cloud instances. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 1877–1894.

[14] Wu, C. J., Raghavendra, R., Gupta, U., ... & Hazelwood, K. (2022). Sustainable AI: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems (MLSys), 4, 795–813.

[15] Davies, M., Srinivasa, N., Lin, T. H., et al. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 38(1), 82–99.

[16] Meta AI. (2024). The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783.

[17] DeepSeek AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948.

[18] Raut Rahul Ganpat and Dr. Sonawane Vijay Ramnath **"**Constructing a User-Cantered Fake News Detection Model by Using Enhance Stacking Ensemble Classification Algorithms in Machine Learning Techniques"JIER, VOL.7,PP 9-13, ISSN (Online) : 2581–6357.

[19] C. Ramakrishna and Dr. Pramod Pandurang Jadhav, **"**Machine Learning Based Sentiment Analysis Implementation for Multiple Reviews Classification**"** JIER, VOL.7,PP 21-25, ISSN (Online) : 2581–6357

[20] Nazeer Shaik, Dr. C. Krishna Priya "Machine Learning Algorithms for Intrusion Detection Systems" JIER, VOL.7,PP 8-18, ISSN (Online) : 2581–6357.

[21] Sarkar Prasanjeet Jyotirmay and Dr. Santosh Pawar, "Advancing Early Detection and Prediction of Diabetes Mellitus: A Robust Soft Voting Ensemble Machine Learning Approach" JIER, VOL.7,PP 12-17, ISSN (Online) : 2581–6357.

[22] Praveen Sharma and Dr. Deepika Pathak,"A Novel Machine Learning Categorial Algorithm with Remote Detecting and GIS Based Decision Support System to Identify Disaster", JIER, VOL.5, PP 25-29, ISSN (Online) : 2581–6357.