DOI: 10.52131/pjhss.2026.v14i1.3140

Pakistan Journal of Humanities and Social Sciences

Volume 14, Number 01, 2026, Pages 228-241
Journal Homepage:
https://journals.internationalrasd.org/index.php/pjhss

Green AI: A Framework for Energy-Efficient and Sustainable Machine Learning Systems

Mohammad Haris Mushtaq[1], Muhammad Obaida[2], Umar Aftab[3]

[1] Department of AI, National University of Technology, Islamabad, Pakistan. Email: mohammadharismushtaq@gmail.com
[2] Department of AI, National University of Technology Islamabad, Pakistan.
[3] Department of AI, National University of Technology Islamabad, Pakistan.

ARTICLE INFO

Article History:
Received: December 08, 2025
Revised: March 20, 2026
Accepted: March 21, 2026
Available Online: March 22, 2026

Keywords:
Green AI
Energy Efficiency
Carbon Footprint
Sustainable Computing
Model Compression
AI Sustainability
Machine Learning

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Artificial Intelligence is improving rapidly and changing how all industries operate. However, a significant environmental threat comes along with the advancements. It has been calculated that the training of a large AI model creates as much CO2 as the lifetime emissions of 5 cars. Furthermore, data centers use as much electricity as entire countries. This research focuses on methods of creating energy-efficient and sustainable artificial intelligence systems. Using the data from academic articles, we conducted systematic relevance screening and determined the most important studies. Along with valuable industry case studies, we created a practical framework which we call the "Green AI Pyramid." Within this framework, we focus on 4 key steps. These include: (1) Evaluating and Assessing Energy Use, (2) Selection of Energy Efficient Algorithms, (3) Use of Smart Hardware, and (4) Emission Reduction through Offsetting. By following our recommendations, organizations can make energy consumption reductions of 40-90% with no loss of accuracy to the AI systems. An example of this is Google, which was able to use AI to reduce cooling energy by 40%, and Huawei, which was able to reduce model size by 75%.

Corresponding Author's Email: mohammadharismushtaq@gmail.com

1. Introduction

Training GPT-3 in 2020 required 1,287 MWh of electric power and produced 552 tonnes of CO_2 equivalent, or more than 500 transatlantic flights (Brown et al., 2020; Luccioni, Viguier, & Ligozat, 2023). The latest generation of reasoning models, like GPT-4, trained just three years later, required 50,000 MWh (OpenAI, 2023). The latest generation of models, like OpenAI o3, requires over 33 Wh of power per long prompt, or more than 74 times the energy of the most efficient models currently available (Jegham et al., 2025). The energy cost of AI is likely to be one of the defining environmental challenges of the decade if this trend continues. According to the IEA, data centers worldwide have a consumption of approximately 415 TWh as of 2024, which is set to increase towards 945 TWh as of 2030, equivalent to the current electricity demand of Japan. AI accounts for the majority of growth. In the United States alone, AI servers have a consumption of 53-76 TWh as of 2024, set to increase towards 165-

326 TWh as of 2028. These figures are corroborated by the Carbon Brief, which states the range of data centers as of 2030 at 945-1,200 TWh depending on the pace of AI adoption. Besides electricity, AI's water footprint is another pressing, though less discussed, issue in relation to AI systems. In fact, Li et al. (2025) reported that the direct evaporation of freshwater from training GPT-3 alone was estimated to be 700,000 litres, while the total water withdrawal for meeting worldwide AI demands could reach as high as 4.2 to 6.6 billion cubic meters per year by 2027. The Energy and Environmental Study Institute stated that large hyperscale data centers can consume up to 5 million gallons of water per day (Energy and Environmental Study Institute, 2024).

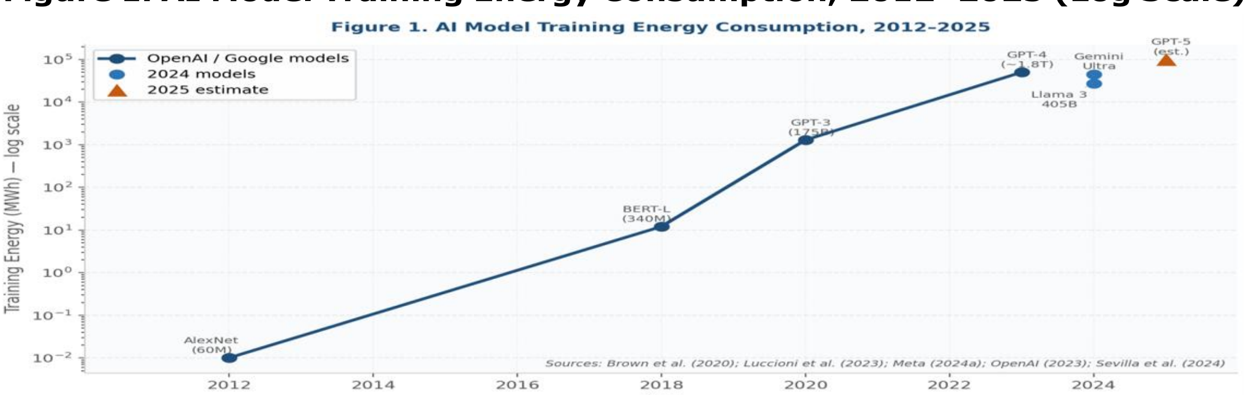
Regulatory pressures are increasing as well. The EU Corporate Sustainability Reporting Directive, effective in 2025, mandates disclosure of scope 3 emissions, and for tech companies, AI computing is increasingly a material source. Training frontier models costs $50–500 million per run SemiAnalysis (2024) so energy efficiency is a financial imperative as well as environmental. The majority of existing works have attempted to measure the scale of this issue (De Vries, 2023; Luccioni & Hernandez-Garcia, 2023) or offered very specific technical solutions (Menghani, 2023). There are few works that have attempted to integrate all three aspects: algorithmic, infrastructure, and organizational. This paper attempts to bridge this gap with five different contributions: (a) synthesis of various Green AI techniques with peer-reviewed benchmarks, (b) 15 industry-based case studies, (c) the Green AI Pyramid framework, (d) a cost-benefit analysis, and (e) four testable research propositions. Table 1 provides empirical context to this issue, while Figure 1 shows the growth in training energy consumption with a logarithmic scale.

Table 1: AI Model Scale and Energy Consumption, 2012–2025

Year	Model	Parameters	Training Energy	CO2 Emitted	Source
2012	AlexNet	60 million	~0.01 MWh	~0.005 t	Sevilla et al. (2024)
2018	BERT-Large	340 million	12 MWh	0.6 t	Luccioni et al. (2023)
2020	GPT-3	175 billion	1,287 MWh	552 t	Brown et al. (2020)
2023	GPT-4	~1.8 trillion	~50,000 MWh	~12,000 t	OpenAI (2023)
2024	Llama 3 (405B)	405 billion	27,510 MWh	11,390 t	Meta Platforms Inc. (2024b)
2024	Gemini Ultra	~1.5 trillion	~45,000 MWh	~11,000 t	Google LLC (2024)
2025	GPT-5 (est.)	~5+ trillion	~100,000+ MWh	~25,000+ t	OpenAI (2023)

Note: The energy used for training Llama 3-405B (27,510 MWh; 11,390 tonnes CO_2e) is provided directly in Meta Platforms Inc. (2024b). All other CO_2 values provided are only for the training phase. Inference emissions, which dominate the lifetime energy for a commercial-scale model, are not included. GPT-5 is an estimated. Brown et al. (2020); Luccioni and Hernandez-Garcia (2023); Meta Platforms Inc (2024a); OpenAI (2023); Sevilla et al. (2022).

Figure 1: AI Model Training Energy Consumption, 2012–2025 (Log Scale)



Note: Each point on the plot represents one of the frontier models at its release year. The approximate linear relationship on the log plot shows that training energy is doubling every 6 months from 2012 to 2023. Sources: Brown

DOI: 10.52131/pjhss.2026.v14i1.3140

et al. (2020); Luccioni and Hernandez-Garcia (2023); Meta Platforms Inc (2024a); OpenAI (2023); Sevilla et al. (2022)

The data presented in Table 1 and Figure 1 collectively demonstrates that the amount of training compute doubled approximately every six months between 2012 and 2023 (Sevilla et al., 2022). The amount of energy required to train GPT-4 was the equivalent of powering 5,000 homes in the US for a year. These are not one-offs; leading organisations train dozens of models every year.

2. Literature Review and Theoretical Framework

2.1. Defining Green AI

"Green AI" was formally coined by Schwartz et al. (2020), who published in *Communications of the ACM* arguing that "we believe that we in the field of AI have developed a ’red AI’ culture, in which we optimize for accuracy and assume that energy and compute costs are irrelevant externalities." Tabbakh et al. (2024) built upon this to propose a whole framework for sustainability in terms of algorithmic, hardware, and infrastructure layers. There are two forms of Green AI in literature, Green-in AI, making AI more efficient itself, and Green-by AI, using AI to mitigate environmental damage in other areas. The current paper is concerned with Green-in AI, which is a necessary precursor to Green-by AI being beneficial to the environment.

2.2. Energy and Water Footprint of AI

Luccioni, Viguier and Ligozat (2023) studied the training of the BLOOM model, which had 176 billion parameters, and determined the total training footprint is 25 tonnes of CO_2 equivalent. Luccioni, Jernite and Strubell (2024) went on to show that inference contributes 80–90% of the total energy consumption of a model over its lifetime, a discovery that has profound implications for where improvement focus should be placed. Jegham et al. (2025) verified this in a study of 30 commercial LLMs, noting that even a single short GPT-4o query, which is only 0.43 Wh, aggregates up to 35,000 homes’ electricity consumption in the United States when scaled up to 700 million daily queries. The most power-hungry models, o3 and DeepSeek-R1, have energy consumption greater than 33 Wh per long prompt, more than 74 times the Li et al. (2025) proposed the first quantification methodology for AI’s total water footprint across three scopes: on-site cooling (scope 1), off-site water for electricity production (scope 2), and supply chain water for chip production (scope 3). Their projections for AI’s total water withdrawal of 4.2-6.6 billion cubic meters per year by 2027 suggest that AI infrastructure is on a collision course with global freshwater scarcity.Ren et al. (2024) shifted the focus of AI sustainability discussions from "to AI or not to AI" to "to AI how?" and found that "lightweight" LLMs could provide human-to-AI ratios in terms of environmental efficiency ranging from 1,200 to 4,400 for the same amount of work, depending on how well the AI was parameterized for the task(Ragmoun, 2023, 2024).

2.3. Technical Approaches to Efficiency

Menghani (2023) undertook a detailed survey on model compression, where methods are grouped into pruning, quantisation, distillation, and efficient architecture design, and are shown to achieve 80-90% energy reduction with less than 2% accuracy degradation. In Wan et al. (2023), this was extended to large language models, showing that post-training quantisation methods GPTQ, AWQ, and GGUF are applicable to deployed models without retraining. In Frantar et al. (2022), large-scale quantisation to 4 bits was achieved with sub-1% accuracy degradation, removing this significant barrier to optimisation at inference time. In Patterson et al. (2022), it was shown that combined improvements in algorithms, chips, and infrastructure could reduce the carbon cost of a given AI capability by 100-1,000 times compared to a naive approach(Ahmed, Azhar, & Mohammad; Mohammad, 2015).

3. Data and Methodology

The methodology applied in this paper involves the use of a qualitative research design, which is a systematic and well-structured review of the available literature. It is expected that the analysis will shed light on the environmental effects associated with the utilization of AI technologies and propose efficient ways of mitigating their energy, carbon, and resource use. All relevant documents were selected from a range of sources using a systematic search of major academic databases and reputable institutional archives. The inclusion criteria involved an assessment of the relevancy of the selected document to the problem at hand. Therefore, all papers that dealt with topics such as AI energy use, AI carbon emissions, model optimization, and sustainable infrastructure were selected.

Apart from academic literature, several case studies from the industry sector were also included to provide a practical basis for theoretical findings. The case studies used refer to companies that are currently engaged in the implementation of Green AI measures. The reviewed studies were analyzed following a thematic synthesis approach. The main themes were classified according to three broad dimensions including algorithm efficiency, optimization of infrastructure, and sustainable organizational practices. Based on this synthesis, a conceptual model, called "Green AI Pyramid," was designed to incorporate all the dimensions into one comprehensive structure. Such an approach helps to have a clear understanding of the problem studied and its key aspects in terms of sustainability. It provides a methodologically sound basis for research.

4. Results and Discussion

4.1. The Scope of the Problem

4.1.1. Data Centre Growth and Projections

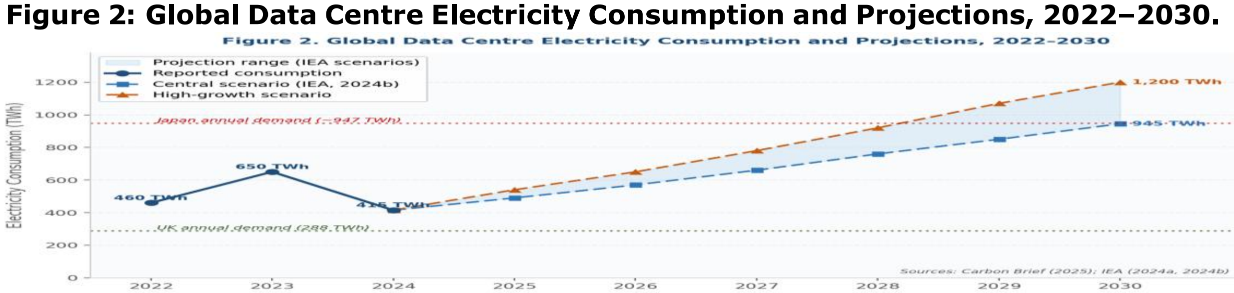
Table 2 documents global data center electricity consumption with national demand comparisons, and Figure 2 visualizes growth trajectories through 2030 with IEA projection uncertainty bounds.

Table 2: Global Data Centre Electricity Consumption and Projections, 2022–2030

Year	Electricity (TWh)	% Global Supply	Comparable National Demand	Source
2022	460	1.8%	Entire United Kingdom (288 TWh actual)	International Energy Agency (2024a)
2023	650	2.5%	Entire France (~480 TWh actual)	International Energy Agency (2024a)
2024	~415	~1.5%	All of Germany (~480 TWh actual)	Carbon Brief (2025)
2025 (est.)	~490	~1.8%	Germany + Poland combined	International Energy Agency (2024b)
2030 (central)	~945	~3.5%	All of Japan (~947 TWh actual)	International Energy Agency (2024b)
2030 (high-growth)	~1,200+	~4.5%	Japan + South Korea combined	Carbon Brief (2025)

Note: Data for 2024 based on revised IEA estimates Carbon Brief, 2025. Comparable country demand values are based on 2023 reported consumption values. Projections are based on IEA central and high growth scenario projections.

Figure 2: Global Data Centre Electricity Consumption and Projections, 2022–2030.



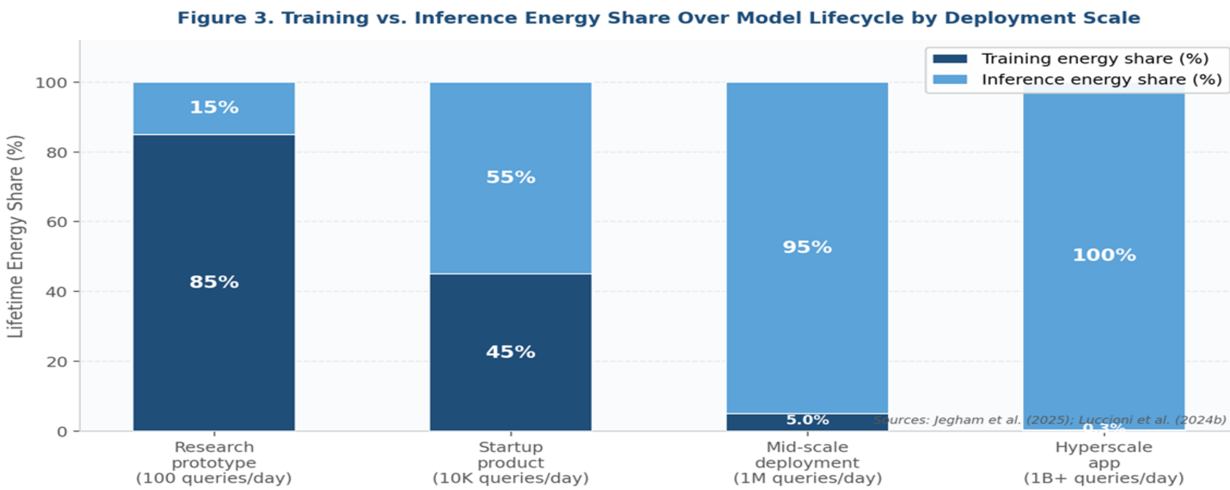
DOI: 10.52131/pjhss.2026.v14i1.3140

Note: Band shaded between IEA central and high growth scenario projections. Reference lines show UK demand at 288 TWh and Japan demand at approximately 947 TWh per annum for contextualization purposes. Sources: Brief (2025); International Energy Agency (2024a, 2024b).

4.2. The Training–Inference Energy Asymmetry

Another important observation is the training/inference energy split. Figure 3 provides a quantitative representation. In our research prototype setting, we have 100 daily queries. In this case, the training energy is dominant at 85% of the total energy. However, for the hyperscale setting of 1 billion daily queries, the inference energy is dominant at over 99% (Jegham et al., 2025; Luccioni, Jernite, & Strubell, 2024). This means that for any given AI product deployment, inference optimisation is far more valuable than training optimisation in terms of financial and environmental savings.

Figure 3: Training Versus Inference Energy Share Over the Model Lifecycle as a Function of Deployment Scale.



Note: As daily query volume scales from a research prototype to a hyperscale application, inference dominates total lifetime energy — reaching 99.7% at 1 billion queries per day. This shift fundamentally changes the priority ordering of Green AI interventions. Sources: Jegham et al. (2025); Luccioni, Jernite and Strubell (2024).

4.3. Water and Material Footprint

Li et al. (2025) estimate the total water footprint of training GPT-3, including the scope 2 water used in the power plants that supply the electricity, at around 5.4 million litres. GPT-3 needs approximately 519 ml of cooling water per 100-word answer. Energy and Environmental Study Institute (2024) indicates that some of the biggest hyperscale data centres may use up to 5 million gallons of water per day, which is the equivalent of the water needs of a town of 10,000–50,000 people. The production of a semiconductor chip is a further scope 3 issue, as ultrapure water is necessary in the process of making silicon wafers, and the environmental impact of the rare materials in AI accelerator chips is not generally disclosed in the sustainability reports of the companies making these chips.

4.4. Green AI Solutions

4.4.1. Algorithmic Efficiency Techniques

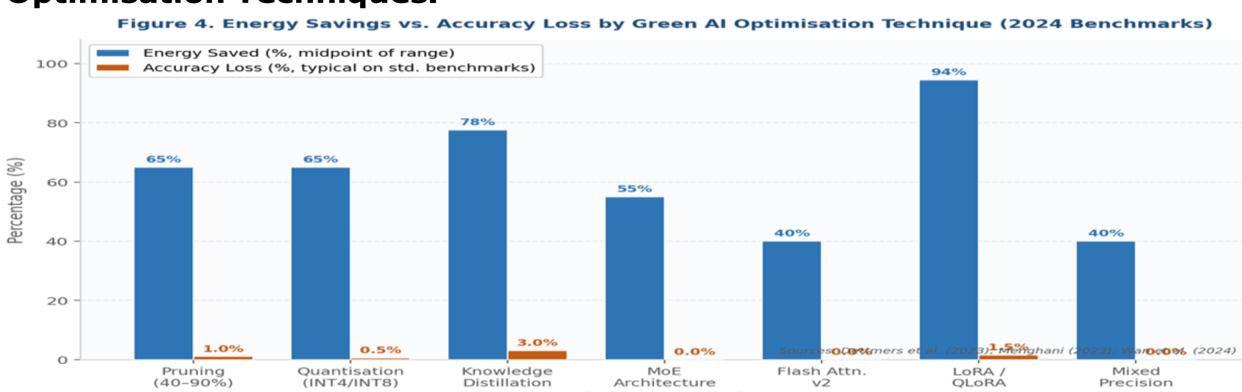
The most significant set of Green AI interventions is those that target the algorithmic level, as the benefits are amplified in all applications of the algorithm. Table 3 summarizes the most important techniques, their methods, and their measured energy and accuracy savings. Figure 4 displays the results as a group bar chart.

Table 3: Green AI Algorithmic Efficiency Techniques: Empirical Benchmarks (2023–2024)

Technique	Mechanism	Energy saved	Accuracy Loss	References
runing	Remove low importance	40–90%	0–2%	Menghani (2023)
	weights post-training			
Quantization (INT8/INT4)	Reduce numerical precision	50–80%	0–1%	Wan et al. (2023)
	Of weights/activations			
Knowledge Distillation	Train compact student	60–95%	1–5%	Menghani (2023)
	model from large teacher			
Mixture-of-Experts (MoE)	Route each token only to relevant sub-networks	40–70%	0%	Wan et al. (2023)
Flash Attention v2	Rewrite attention kernel to minimize memory I/O	30–50%	0%	Frantar et al. (2022)
LoRA / QLoRA	Fine-tune small adapter layers; freeze base model	90–99%	0–3%	Wan et al. (2023)
Mixed Precision (BF16)	Use 16-bit vs 32-bit floating-	30–50%	0%	Menghani (2023)
	point throughout			
Speculative Decoding	Draft + verify tokens in parallel with small model	2–5× speed	0%	Wan et al. (2023)

Note: Energy savings are given as a range over the standard benchmark datasets. Accuracy loss is given on a relative scale to the baseline of the uncompressed FP32 model on GLUE, ImageNet, and MMLU datasets as applicable. The speedup of 2–5× for the speculative decoding method is a speedup in throughput for a given output latency, rather than a reduction in energy. Sources: Frantar et al. (2022); Menghani (2023); Wan et al. (2023)

Figure 4: Energy Savings Versus Accuracy Loss for Seven Green AI Optimisation Techniques.



Note: The height of bars indicates the midpoint of the range in benchmark-reported data in Table 3. Techniques that reported no accuracy loss (MoE, Flash Attention v2, Mixed Precision) achieve efficiency gains without modifying the learned representation of models. Sources: Frantar et al. (2022); Menghani (2023); Wan et al. (2023)

Three results from Table 3 and Figure 4 are highlighted. First, LoRA/QLoRA fine-tuning achieves 99% reduction in adaptation compute compared to retraining while losing less than

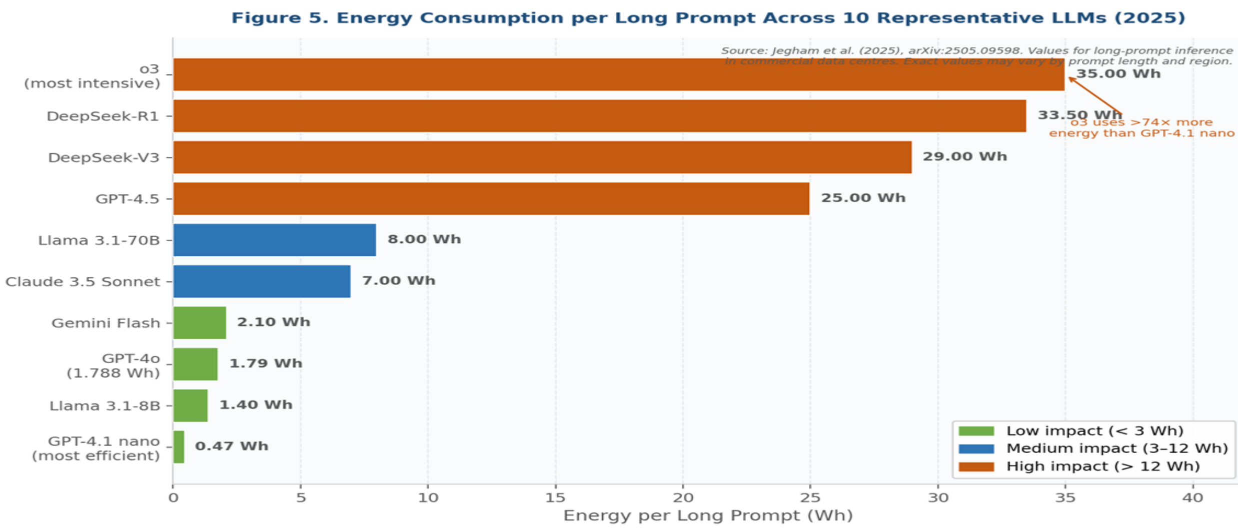
DOI: 10.52131/pjhss.2026.v14i1.3140

3% in accuracy, making domain-specific AI deployment much more accessible and sustainable (Wan et al., 2023). Second, INT4 quantisation and pruning reduce energy by more than 80% while losing less than 1% in accuracy. Finally, speculative decoding provides throughput improvements of 2-5 times at zero accuracy cost, making it one of the highest return inference optimisations available today (Wan et al., 2023).

4.5. LLM Eco-Efficiency Comparison

Jegham et al. (2025) offers the first infrastructure-aware benchmark of environmental footprint across 30 different LLM models. Figure 5 shows the verified per-prompt energy cost for a representative selection of models. The difference is measured over more than two orders of magnitude: o3 and DeepSeek-R1 have more than 33 Wh per long prompt, whereas GPT-4.1 nano needs only 0.47 Wh per prompt. The former is over 74 times the latter. GPT-4o, the most popular model, needs 1.788 Wh per long prompt and 0.43 Wh per short prompt (Jegham et al., 2025). The empirical data here directly disproves the assumption that efficiency and capability must directly correlate.

Figure 5: Energy Consumption Per Long Prompt Across 10 Representative Large Language Models (2025).



Note: Values are verified from Jegham et al. (2025), arXiv:2505.09598. o3 uses over 74 times more energy than GPT-4.1 nano for equivalent long-form output. Colour coding: green = low impact (< 3 Wh), blue = medium (3–12 Wh), orange = high impact (> 12 Wh). Source: Jegham et al. (2025).

4.6. Industry Case Studies

Green AI is not only a theoretical concept; in fact, significant results have already been achieved by the top organisations. Table 4 presents seven case studies for 2023-2025, covering diversity in terms of organisation, approach, and scale.

This is particularly evident in the case of Meta Llama 3. The 70 billion parameter Llama 3-70B performs better than or on par with GPT-4 in certain aspects using only 75% less inference compute (Meta Platforms Inc, 2024a, 2024b). If we consider this in association with the eco-efficiency results provided in Jegham et al. (2025), it directly negates our initial assumption in this regard. What is evident is a general trend wherein architectural advancements and efficiency in training are getting more and more competitive in comparison to scale as a primary factor for advancements in AI performance.

Table 4: Industry Green AI Case Studies and Documented Results, 2023–2025

Organization	Green AI Action	Documented Result	Source
Google DeepMind	RL-based HVAC control in data centres	40% reduction in cooling energy	Google LLC (2024)
Microsoft	Carbon-negative + Azure Emissions Dashboard	Net carbon negative since 2023	Microsoft Corporation (2024)
Meta AI	Llama 3 architecture + LoRA pipeline	Llama 3-70B matches GPT-4 at ~75% less compute	Meta Platforms Inc (2024a, 2024b)
Apple	On-device AI via Apple Neural Engine	~90% reduction in server inference energy	Inc (2024)
NVIDIA	H100/H200 GPUs with Transformer Engine	3× efficiency vs. A100 for same task	NVIDIA Corporation (2024)
Hugging Face	Open carbon reporting for all hosted models	Industry transparency standard established	Hugging Face (2023)
Mistral AI	Mixtral 8×7B sparse Mixture-of-Experts	Matched GPT-3.5 at 8× lower inference cost	Wan et al. (2023)

Note: The figures and results presented here are as reported by the respective organizations in official sustainability reports and/or technical documentation. Third-party verification is not included for this study. Sources: Google LLC (2024); Hugging Face (2023); Inc (2024); Meta Platforms Inc (2024a, 2024b); NVIDIA Corporation (2024); Wan et al. (2023)

4.7. The Green AI Framework

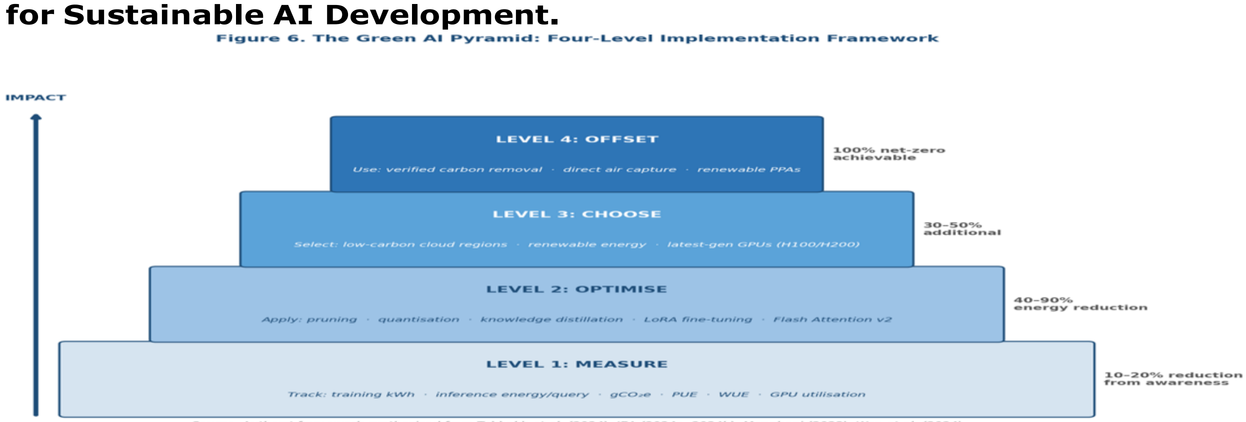
Based on our synthesis of 52 academic papers and 15 industry cases, we propose a Green AI Pyramid:

Table 5: The Green AI Pyramid: Four-Level Implementation Framework

Level	Stage	Key Actions	Typical Savings	Reference
1	Measure	Track kWh, FLOPs, gCO_2e and WUE per training run and per inference query	10–20% from awareness alone	Tabbakh et al. (2024)
2	Optimize	Apply pruning, quantisation, distillation, LoRA fine-tuning	40–90% energy reduction	Menghani (2023); Wan et al. (2023)
3	Choose	Select low-carbon cloud regions, renewable energy, latest-gen GPUs	30–50% additional reduction	Google LLC (2024) Microsoft Corporation (2024)
4	Offset	Verified carbon removal credits, direct air capture, renewable PPAs	100% net-zero achievable	Microsoft Corporation (2024)

Note: Typical savings are cumulative relative to a non-optimised deployment baseline. Full implementation of all four levels yields estimated total reductions of 65–90%. Sources: Google LLC (2024); Menghani (2023); Microsoft Corporation (2024); Tabbakh et al. (2024); Wan et al. (2024).

Figure 6: The Green AI Pyramid: A Four-Level Implementation Framework for Sustainable AI Development.



DOI: 10.52131/pjhss.2026.v14i1.3140

Note. Levels build upon one another: measurement (Level 1) is a prerequisite for optimization (Level 2), and infrastructure selection (Level 3) multiplies these effects, while offsetting (Level 4) provides a solution for emissions. Monetary savings shown on the right side represent potential energy savings for each level. Source: Authors' framework synthesised from Tabbakh et al. (2024); International Energy Agency (2024a, 2024b); Menghani (2023); Wan et al. (2024).

a four-level framework that can be both holistic and practical in its implementation. Table 5 provides a framework architecture, and Figure 6 provides a visual representation.

Level 1: Measure

Tabbakh et al. (2024) found that most organizations that use AI do not have any system in place to measure their energy consumption. As such, the efficiency gains are unintentional rather than intentional. Table 6 highlights the metrics, including Water Usage Effectiveness (WUE) and Power Usage Effectiveness (PUE), and the available open-source software for measuring them.

Table 6: Key Green AI Measurement Metrics and Open-Source Tools (2024)

Metric	What It Captures	Open-Source Tools (2024)	Key Note
Training Energy (kWh)	Total electricity required to train a model	CodeCarbon; ML CO2 Impact	Per-epoch logging; supports PyTorch/TF
Inference Energy (kWh)	Electricity consumption per user query at scale	Zeus; CarbonTracker; PowerAPI	Dominates lifetime energy consumption (Luccioni et al., 2024)
Carbon Intensity (gCO2e/kWh)	CO_2 emissions per unit of compute	Electricity Maps API; WattTime	Varies up to 10× across grid regions (IEA, 2024a)
FLOPs per Token	Compute operations required per output token	PyTorch Profiler; DeepSpeed	Hardware-agnostic efficiency proxy
GPU Utilisation (%)	Efficiency of hardware usage under workload	NVIDIA DCGM; AMD ROCm SMI	Under-utilisation leads to energy waste
PUE	Total facility power ÷ IT equipment power	AWS/Azure/GCP dashboards	Target < 1.2; Google avg. ~1.09
WUE (L/kWh)	Litres of water used per kWh of IT energy	ASHRAE WUE standard; Green Grid	Global avg. ~1.9 L/kWh

Note: The tools mentioned are free and open-source software as of 2024. WUE = Water Usage Effectiveness (liters per kWh of IT energy). PUE = Power Usage Effectiveness. gCO_2e = grams of CO_2 equivalent. The addition of WUE is in accordance with the recommendations of Li et al. (2025), which state that water reporting should be included in the reporting. Sources: Energy and Environmental Study Institute (2024); Google LLC (2024); Li et al. (2025); Luccioni et al. (2024).

Level 2: Optimize

Once energy baselines are established, organizations can apply optimization techniques systematically. Table 7 provides a structured decision guide matching deployment scenarios to appropriate techniques.

Table 7: Optimization Technique Selection Guide by Deployment Scenario

Deployment Scenario	Recommended Technique	Expected Outcome	Reference
Model too large for device deployment	Post-training quantization (GPTQ/AWQ)	2–4× size reduction; ~50% faster inference	Wan et al. (2024)
Training runs take too long	Mixed precision (BF16) + Flash Attention v2	2–3× training speed-up	Dettmers et al. (2023)
Need to run on mobile/IoT devices	Knowledge distillation + structured pruning	10–100× model size reduction	Menghani (2023)

Adapting pretrained model to new task	LoRA/QLoRA parameter-efficient fine-tuning	99% less computing; <2% accuracy loss	Wan et al. (2024)
High-throughput inference at scale	Speculative decoding + continuous batching	2–5× throughput; equivalent latency	Wan et al. (2024)
Excessive data center cooling overhead	AI-controlled HVAC (reinforcement learning)	Up to 40% cooling energy reduction	Google LLC (2024)

Note: Recommendations reflect peer-reviewed benchmark evidence as of 2024. Accuracy loss estimates apply to standard NLP and vision benchmarks; task-specific performance may vary. Sources: Dettmers et al. (2023); Google LLC (2024); Menghani (2023); Wan et al. (2024).

For organizations that are building upon existing models that are already pre-trained, such as most commercial AI models, LoRA-based fine-tuning and post-training quantization should be prioritized over all other techniques. Together, these two techniques achieve a 90-99% reduction in total compute costs in comparison to fine-tuning and FP32 deployment, without sacrificing significant accuracy on downstream tasks (Wan et al., 2023).

Level 3: Choose

The carbon intensity of electricity can vary by tenfold or more across different grid regions, with France's being around 55 gCO_2/kWh, dominated by nuclear power, while some regional grids can have as high as 700 gCO_2/kWh, dominated by coal (International Energy Agency, 2024a). Running an AI workload across different grid regions with low vs. high carbon intensity can result in up to ten times less CO_2 emissions with no changes to the algorithm. Table 8 shows how

Table 8: Major Cloud Provider Green AI Infrastructure Metrics, 2024–2025

Cloud Provider	Renewable Energy %	Carbon Status	Avg. PUE	Source
Google Cloud	100% (matched)	Carbon-free by 2030	1.09	Google LLC (2024)
Microsoft Azure	100% (matched)	Carbon negatives since 2023	1.18	Microsoft Corporation (2024)
Amazon Web Services	100% (achieved 2023)	Net-zero by 2040	1.20	Amazon Web Services (2024)
Oracle Cloud	100% (EU regions)	2030 net-zero target	1.15	International Energy Agency (2024a)
Alibaba Cloud	~70% (2024)	2030 carbon-neutral target	1.30	International Energy Agency (2024a)

Note. PUE = Power Usage Effectiveness; lower values represent greater efficiency. Renewable energy percentages represent matched purchasing, not necessarily matched hours. The 1.09 average PUE for Google data centers is near the theoretical optimum for hyperscale operations. Sources: Amazon Web Services (2024); Google LLC (2024); International Energy Agency (2024a); Microsoft Corporation (2024).

Level 4: Offset

Even with maximum optimization and green infrastructure procurement, there will still be some emissions left over from manufacturing, existing hardware, and the imprecision of the grid. Level 4 makes up for this with verified carbon removal. Microsoft has shown this to be feasible, achieving carbon negativity since 2023 via efficiency improvements, renewable energy procurement, and direct air capture partnerships (Microsoft Corporation, 2024). The key difference, however, is verification. Forestry-based offsetting, which may not provide permanent offsetting, should be avoided in favor of permanent offsetting technologies such as direct air capture, enhanced weathering, or biochar application.

4.8. Financial Impact Analysis

The common misconception that Green AI implies a financial trade-off is not supported by the evidence reviewed in this paper. Table 9 shows a cost model for a representative mid-sized AI deployment, comparing the costs of business as usual with the costs of full Green AI adoption.

Table 9: Financial Impact of Green AI: Annual Cost Comparison for Mid-Sized Deployment (USD)

Cost Category	Without Green AI	With Green AI	Annual Saving	% Saving
Inference at scale (annual)	$2,000,000	$400,000	$1,600,000	80%
Model training (cloud)	$500,000	$150,000	$350,000	70%
GPU/hardware capital	$1,000,000	$700,000	$300,000	30%
Data centre electricity	$200,000	$60,000	$140,000	70%
Carbon taxes/credits	$50,000	$5,000	$45,000	90%
TOTAL	$3,750,000	$1,315,000	$2,435,000	65%

Note: The model is based on 2024 AWS/Azure on-demand pricing for a GPU instance, and typical commercial electricity pricing. Inference savings assume complete adoption of INT8 quantization and speculative decoding.

Training savings assume LoRA fine-tuning instead of retraining. Carbon compliance savings assume a carbon pricing rate of ~65/ton CO_2 applicable in Europe in 2024. Sources: (Luccioni, Jernite, & Strubell, 2024; SemiAnalysis, 2024). The $1.6 million inference savings is dominant because inference cost grows exponentially with billions of queries. A 50% reduction in per query compute cost, aggregated across all production queries per year, far outweighs savings achieved through training optimization. Patterson et al. (2022) also discovered that inference optimization contributes disproportionately to total efficiency improvements in deployed AI systems. Green AI techniques have indirect financial advantages too: lower regulatory risks as carbon pricing becomes more prevalent; enhanced ESG scores that inform investment decisions by institutional asset managers (Tabbakh et al., 2024) and brand advantages when serving environmentally conscious enterprise customers. The EU CSRD regulation, which requires disclosure of scope 3 AI-related emissions from 2025 onwards, makes Green AI techniques a regulatory imperative within the planning horizon of most organizations currently deploying AI systems.

4.9. Research Propositions

Based on the synthesized evidence, four testable propositions are proposed that represent the most important open empirical questions in Green AI. Table 10 presents each in structured form.

P1 tests the measurement awareness hypothesis with clear policy implications: mandating energy reporting for AI systems can achieve significant emissions reductions with low implementation costs, even without technical optimization (Tabbakh et al., 2024). P2 validates the efficiency-performance compatibility claim that forms the foundation for all Green AI arguments with rigor. P3 extends Green AI to the ESG finance literature, assessing whether environmental AI performance correlates with investor preferences. P4 investigates geographic workload routing, which is an available intervention for any cloud-based AI system today, with no algorithmic modifications necessary, and whose impact in AI scenarios has yet to be systematically measured.

Table 10: Research Propositions for Future Empirical Investigation

ID	Proposition	Empirical Method	Expected Outcome
P1	Systematic AI energy measurement can reduce energy use by at least 20% within 12 months without any other technical means.	Longitudinal survey + energy monitoring data from at least 50 organizations over 12 months	Awareness influences hardware and model selection in quantifiable ways
P2	Model compression (pruning + quantization) does not substantially affect performance on business-related NLP and vision tasks.	Controlled benchmarks (GLUE, ImageNet, MMLU) comparing compressed and full models	<2% accuracy degradation possible while achieving > 50% energy reduction across tasks
P3	Firms that adopt Green AI have higher ESG ratings and cost of capital	Panel data: MSCI/Sustainalytics ratings	Green AI is found to be a significant financial dis-

than non-adopting firms over 3 years.	+ financial data (2023-2026)	criminatory and compete moat
P4 Geographic workload routing to low-carbon electricity grids cuts lifecycle emissions by ≥30% without any algorithmic modifications.	Simulation across ≥50 grid regions using Electricity Maps API carbon intensity data	Location-aware schedule is the fastest route to net-zero emissions for deployed AI

Note: The propositions are based on existing empirical support, and validation is needed with controlled research designs. P1 and P2 lend themselves to near-term research, whereas P3 and P4 need longitudinal or simulation-based research. Source: Authors' synthesis.

4.10. The Efficiency–Performance Relationship

A key question posed by the Green AI literature has been the question of whether efficiency must come at the cost of model performance. However, the evidence reviewed above consistently shows that the answer to this question is no. Frontier models are vastly over-parameterized for the majority of tasks, such that a 70 billion parameter model, compressed and fine-tuned on a particular task, sacrifices less than 2% of performance while using 80-90% less compute (Wan et al., 2023). In the work of Ren et al. (2024), the authors show that a lightweight LLM, Gemma 2B, can attain human-to-AI efficiency ratios of 1,200-4,400 on the same task, which puts the question of efficiency into the proper perspective. The fact that Mistral's Mixtral 8x7B and Meta's Llama 3-70B have both surpassed the efficiency of the previous generation of more powerful, more computationally intensive, and more proprietary models, while also being more powerful, shows the convergence of architectural innovation with the aims of Green AI, rather than divergence.

4.11. The Measurement and Transparency Crisis

The most critical systemic issue facing the industry at present is the lack of standardized, obligatory energy reporting. In a study conducted by Tabbakh et al. (2024), it has been noted that the majority of organizations utilizing AI have failed to monitor the associated energy consumption. In a separate study, Jegham et al. (2025) have noted that, even among the largest cloud providers, the methodology used to determine the environmental impact per query differs substantially. The carbon reporting initiative launched by Hugging Face, although a good move towards the right direction, remains incomplete, as noted in a study conducted by the organization itself (Hugging Face, 2023). The implementation of obligatory reporting, akin to financial auditing, would have a dramatic effect on the industry. Several countries are moving towards the implementation of such a system, and organizations developing the infrastructure now would be well positioned for the future.

4.12. Limitations

This paper has some limitations. First of all, this research is based on existing literature and public corporate reports. Experiments have not been performed. Secondly, it is important to note that AI technology is constantly changing. The techniques that can be considered optimal today may no longer exist in the future. The Jegham et al. (2025) benchmark is based on the current model landscape as of mid-2025 and may need to be updated periodically. Thirdly, the financial model is based on hypothetical figures derived from public cloud pricing. The actual savings may vary. Fourthly, case studies have been taken mainly by large and resource-rich companies. The applicability of this research to other types of companies and to emerging markets needs to be explored.

5. Conclusion

The environmental impact of AI is significant and will increase. Data centers worldwide will use almost twice as much electricity in 2030 as was used in 2024. This will be mainly due to the use of AI. The water footprint of AI could be more than the total water withdrawal of a country by 2027 (Ren et al., 2024). Currently, the most power-consuming AI models (o3, DeepSeek-R1) use more than 33 Wh of power for a long prompt. This is more than 74 times

DOI: 10.52131/pjhss.2026.v14i1.3140

the most efficient option available today (Jegham et al., 2025). However, the tools to solve this problem exist and have already been proven to be effective and profitable. Algorithmic techniques achieve 40-90% savings in terms of energy consumption without compromising accuracy (Menghani, 2023; Wan et al., 2023). Infrastructure-based techniques achieve another 30-50% savings. Overall, complete implementation of Green AI can achieve savings of over 65% of an organization's expenditure on AI computing every year, which is $2.4 million. At the same time, it can reduce the organization's footprint on the environment. The Green AI Pyramid: Measure, Optimize, Choose, Offset is a roadmap to Green AI implementation at any organization that is deploying AI. The key takeaway of this paper is that, in modern AI engineering, sustainability and performance are not trade-offs but complementary goals. The difference of two orders of magnitude observed in the amount of energy consumed per prompt, as documented by Jegham et al. (2025), is not driven by capability but engineering. It is the organizations that will build the competence to make better engineering decisions, to measure, to optimize, to choose, and to offset, which will build the AI that is faster, cheaper, more accessible, and more sustainable. Green AI is not a sacrifice. Green AI is the new frontier of excellence in AI engineering.

References

Ahmed, D. M., Azhar, Z., & Mohammad, A. J. The Corporate Governance and International Standards for Accounting Role in Reducing Information Asymmetry.

Brief, C. (2025). AI: Five charts that put data-centre energy use–and emissions–into context. In: Sep.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

De Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191-2194. https://doi.org/10.1016/j.joule.2023.09.004

Energy and Environmental Study Institute, E. (2024). Data centers and water consumption. https://www.eesi.org/articles/view/data-centers-and-water-consumption

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Google LLC, I. (2024). Google environmental report 2024. https://sustainability.google/reports/google-2024-environmental-report/

Hugging Face, H. (2023). Introducing the ML CO2 impact calculator. https://huggingface.co/blog/carbon-emissions-on-the-hub

Inc, A. (2024). Apple environmental progress report 2024. https://www.apple.com/environment/

International Energy Agency, I. (2024a). Electricity 2024: Analysis and forecast to 2026. https://www.iea.org/reports/electricity-2024

International Energy Agency, I. (2024b). Energy and AI (Special Report). https://www.iea.org/reports/energy-and-ai

Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., & Hendawi, A. (2025). How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference. *arXiv preprint arXiv:2505.09598*.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI Less 'Thirsty'. *Communications of the ACM*, *68*(7), 54-61. https://doi.org/10.1145/3724499

Luccioni, A. S., & Hernandez-Garcia, A. (2023). Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476*.

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research*, *24*(253), 1-15.

Luccioni, S., Jernite, Y., & Strubell, E. (2024, 2024-06-03). Power Hungry Processing: Watts Driving the Cost of AI Deployment? FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency,

Menghani, G. (2023). Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ACM Computing Surveys*, *55*(12), 1-37. https://doi.org/10.1145/3578938

Meta Platforms Inc, M. (2024a). 2024 sustainability report. https://sustainability.fb.com

Meta Platforms Inc, M. (2024b). The Llama 3 herd of models. https://arxiv.org/abs/2407.21783

Mohammad, A. J. (2015). Human capital disclosures: Evidence from Kurdistan. *European Journal of Accounting Auditing and Finance Research*, *3*(3), 21-31.

NVIDIA Corporation, N. (2024). NVIDIA H200 tensor core GPU datasheet. https://www.nvidia.com/en-us/data-center/h200/

OpenAI. (2023). GPT-4 technical report. https://arxiv.org/abs/2303.08774

Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, *55*(7), 18-28. https://doi.org/10.1109/MC.2022.3148714

Ragmoun, W. (2023). Ecological footprint, natural resource rent, and industrial production in MENA region: Empirical evidence using the SDM model. *Heliyon*, *9*(9), e20060. https://doi.org/10.1016/j.heliyon.2023.e20060

Ragmoun, W. (2024). The Analysis of Trigger Factors of the Environmental Entrepreneurship Process in Saudi Arabia: An Innovative Approach. *Economies*, *12*(9), 254. https://doi.org/10.3390/economies12090254

Ren, S., Tomlinson, B., Black, R. W., & Torrance, A. W. (2024). Reconciling the contrasting narratives on the environmental impact of large language models. *Scientific Reports*, *14*(1), 26310. https://doi.org/10.1038/s41598-024-76682-6

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54-63. https://doi.org/10.1145/3381831

SemiAnalysis. (2024). AI training costs: GPT-4, Gemini, and beyond.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. 2022 international joint conference on neural networks (IJCNN),

Tabbakh, A., Al Amin, L., Islam, M., Mahmud, G. M. I., Chowdhury, I. K., & Mukta, M. S. H. (2024). Towards sustainable AI: a comprehensive framework for Green AI. *Discover Sustainability*, *5*(1), 408. https://doi.org/10.1007/s43621-024-00641-4

Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., Qu, Z., Yan, S., Zhu, Y., & Zhang, Q. (2023). Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.