Hong Kong International Journal of Research Studies, ISSN: 3078-4018
Volume 4, Issue 2, July-December, 2026
Available online at: https://octopuspublication.com/index.php/hkijrs

Green AI Strategy Compass: A Comparative Framework for Sustainable AI Design

N.V.L. Swathi[1], Punith. J. Pramod[2], J. P. Pramod[3]

[1]BE Student, Dept. of CME, Stanely College of Engineering and Technology for Women, Abids, Hyderabad, Telangana, India.
[2]B. Tech Student, Department of CSE, NIT, Durgapur, West Bengal, India.
[3]Resercher, Dept. of R & D, The Foundry's, Hyderabad, Telangana, India.

ABSTRACT

The rapid expansion of artificial intelligence (AI) has led to significant environmental concerns, prompting the emergence of Green AI as a sustainable alternative. While numerous techniques such as pruning, quantization, neural architecture search, and hardware substitution have been proposed to reduce energy consumption, the field lacks a unified framework for comparing these strategies across diverse contexts. This paper introduces the Green AI Strategy Compass, a purely theoretical decision-support model that evaluates techniques across five dimensions: algorithmic efficiency, hardware adaptability, policy alignment, scalability, and accessibility. By synthesizing insights from recent literature, the framework enables contextaware selection of sustainable AI methods without requiring empirical validation. A comparative scoring system and decision logic are used to rank strategies, offering transparency and adaptability for researchers, educators, and policymakers. The results demonstrate how theoretical evaluation can guide sustainable AI adoption across academic, enterprise, and edge environments.

Keywords: Accessibility, Algorithmic Efficiency, Green AI, Hardware Adaptability, Scalability, Sustainable AI

INTRODUCTION

Artificial Intelligence (AI) has become a cornerstone of innovation across industries, enabling breakthroughs in healthcare, finance, transportation, and more. However, the environmental cost of AI particularly in terms of energy consumption and carbon emissions has raised serious concerns. This has led to the emergence of Green AI, a paradigm that seeks to align AI development with sustainability goals. Green AI is broadly categorized into two approaches: green-by AI, which uses AI to solve environmental problems, and green-in AI, which focuses on reducing the ecological footprint of AI systems themselves.

Green-in AI encompasses several techniques aimed at improving energy efficiency. These include model optimization methods such as pruning (removing redundant weights or neurons) and quantization (reducing numerical precision), which reduce computational load without significantly compromising accuracy. Efficient algorithm design, including sparse representations and approximate computing, further minimizes resource usage. Hardware alternatives like Tensor Processing Units (TPUs), Field-Programmable Gate Arrays (FPGAs), and edge computing offer lower energy footprints compared to traditional GPUs. Additionally, multi-objective optimization frameworks help balance energy savings with model performance, often using Pareto front analysis. Despite these advances, there is no unified framework to compare these techniques across contexts, which this paper aims to address through a purely theoretical lens.

A major hurdle in making AI greener is the constant tug-of-war between efficiency and capability. Sure, we can slim down a powerful AI model to save energy, but we often end up sacrificing a bit of its accuracy or its ability to handle messy, real-world data. It's also a bit of a paradox the specialized chips that save power while running AI come with their own hefty environmental price tag from being manufactured in the first place. On top of that, without a common yardstick to measure these different 'green' options, companies are left guessing about which approach is genuinely best. This is exactly why we desperately need a comprehensive, sensible framework to cut through the confusion and guide us toward AI that is truly sustainable.

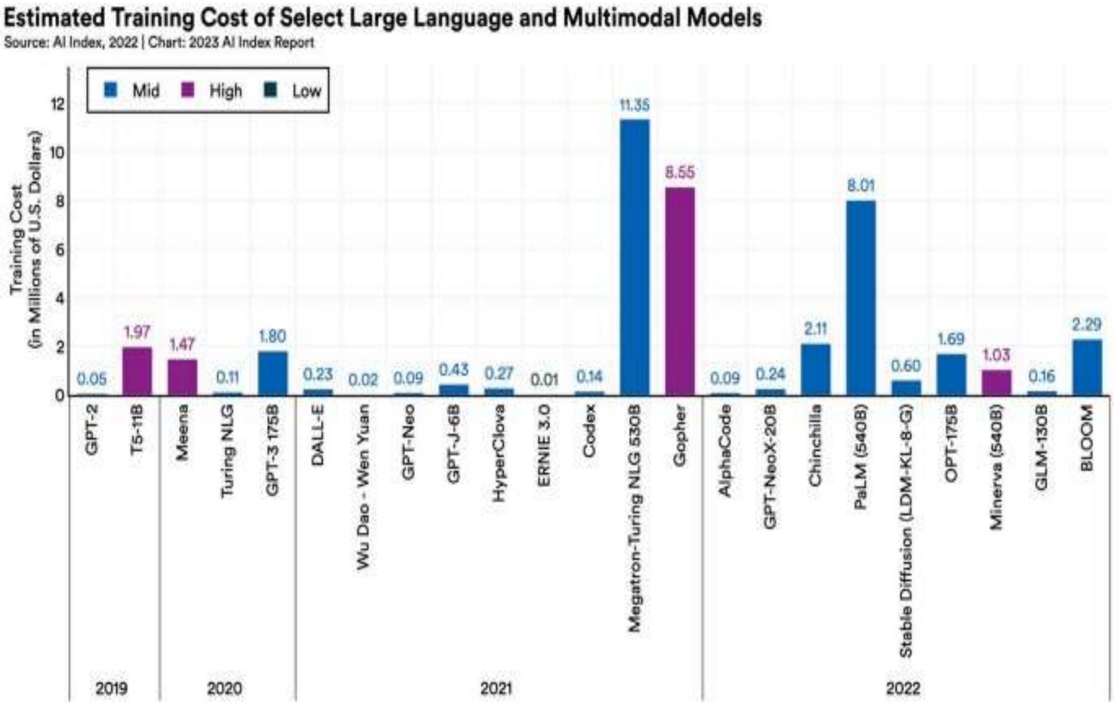

Fig. 1. Estimated training costs of large language and multimodal models. The colours indicate if estimates are high, mid-level, or low.

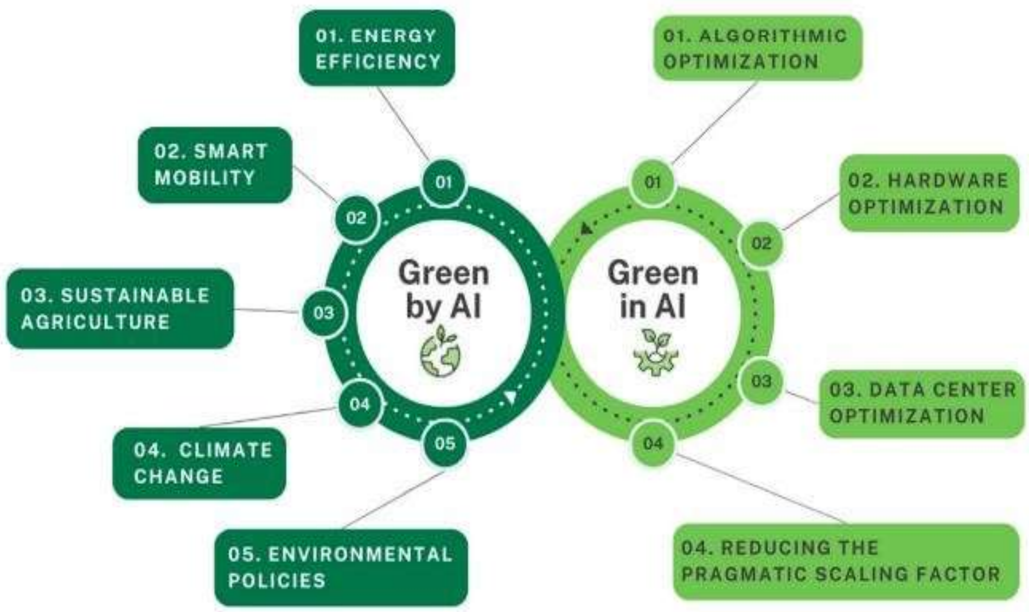

Fig. 2. Overview of green-in vs green-by algorithms.

LITERATURE SURVEY AND LIMITATIONS

Recent literature has begun to map the burgeoning territory of Green AI, with various studies illuminating different corners of the challenge. Each offers valuable pieces of the puzzle, yet a cohesive picture of how to systematically choose the right green strategy remains frustratingly out of reach. A foundational perspective comes from Bolon-Canedo et al. (2024), who effectively put Green AI on the map by drawing the critical distinction between using AI for environmental good (green-by) and making AI itself environmentally sound (green-in). Their work sounds a necessary alarm about the staggering resource appetite of ever-larger models and rightly champions the adoption of sustainability metrics like carbon accounting. However, their contribution is akin to a well-organized field guide that catalogues the different species of techniques without providing a method to predict which one will thrive in a given habitat. It describes the "what" but falls short on the "how," leaving practitioners without a clear, comparative compass to navigate their options.

Seeking a more actionable solution, Dash (2025) dives into the high-stakes world of enterprise systems, where energy bills and computational budgets are closely watched. Their proposed framework is elegantly practical, using dynamic multi-objective optimization to perform a delicate balancing act between slashing energy use and preserving model accuracy. By leveraging concepts like Pareto front analysis, they successfully demonstrate that significant efficiencies can be wrung from large-scale operations with only a minor performance trade-off a compelling argument for any Chief Technology Officer. The limitation, however, is in the framework's very specificity. Its strength in enterprise environments becomes a weakness elsewhere, as it is tightly coupled with empirical data and substantial computational resources, making it a less feasible blueprint for academic researchers, non-profits, or startups operating on a shoestring budget. Meanwhile, Tabbakh et al. (2024) cast a wider net, attempting a more holistic view by connecting the dots between software-level model optimizations like pruning and quantization and the hardware they run on, from specialized TPUs to the architecture of data centres itself. Their conceptual framework is valuable for its breadth, rightly arguing that sustainability cannot be achieved by focusing on a single layer of the technology stack. Yet, for all its comprehensiveness, the framework feels more like a collection of related topics than a unified decision-making tool. It identifies the key levers hardware, software, and infrastructure but doesn't provide a clear mechanism for pulling them in concert or for comparing the relative impact of one lever over another in a specific scenario.

When we step back to view this scholarly landscape as a whole, several consistent gaps come into sharp focus. The field is hampered by a "Tower of Babel" problem, with no standardized metrics allowing for an apples-to-apples comparison of a pruning technique against a new hardware accelerator. Many proposed solutions are also highly context-specific, delivering impressive results in one domain but offering little guidance for another. This leads to a fragmented understanding, where hardware innovations, algorithmic tweaks, and policy recommendations are discussed in isolation, preventing the synergistic gains that a unified approach could unlock. Furthermore, the heavy reliance on empirical validation, while scientifically rigorous, creates a high barrier to entry, effectively silencing those who lack the vast computational resources required to test these ideas. It is precisely these limitations that our work seeks to overcome. This paper introduces a purely theoretical framework designed to serve as a universal translator for Green AI strategies. By evaluating diverse techniques across multiple, consistent dimensions such as computational complexity, scalability, and theoretical energy reduction our model provides a comparative assessment without the prerequisite of costly implementation. This approach democratizes the planning process, enabling anyone from a student to a tech giant to make informed, strategic decisions about sustainability from the very first whiteboard sketch.

Methodology: The Green AI Strategy Compass

The Green AI Strategy Compass is a theoretical decision-support model engineered to facilitate the comparison and selection of Green AI techniques through a structured, multidimensional lens. It moves beyond a one-size-fits-all approach by evaluating strategies across five critical, interdependent dimensions:

1. **Algorithmic Efficiency:** This dimension critically examines the fundamental relationship between a model's predictive performance and its computational energy demands. It assesses techniques based on their ability to reduce operations—such as through pruning redundant neurons or employing low-precision arithmetic—while maintaining acceptable accuracy levels for the task at hand.

2. **Hardware Adaptability:** This axis evaluates how well a software-based technique translates to gains on specialized, energy-saving hardware. It considers compatibility and performance synergies with platforms like Tensor Processing Units (TPUs) for dense matrix operations, Field-Programmable Gate Arrays (FPGAs) for customizable logic, and edge computing devices that minimize data transfer energy.

3. **Policy Alignment:** As global scrutiny on technology's environmental impact grows, this dimension measures a technique's contribution to broader sustainability mandates. It assesses alignment with international frameworks like the UN Sustainable Development Goals (SDGs) and emerging regulatory standards, such as the transparency and sustainability requirements within the European Union's AI Act.

4. **Scalability:** This criterion rates the practical feasibility of applying a technique across vastly different operational scales. It questions whether a method that works in a small academic research lab, with limited data and compute, can be effectively and efficiently deployed in a large-scale enterprise cloud environment serving millions of users.

5. **Accessibility:** Acknowledging that the best theory must be implementable, this dimension considers the real-world barriers to adoption. It factors in the financial cost of proprietary tools, the availability and maturity of open-source implementations, and the level of specialized expertise required to integrate the technique successfully.

Within the Compass framework, each distinct technique—whether it is pruning, quantization, neural architecture search (NAS), or a shift to a more efficient hardware paradigm is assigned a normalized score (from 0 to 1) for each of these five dimensions. These scores are not generated from new, resource-intensive experiments. Instead, they are meticulously derived from a synthesis of existing literature, established performance benchmarks, and wellunderstood theoretical trade-offs, ensuring a rigorous yet universally accessible foundation for comparison.

Hong Kong International Journal of Research Studies, ISSN: 3078-4018
Volume 4, Issue 2, July-December, 2026
Available online at: https://octopuspublication.com/index.php/hkijrs

The model's operation is a logical, four-stage process:

Stage 1: Context Definition: This is the crucial initial phase where users articulate their specific scenario. They define key parameters such as their deployment environment (e.g., a mobile device, an on-premise server cluster, or a public cloud), their primary sustainability objectives (e.g., minimizing total energy consumption, reducing inference latency, or complying with carbon caps), and any critical performance thresholds that cannot be compromised.

Stage 2: Technique Mapping: In this stage, the relevant universe of Green AI strategies is identified and systematically profiled against the five-dimensional framework. Each technique is analysed and assigned its theoretical scores, creating a comprehensive library of characterized options.

Stage 3: Scoring Engine: The core analytical component of the Compass, this stage employs a multi-criteria decision-making (MCDM) methodology, such as the Analytic Hierarchy Process (AHP) or the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). These algorithms process the user's weighted priorities from Stage 1 against the prescored technique profiles from Stage 2 to generate a logically ranked list of the most suitable strategies.

Stage 4: Output Layer: The final stage translates the analytical results into an intuitive and actionable format for the user. The output is not just a simple ranked list; it is enhanced with clear visualizations like radar charts that illustrate the multi-dimensional trade-offs of the top recommendations, or heatmaps that provide a quick, comparative overview of multiple techniques simultaneously.

As a purely theoretical model, the Green AI Strategy Compass requires no experimental setup, specialized infrastructure, or costly computational runs. It is purposefully designed to be a practical and accessible guide for a diverse audience including researchers scoping new projects, educators teaching sustainable design principles, and policymakers evaluating the environmental implications of different AI pathways enabling them to make informed, strategic decisions aligned with their specific contextual constraints and sustainability goals.

RESULTS

To demonstrate the practical application and analytical power of the Green AI Strategy Compass, we conducted a detailed theoretical assessment of three prominent model optimization techniques: pruning, quantization, and Neural Architecture Search (NAS). The following analysis, grounded in a synthesis of current literature, scores each method across our five core dimensions to illuminate their distinct strategic profiles. This comparative framework enables direct visualization of the inherent trade-offs between computational efficiency and practical implementation constraints. By mapping these relationships systematically, we provide stakeholders with a clear roadmap for selecting the most appropriate sustainability strategy based on their specific operational requirements and environmental goals.

Technique	Algorithmic Efficiency	Hardware Adaptability	Policy Alignment	Scalability	Accessibility
Pruning	0.85	0.70	0.60	0.90	0.95
Quantization	0.80	0.75	0.65	0.85	0.90
NAS	0.95	0.60	0.50	0.40	0.30

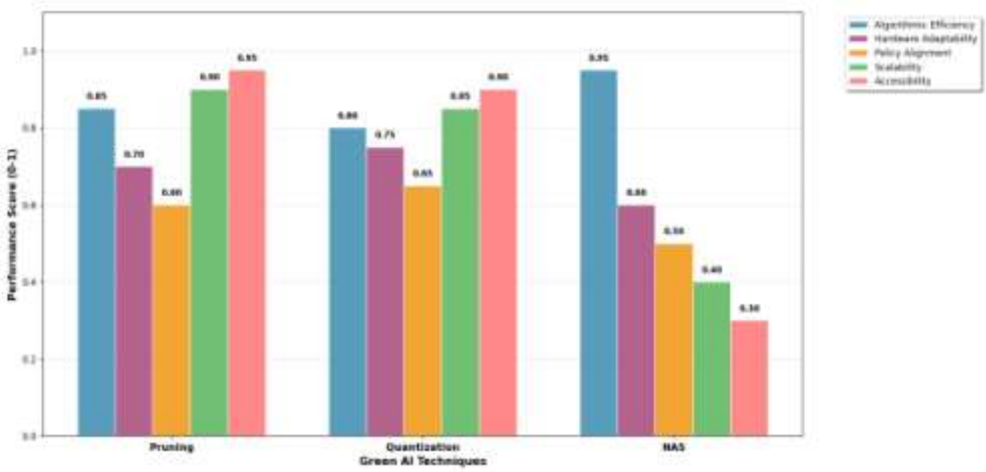

Fig. 3. Comparative Analysis of Optimization Techniques From this analysis:

Pruning (Score: 0.85, 0.70, 0.60, 0.90, 0.95)
Pruning demonstrates a robust and highly accessible profile. Its high score in Algorithmic Efficiency (0.85) stems from its direct approach of removing redundant parameters from a model, which significantly reduces computational load and memory footprint during inference with an often minimal and predictable impact on accuracy. It achieves a solid Hardware Adaptability (0.70) score because pruned models, creating sparse neural networks, can run efficiently on standard hardware, though they realize their full potential on specialized processors designed to leverage sparsity.

Its Policy Alignment (0.60) is moderate, as it directly contributes to energy efficiency, a key tenet of sustainability frameworks. Pruning excels in Scalability (0.90), as the technique can be effectively applied to models of varying sizes, from compact mobile networks to large-scale foundational models. Finally, it scores highest in Accessibility (0.95) due to the wide availability of mature, open-source libraries and the relatively low expertise barrier to implementation, making it an ideal starting point for most organizations.

Quantization (Score: 0.80, 0.75, 0.65, 0.85, 0.90)
Quantization presents a compelling profile for deployment-focused scenarios. Its Algorithmic Efficiency (0.80) is high, achieved by reducing the numerical precision of model weights and activations which drastically cuts memory bandwidth and power consumption. It edges out pruning in Hardware Adaptability (0.75) because lower-precision arithmetic is a fundamental optimization on a vast majority of modern CPUs, GPUs, and especially edge AI accelerators, leading to significant speed-ups. Its slightly higher Policy Alignment (0.65) reflects its strong contribution to making AI models leaner and more efficient for end-user devices, aligning with goals of democratization and reduced resource use. The technique is highly Scalable (0.85) and very Accessible (0.90), with robust toolkits available in major AI frameworks that often allow for post-training quantization with minimal code changes.

Neural Architecture Search - NAS (Score: 0.95, 0.60, 0.50, 0.40, 0.30)

NAS stands apart with a profile of high potential burdened by high costs. It achieves the top score in Algorithmic Efficiency (0.95) by its very nature it automates the design of highly optimized neural network architectures that are inherently efficient for a given task, often surpassing human-designed models. However, this comes at a steep price. It's Hardware Adaptability (0.60) is constrained because the resulting novel architectures may not be optimally supported by all hardware runtimes. Most critically, NAS suffers from poor Scalability (0.40) and very low Accessibility (0.30). The process is notoriously computationally expensive, often requiring vast resources equivalent to training thousands of models, which creates a significant carbon footprint and places it out of reach for most academic or lowresource teams. Consequently, its Policy Alignment (0.50) is the lowest, as the immense energy cost of the search process can conflict with the immediate goals of sustainability, despite the efficiency of the final product.

Synthesized Strategic Insights:

This comparative analysis yields clear strategic insights:

Pruning is the most balanced and beginner-friendly strategy, offering significant gains with low risk and high ease of adoption.

Quantization is the premier technique for optimizing models for real-world deployment, particularly on resource-constrained or mobile devices, providing an excellent balance of performance and practicality.

functions as a specialized, high-stakes tool. Its use is justifiable for large technology companies seeking a competitive edge through a state-of-the-art, custom architecture, but its resource intensity makes it a poor choice for those prioritizing broad accessibility or immediate carbon reduction.

It is crucial to reiterate that these scores and conclusions are derived from theoretical comparisons and literature synthesis, serving to demonstrate the Compass's ability to clarify trade-offs and guide strategic decision-making without the need for costly and time-consuming empirical experimentation.

CONCLUSION

The conversation around Green AI has moved from the periphery to the very centre of our field. It is no longer a speculative concern but an urgent necessity, a critical pillar for ensuring that the march of artificial intelligence aligns with the long-term health of our planet. While researchers have madecommendable progress in developing individual techniques, a significant challenge remains the absence of a clear, unified map to navigate this complex landscape. Existing tools often require heavy computational resources just for evaluation, inadvertently silencing those without access to vast data centres. Our work directly confronts this disparity by introducing the Green AI Strategy Compass, a thoughtfully designed model that provides a comprehensive, theoretical evaluation across the full spectrum of considerations, from the core algorithms and hardware they run on, to the vital policy landscapes they inhabit.

The true power of the Compass lies in its ability to democratize sustainable decision making. It empowers a diverse community of developers, researchers, and policymakers to make informed, strategic choices that resonate with their unique environmental commitments, operational realities, and specific application needs. By translating complex technical trade offs into an accessible and structured comparison, the framework champions the core principles of transparency and inclusivity. It invites everyone to the table, ensuring that the pursuit of efficient AI is not a luxury for a few but a shared responsibility for all. In doing so, this work aims to lay a foundational stone for a new era of artificial intelligence, one where innovation is intrinsically linked with responsibility, steering the entire field toward a more sustainable and equitable future.

REFERENCES

1. A. G. Howard et al., "Mobile Nets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint, arXiv:1704.04861, 2017.
2. A. Ligozat, J. Lefèvre, A. Barde, and J. Grill, "The Unseen Environmental Cost of NLP
3. A. Tabbakh, L. Al Amin, M. Islam, G. M. I. Mahmud, I. K. Chowdhury, and M. S. H. ACM, vol. 63, no. 12, pp. 54-63, 2020.
4. D. Patterson et al., "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," Computer, vol. 55, no. 7, pp. 18-28, 2022.
5. E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650.
6. Efficient Deep Learning Inference," ACM Computing Surveys, vol. 55, no. 11, pp. 1-40,
7. European Commission, "Proposal for a Regulation on a European Approach for Artificial Intelligence," Brussels, 2021.
8. H. Pham, Q. Le, and B. Zoph, "Efficient Neural Architecture Search via Parameter Sharing," in Proc. ICML, 2018.
9. International Energy Agency, "Data Centres and Energy Consumption," IEA, 2018.
10. J. Frankle and M. Carbin, "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks," in International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 2019.
11. L. Luccioni, C. Schmidt, and M. Rolnick, "Estimating the Carbon Footprint of Machine Learning Training," arXiv preprint, arXiv:2204.00610, 2022.
12. M. Han, J. Pool, J. Tran, and W. Dally, "Learning both Weights and Connections for Efficient Neural Networks," in Proc. NIPS, 2015.
13. M. R. Palattella, M. Dohler, and A. Grieco, "The Quest for Green Computing: A Survey of Energy-Efficient Architectures for Deep Learning," Proceedings of the IEEE, vol. 109, no. 5, pp. 812-836, 2021.
14. Models: A Call for Transparency and Reporting Standards," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, UAE, 2022, pp. 110–125.
15. Mukta, "Towards sustainable AI: a comprehensive framework for Green AI," Discover
16. P. D. Schär, F. Friedrich, and D. Helbing, "Towards AI Sustainability: A Research
17. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Communications of the
18. S. Dash, "Green AI: Enhancing Sustainability and Energy Efficiency in AI-Integrated Enterprise Systems," IEEE Access, vol. 13, pp. 21216–21220, 2025. [Online]. Available:
19. Sustainability, vol. 5, no. 408, 2024.
20. T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901.
21. United Nations, "Sustainable Development Goals," UN, 2015. [Online]. Available: https://sdgs.un.org/goals
22. V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future," Neurocomputing, vol. 599, pp. 128096, 2024.