

The Path to Green AI – Challenges and Solutions

Nikhil Bhadra

Staples High School, 70 North Avenue, Westport, CT-06880, USA; nbhadra60@gmail.com

ABSTRACT: Recent growth in Artificial Intelligence (AI) systems has caused an unprecedented demand for large-scale, high-speed computing resources. This has led to increased concerns about the sustainability of AI development and its environmental impact. This study examines different strategies through a literature review and uses key metrics such as efficiency, cost, and feasibility to compare them. The findings revealed that software solutions such as algorithm optimization, the use of small language models, and data curation achieve the highest efficiency by reducing computational demand. Hardware solutions provide a baseline level of efficiency while workload optimization minimizes energy consumption through distributed learning and energy-aware scheduling. Finally, infrastructure solutions offer a downstream option to minimize energy use in data centers through efficient thermal management. The study also emphasizes challenges in standardizing carbon reporting and establishing enforceable regulatory systems and stresses that AI researchers, data center engineers, and energy policy experts must collaborate to translate technical efficiency gains into real-world emissions reductions for sustainable AI development.

KEYWORDS: Green AI, Hardware Optimization, Algorithm Optimization, Small Language Model, Data Curation, Workload Optimization, Grid Optimization.

■ Introduction

AI has become an important component of modern society. As AI capabilities evolve, the power required to train large models has increased faster than improvements in energy efficiency, causing a significant impact on the environment.¹

This paper addresses the critical question: Can we keep up with the growing demand for AI without negatively impacting the environment?

Green AI has emerged as the field to address this challenge. It refers to the development and use of AI systems that reduce energy consumption, carbon emissions, and computational waste while maintaining strong performance. It aims to address the rising environmental and economic costs of training and deploying large AI models. Although Green AI is still an emerging field, it has gained increasing attention from researchers, technology companies, and policymakers.

The present study aims to investigate various green AI strategies by comparing them against key metrics such as efficiency, cost, and feasibility. By synthesizing current strategies and reviewing their merits and shortcomings, this work will contribute to current efforts in Green AI. While previous literature has extensively discussed each strategy in isolation, this study aims to provide a comparative analysis of the strategies through both literature review and qualitative diagnostics, to examine their combined benefits in reducing energy consumption.

By analyzing these approaches alongside macro- and policy-level discussions, this study aims to emphasize the distinct natures of Green AI strategies and their collective contribution to sustainable AI development.

■ Green AI Solutions

Sustainable AI development is a multi-solution approach built on different concepts and technologies. They can be

categorized into three main areas: technical, operations, and infrastructure. We will look at the different options under each category and compare them.

Technical Solutions:

Technical solutions include a combination of hardware and software solutions that can provide significant energy savings.

1. Hardware Optimization:

Hardware optimization covers efficiency techniques such as specialized accelerators, low-power AI chips, and application-specific integrated circuits (ASICs).

Specialized Hardware Accelerators:

Special hardware accelerators, like Tensor Processing Units (TPUs), achieve much higher performance per watt than regular GPUs by optimizing the matrix-multiplication architecture. Experimental and production deployments have shown that TPUs can improve inference performance per watt by up to 15×, with gains of 25–30× TOPS/Watt across different neural network architectures.²

Low-Power AI Chips and Edge Accelerators:

Low-power AI hardware such as NVIDIA's Jetson Nano and Google's Coral Edge TPU can do about 4 trillion operations per second (TOPS) while using only 2 watts of power.³ Compared to regular 64-bit ARM CPUs, Edge TPUs consume less energy while keeping good classification accuracy for deep feedforward neural networks.

Application-Specific Integrated Circuits (ASICs):

ASICs are custom-made for specific AI tasks, enabling significant improvements in both performance and power con-

sumption. ASICs use much less energy than GPUs for steady, repetitive AI tasks.⁴

Table 1 provides a comparison of different hardware optimization strategies. Hardware optimization by itself only provides baseline efficiency. We will discuss next the different software solutions that reduce the need for general-purpose computation and make hardware solutions more effective.

Table 1: Comparison of different hardware optimization strategies.

Hardware Type	Primary Use Case	Energy Efficiency	Flexibility	Cost
GPU	Training & Inference	Low	High	Medium
TPU / Accelerator	AI-specific workloads	High	Medium	Medium
Edge AI Chips	On-device inference	Very High	Low	Low
ASIC	Fixed AI workloads	Very High	Very Low	Very High

Table 1 - Trade-off between energy efficiency, flexibility, and cost for different hardware optimization solutions. ASICs and edge accelerators deliver the highest energy efficiency. However, they offer limited flexibility and are mainly used for well-defined workloads. GPUs, on the other hand, offer high flexibility but incur significant energy costs.

2. Algorithm Optimization:

As AI systems continue to scale, algorithm-level optimization, for both training and inference, has emerged as a critical pathway toward sustainable AI.

Training Optimization:

Energy efficiency during the training phase can be improved through data-level, model-level, and algorithm-level optimization strategies.⁵ At the data level, **Stochastic mini-batch dropping (SMD)** reduces computational overhead by selectively skipping less informative mini-batches during training. At the model level, **Input-dependent selective layer update (SLU)** minimizes redundant computation by updating only the most relevant layers for a given input. At the algorithm level, **Predictive sign gradient descent (PSG)** reduces training energy costs by approximating gradient directions rather than computing exact gradient values.⁵ Complementary techniques, such as **Adaptive learning rate scheduling** and **Early stopping**, further reduce energy usage by avoiding unnecessary computation.⁵

Inference Optimization:

Inference optimization strategies include **Parameter pruning**, which removes redundant or low-importance weights from neural networks. **Quantization** lowers numerical precision, enabling faster execution with minimal impact on accuracy. Additional techniques, such as **Convolutional filter compression** and **Matrix factorization**, simplify network operations, while **Network architecture search (NAS)** automatically identifies model structures optimized for efficiency.⁶ **Knowledge distillation** further improves efficiency by transferring knowledge from large “teacher” models to smaller “student” models, enabling comparable performance at a fraction of the computational and energy cost.⁶

Table 2 provides a comparison of the different algorithm optimization solutions. While algorithm optimization is beneficial, it is still extensive in the context of large language models.

In the next section, we will discuss small language models for specific tasks, which may significantly limit the scope of training by restricting the model's size.

Table 2: Comparison of different training and inference strategies.

Phase	Optimization Technique	Primary Objective	Energy Efficiency
Training	Stochastic Mini-Batch Dropping (SMD)	Reduce the number of training samples per iteration	Medium
Training	Selective Layer Update (SLU)	Avoid unnecessary layer updates	High
Training	Predictive Sign Gradient Descent (PSG)	Simplify gradient computation	Medium
Inference	Parameter Pruning	Remove redundant weights	High
Inference	Quantization	Reduce numerical precision	High
Inference	Knowledge Distillation	Transfer knowledge to smaller models	High

Table 2 - Primary objective and energy efficiency of different training and inference optimization solutions. Most strategies provide significant energy savings when properly implemented and when the primary objective is clear.

3. Small Language Models:

Small language models reduce computing power, memory use, and training time by reducing the number of parameters and optimizing architectural design. DistilBERT, a distilled version of BERT, is 40% smaller and retains 97% of its language understanding while being 60% faster.⁷ SLMs are increasingly being used in on-device intelligence systems, requiring less data transfer and thereby using less energy and providing better data protection.

Training Efficient Models from Scratch:

One approach to obtaining SLMs involves training models from scratch using efficiency-oriented architectures. There are three strategies: **1. Efficient feed-forward networks (FFNs)**, which reduce parameter redundancy in dense layers; **2. Efficient attention mechanisms**, which mitigate the quadratic complexity of standard self-attention through sparse or linear approximations; and **3. Attention-free architectures**, which replace attention mechanisms with alternative sequence modeling techniques.⁸

Model Compression from Large Language Models:

An alternative approach derives SLMs from pretrained LLMs using model compression techniques such as **Quantization**, **Pruning**, and **Knowledge Distillation**, as discussed in the Algorithm Optimization section.⁸

Collaborative SLM-LLM Frameworks:

SLMs and LLMs can be combined, with LLMs providing general thinking and language knowledge, while SLMs act as experts for domain-specific tasks.⁸ This division reduces the need for repeated fine-tuning of large models and lowers overall computational cost.

Optimizing training and limiting model size can save significant computing power. In the next section, we will review another strategy for curating training data and reducing its size to improve training efficiency.

4. Data Curation:

Large Language Models (LLMs) are commonly trained on massive web-scale corpora such as C4, Refined Web, and The Pile. However, prioritizing data quality through curated, smaller datasets can enable effective LLM training with lower energy consumption.⁹

Data Curation as an Energy Saving Strategy:

Data curation focuses on extracting high-quality and informative subsets from large corpora. By doing so, models can be trained on smaller yet more representative datasets that facilitate learning important linguistic patterns without unnecessary tokens. Subramanyam *et al.*¹⁰ through experimental work, proposed a new quality-aware scaling law that extends the widely used Chinchilla framework, predicting pretraining loss as a joint function of model size, data volume, and data quality.

Table 3 compares the size and the number of hours needed to train a model. Approximately 15 to 20 domain-specific models can be trained with the same amount of energy needed to train a single LLM. Focusing on building smaller, domain-specific models rather than general-purpose LLMs can result in significant energy savings.

Table 3: Comparison of hours and energy needed to train models of different sizes.

Dataset Type	Size (B Tokens)	GPU Hours	Estimated Energy (MWh)	Model Type
Large-scale Web Dataset (Uncurated)	300	50,000	2,500	Full C4/Pile-style corpora
Curated High-Quality Subset	50	8,000	400	SLM-evaluated, smaller but informative
Task-Specific Curated Dataset	20	3,000	150	Domain-specific, highly filtered

Table 3 - Estimated GPU hours and associated energy usage for training different types of datasets. Large-scale web datasets are typically used to train LLMs like GPT-3 (300B tokens). SLMs and other domain-specific models need much smaller datasets and considerably less energy for training.

Operations Solutions:

In this section, we will discuss the benefits of distributed learning and proactive scheduling to maximize the efficiency of training models.

1. Workload Optimization:

Federated Learning:

Federated Learning (FL) creates machine learning models from data distributed across many devices. Each device uses its local data to update the model, and only that update is sent back to the server.¹¹ FL also makes it simpler to follow data privacy and security rules as data stays localized.¹²

Energy-Aware Scheduling:

Energy-aware scheduling techniques consider power consumption, cooling requirements, and workload characteristics to optimize energy use.¹³ They maintain workload profiles to schedule tasks during periods when energy demand is low or when large amounts of renewable energy are available. P. Krishnamurthy did a study where, at a peak utilization setting of 0.69%, the proposed workload optimization framework

achieved a 13% time-saving efficiency compared to static scheduling, corresponding to an estimated energy savings of 364 MW for LLMs with over 10 billion parameters.¹⁴

Infrastructure Solutions:

Infrastructure solutions provide the most downstream optimization solution in the Green AI framework. In the following sections, we will discuss strategies that can provide efficient solutions for managing the increasing workload in AI data centers.

1. Efficient Heat Management:

Cooling systems at data centers consume nearly 50% of electricity. Techniques such as immersion and liquid cooling are increasingly being adopted in place of traditional air-cooling systems to manage the high cooling requirements and associated power demand in AI data centers.

Cooling Strategies:

Immersion cooling submerges servers or components in a thermally conductive, electrically insulating dielectric fluid. Heat is directly absorbed by the fluid, eliminating the need for fans or air-based heat removal. Immersion cooling can offer energy efficiency gains of 40% compared to conventional air cooling.¹⁵

Liquid cooling circulates a coolant, often a mixture of ethylene glycol and deionized water, directly to heat-generating components via cold plates. Heat generated by the equipment is efficiently conducted through the liquid, which is then transported to the heat exchanger, where it is dissipated via the external cooling system. Liquid cooling can offer energy efficiency gains of 60-90% compared to conventional air cooling.¹⁶

Other Heat Management Strategies:

Microfluidic cooling delivers coolant directly inside silicon chips via etched microchannels. By removing heat at its source, it dramatically reduces thermal resistance compared to conventional cold-plate solutions. Rajan *et al.*¹⁷ used an Intel Core i7-8700K CPU with micropin fins etched directly on the back and overclocked it to dissipate up to 215W of power while cooled by room-temperature de-ionized (DI) water. The experiment demonstrated up to 44.4% reduction in junction-to-inlet thermal resistance while using only 0.3× of volumetric coolant flow per Watt of power dissipated in the CPU compared to a conventional cold plate.¹⁷

Waste Heat Reuse (WHR), on the other hand, improves data center sustainability by redirecting excess heat to external applications, such as district heating, hot water supply, or industrial processes.¹⁸ Case studies like Stockholm Data Parks (Sweden) and the University of Notre Dame (USA) demonstrate that WHR not only reduces environmental impact but also creates economic value from otherwise wasted thermal energy.

The largest gains in cooling efficiency come from more advanced techniques, such as liquid and immersion cooling, as shown in Table 4. However, they provide diminishing re-

turns when considered in isolation, as these gains come with increased system complexity and deployment challenges. Cooling in general is a downstream optimization layer whose effectiveness depends on upstream computational demand.

Table 4: Comparison of different data center cooling strategies.

Cooling Method	Efficiency	Complexity	Sustainability	Typical Use Case
Air Cooling	Low	Low	Minimal	Standard servers
Liquid Cooling	Very High (approx. 60% energy savings)	Medium	Moderate	AI racks
Immersion Cooling	High (20–40% energy savings)	Medium	Moderate	AI clusters
Microfluidic Cooling	Very High	High	Moderate	Next-gen AI chips

Table 4– Complexity and Sustainability of various data center cooling strategies. Liquid Cooling and Immersion Cooling provide the best balance of complexity and sustainability, whereas next-generation solutions like Microfluidic Cooling are still technically complex and difficult to implement at scale.

2. Alternate Strategies:

As rack power densities in AI data centers continue to increase, alternative approaches that completely reimagine data center placement and cooling environments are gaining attention. **1. Underwater data centers** leverage the naturally low and stable temperatures of deep ocean environments to dissipate their heat efficiently.¹⁹ **2. Space-based data centers** deploy computational infrastructure into Earth's orbit using solar power as a primary energy source and radiative cooling to dissipate heat directly into space.²⁰ **3. High-altitude platforms** benefit from passive thermal dissipation due to low ambient temperatures, as they operate in the stratosphere at altitudes of 18–25 km.²¹ **4. Intelligent control mechanisms** reduce cooling energy consumption by using machine learning models to predict thermal behavior.²²

Table 5 illustrates a clear distinction between theoretical efficiency and deployment feasibility for alternate infrastructure solutions. Solutions that can be integrated within existing infrastructure constraints provide the best path to sustainable AI when compared with the most extreme or novel solutions.

Table 5: Comparison of different alternate infrastructure strategies.

Approach	Cooling Demand	Energy Efficiency	Deployment Feasibility
Conventional Terrestrial DC	High	Low	High
Underwater DC	Very Low	High	Medium
Space-Based DC	Minimal	Very High	Low
High-Altitude DC	Low	Moderate	Medium
AI-Driven Cooling	Moderate	High	High

Table 5 - Energy efficiency and deployment feasibility for various alternate infrastructure solutions. Space-based and underwater data centers offer extremely high cooling efficiency. However, they are unlikely to make a significant contribution to near-term sustainability goals due to their low feasibility and high cost. In contrast, AI-driven cooling and high-altitude platforms offer more balanced trade-offs between efficiency and deployability.

■ Comparative Trade-Off Analysis

Each Green AI strategy presents distinct benefits and limitations, and no single approach provides the best optimization

solution across efficiency, cost, and feasibility. Rather than competing, these approaches provide the best results when applied collectively within a Green AI framework.

Table 6 compares all major Green AI strategies across multiple dimensions. To maximize overall system sustainability, a layered optimization framework is needed in which software efficiency reduces demand and enables more effective use of hardware and infrastructure solutions.

Table 6: Comparison of different Green AI strategies.

Solution	Efficiency	Feasibility	Cost	Complexity	Current Adoption	References
Hardware Optimization (GPUs, TPUs, ASICs)	High	High	Medium–High	Medium–High	Growing (production use)	[2], [3], [4]
Algorithm Optimization (Training, Inference)	High	High	Low–Medium	Medium	Moderate–High	[5], [6]
Small Language Models (SLMs)	Very High (task-specific)	High	Low–Medium	Medium	Growing	[7], [8]
Data Curation	Medium–High	High	Low–Medium	Medium	Early Growing	[9], [10]
Workload Optimization	Medium	Medium–High	Medium	High	Early	[11], [12], [13], [14]
Efficient Heat Management	High	Medium–High	High	High	Moderate	[15], [16], [17], [18]
Alternate Strategies	Very High (theoretical)	Low–Medium	Very High	Very High	Experimental	[19], [20], [21], [22]

Table 6 - Green AI strategies compared against efficiency, feasibility, cost, and complexity. Hardware optimization, algorithm optimization, and efficient heat management provide the best near-term benefits with a moderate to high level of adoption. Data curation and workload optimization show early promise and are expected to be increasingly adopted. Alternate strategies provide the best efficiency benefits but are limited by extremely high complexity and cost.

■ Macro Analysis

This section analyzes sustainability challenges at the macro level by comparing current and future energy use for the two global AI powers, the US and China.

Table 7 and Figure 1 show the current and projected data center energy consumption in the U.S. and China. In both countries, increased scale and adoption of AI are projected to outpace efficiency gains. The primary energy source in the US is natural gas, and in China it is coal. The same computational workload, therefore, can yield significantly different carbon emissions in these two countries based on their underlying energy infrastructure. The adoption of renewable energy and grid optimization, therefore, must be integrated with technical efficiency to enable sustainable AI in both countries.

Table 7: Annual electricity consumption of data centers in the US and China. Data sourced from Pew Research Center and Bloomberg NEF-reported capacity estimates.

Metric	United States	China
2024 Base Energy Use	183 TWh (Approx. 4% of U.S. total)	Approx. 20 GW aggregate (national)
2030 Projection	Approx. 426 TWh (approx. +133%)	Approx. 40 GW (approx. +100% from 2024)
Share of Electricity Demand	Potentially 8–12% by late decade (industry estimates)	Approx. 3–6% of the national grid by 2030 (industry & policy forecasts)
Contribution to Global Demand Growth	Largest single national driver	Second largest driver

Table 7 - Energy consumption by U.S. and China data centers. In both countries, the energy usage is projected to more than double by 2030. The share of overall electricity demand is still low, but is expected to increase by 2030 in both countries.

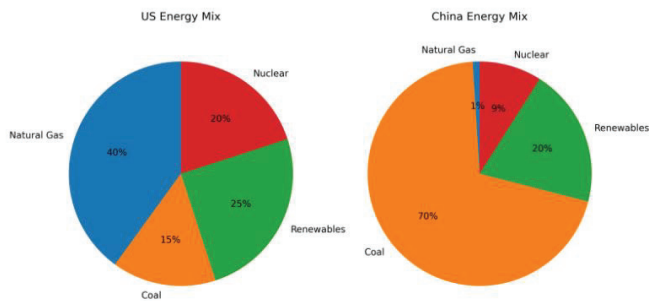


Figure 1: Energy source mix for data centers in the US and China reported by the Pew Research Center and the IEA.

Figure 1 - US data centers get 25% of their energy from renewable sources, while data centers in China get 20%. Natural gas is the largest source of power for US data centers at 40%, while in China, it's coal at 70%. Both countries plan to invest heavily in renewables and nuclear energy for powering AI data centers in the future.

US Strategy:

US energy companies and data center businesses are worried about the strain on the power grid. Delays in upgrading transmission lines are a major obstacle to increasing capacity to support AI data centers. Large companies like Microsoft are promising to pay full energy rates and replace the water used for cooling their data centers. Google and other companies have agreed with energy providers to reduce data center power use during periods of high grid demand. There is growing interest in increasing nuclear and renewable energy use in the US.

China Strategy:

China is focusing on quickly building energy grids and infrastructure to support energy-intensive AI data centers, without incurring long delays due to regulations. Regional plans are increasingly linking data center growth with local renewable energy sources. Areas with abundant renewable energy are given priority for large-scale facilities. Data centers in China are aiming for Power Usage Effectiveness (PUE) benchmarks (e.g., ≤ 1.25) to improve energy efficiency compared to global standards.

Results and Discussion

Implementation of a sustainable AI strategy involves the integration of multiple complementary strategies rather than the deployment of any single solution. Tables 1 to 4 list the different optimization strategies along with their merits and shortcomings. Hardware optimization improves performance per watt and provides baseline efficiency, but is constrained by flexibility and cost. Algorithmic optimization and small language models provide better scalable efficiency gains by directly reducing computational demand. Data curation reduces unnecessary training overhead by prioritizing quality

over quantity. Workload optimization extends efficiency into deployment. Cooling and infrastructure strategies act as downstream enablers whose effectiveness depends on the upstream computational workload.

Table 5 highlights that at the macro level, rising global AI energy demand will outpace isolated efficiency gains. Green AI, therefore, must be understood as a unified framework where grid optimization and increased use of renewable energy are as important as software, hardware, and infrastructure optimizations.

Future studies are needed to assess the real-world feasibility and to quantify the combined effects of multiple Green AI strategies deployed simultaneously. Additional research is also needed on standardized reporting of energy and carbon costs. AI researchers, data center engineers, and energy policy experts need to collaborate to translate technical efficiency gains into real-world emissions reductions.

Conclusion

The paper is primarily based on a literature synthesis from previous work in this field. Green AI is an evolving field, and sufficient empirical data from experiments is not available to analyze the combined effect of multiple Green AI strategies in real-world settings. Hence, most of the comparative analysis in this paper is based on qualitative diagnostics. The optimization solutions listed mainly focus on technical and infrastructure efficiencies and do not fully quantify lifecycle carbon emissions or economic trade-offs. In addition, rapidly evolving hardware and AI models may render some findings outdated and pose new challenges. Finally, regional comparisons are limited to the U.S. and China, the two major AI powerhouses. Increased AI adoption globally may need a more holistic evaluation of broader energy usage and policy environments.

Conclusion

The rapid expansion of AI has placed considerable computational and energy demands, raising major concerns about the sustainability of current AI development practices. Hardware and algorithm optimization, small language models, data curation, workload management, and advanced thermal regulation systems provide distinct but complementary approaches to reducing the environmental impact of AI. The best approach involves implementing a layered optimization framework that incorporates all the solutions discussed, with software efficiency providing the maximum benefit by reducing overall computational demand.

Though progress has been made in the Green AI initiative, major challenges remain in standardizing carbon reporting, ensuring accuracy with efficiency, and establishing enforceable regulatory systems.

Governments, industry, and academia are all investing significant time and effort in Green AI. AI workload will continue to increase, but with the combined effort from all parties, the challenge for sustainable AI development can be met, and future AI growth can be achieved without negatively impacting the environment.

■ Acknowledgments

I would like to thank my mentor, Dr. Samuel Lefcourt, PhD, from Johns Hopkins University, for his guidance and support in helping me write my first academic research paper. Dr. Lefcourt provided valuable input and consistently challenged me on my thesis, prompting me to think more deeply, conduct further research, and be more innovative in my approach. I would also like to thank Joe Xiao for guiding me through the editing process. Finally, I want to acknowledge my parents; without their encouragement, I would not have been able to complete this paper.

■ References

- Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 3645–3650, doi:10.18653/v1/P19-1355.
- Pothukuchi, Nikhila. (2025). "Understanding Tensor Processing Units: The Specialized Hardware Revolutionizing AI Computing." International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 11. 2349-2357. 10.32628/CSEIT23112565.
- Noorbin, Faezeh & Layeghy, Siamak & Kusy, Brano & Jurdak, Raja & Portmann, Marius. (2023). "Exploring Edge TPU for deep feed-forward neural networks." Internet of Things. 22. 100749. 10.1016/j.iot.2023.100749.
- Wang, Yujie. (2023). "Artificial-Intelligence integrated circuits: Comparison of GPU, FPGA, and ASIC." Applied and Computational Engineering. 4. 99-104. 10.54254/2755-2721/4/20230358.
- Wang, Yue, Jiang, Ziyu, Chen, Xiaohan *et al.* "E2-Train: Training State-of-the-art CNNs with Over 80% Energy Savings." Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 5139–5151. NeurIPS 2019, arXiv:1910.13349
- Goel, Abhinav, Tung, Caleb, Lu, Yung-Hsiang *et al.* "A Survey of Methods for Low-Power Deep Learning and Computer Vision." arXiv:2003.11066 (2020).
- Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC²), Association for Computational Linguistics, 2019, pp. 36–40. arXiv:1910.01108, 2020.
- Zhang, Qin, Ziqi Liu, and Shirui Pan. "The Rise of Small Language Models." Shenzhen University and Griffith University, 2025.
- Sajith, Aryan, and Krishna Chaitanya Rao Kathala. "Is Training Data Quality or Quantity More Impactful to Small Language Model Performance?" Neural Computing for Advanced Applications, Springer Nature Singapore, 2025, pp. 75–87. Crossref, https://doi.org/10.1007/978-981-95-3739-6_6.
- Subramanyam, Anirudh, Yuxin Chen, and Robert L. Grossman. "Scaling Laws Revisited: Modeling the Role of Data Quality in Language Model Pretraining." arXiv, Sept. 2025.
- McMahan, Brendan, Moore, Eider, Ramage, Daniel *et al.* "Communication-Efficient Learning of Deep Networks from Decentralized Data." Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282. arXiv:1602.05629, 2016.
- Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong. "Federated Machine Learning: Concept and Applications." ACM Transactions on Intelligent Systems and Technology, vol.10, no.2, 2019, pp.1–19, doi:10.1145/3298981.
- Dayarathna, Miyuru & Wen, Yonggang & Fan, Rui. (2015). "Data Center Energy Consumption Modeling: A Survey." IEEE Communications Surveys & Tutorials. 1-1. 10.1109/COMST.2015.2481183.
- P. C. M. Krishnaram, "Green AI: Optimizing Energy Efficiency of Workloads for Sustainable Data Centers." 2025 IEEE Green Technologies Conference (GreenTech), Wichita, KS, USA, 2025, pp. 148-152, doi: 10.1109/GreenTech62170.2025.10977613.
- Pambudi, Nugroho Agung, Sarifudin, Alfian, Firdaus, Ridho Alfian *et al.* "The Immersion Cooling Technology: Current and Future Development in Energy Saving." Alexandria Engineering Journal, vol. 61, no. 12, 2022, pp. 9509–9527, doi:10.1016/j.aej.2022.02.059
- Junliang Liu, Xiaoya Zhang, Lique Shi. "Overview of Liquid Cooling Technology Development in Data Centers." Academic Journal of Science and Technology, October 2025 17(2):33-36, doi: 10.54097/p00b6t25
- S. Kochupurackal Rajan, B. Ramakrishnan, H. Alissa, W. Kim, C. Belady, and M. S. Bakir, "Integrated Silicon Microfluidic Cooling of a High-Power Overclocked CPU for Efficient Thermal Management." in IEEE Access, vol. 10, pp. 59259–59269, 2022, doi: 10.1109/ACCESS.2022.3179387
- Yuan, Xiaolei, Liang, Yumin, Hu, Xinyi *et al.* "Waste Heat Recoveries in Data Centers: A Review." Renewable and Sustainable Energy Reviews, vol. 188, 2023, article 113777, doi:10.1016/j.rser.2023.113777
- Periola, A.A., Alonge, A.A. & Ogudo, K.A. "Heat Wave Resilient Systems Architecture for Underwater Data Centers." Sci Rep 12, 17161 (2022). <https://doi.org/10.1038/s41598-022-21293-2>
- Periola, A.A., Kolawole, M.O. "Space-Based Data Centers: A Paradigm for Data Processing and Scientific Investigations." Wireless Pers Commun 107, 95–119 (2019)
- W. Abderrahim, O. Amin, and B. Shihada, "How to Leverage High Altitude Platforms in Green Computing?" IEEE Communications Magazine, vol. 61, no. 7, pp. 134-140, July 2023, doi: 10.1109/MCOM.001.2200732.
- H. Liu, A. Aljbri, J. Song, J. Jiang, and C. Hua, "Research advances on AI-powered thermal management for data centers." in Tsinghua Science and Technology, vol. 27, no. 2, pp. 303-314, April 2022

■ Author

Nikhil Bhadra is an aspiring entrepreneur with a strong interest in business and finance. Nikhil plans to study at a top US university to pursue a career in financial services. Nikhil is a swim instructor and founder of WaterSafety4Kids, a growing non-profit that sponsors swim lessons for children from underserved communities.