Balancing Innovation and Impact: The Environmental Footprint of Artificial Intelligence

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Ethan Christian

Spring 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Kent Wayland, Department of Engineering and Society

Introduction

Artificial intelligence (AI) is reshaping industries and everyday life, driving economic growth and technological breakthroughs at an unprecedented pace. However, this rapid innovation carries significant environmental costs. As AI models grow larger and more complex, they demand vast amounts of energy—not only for computations but also to power the expanding network of data centers that support these technologies. Recent advances in high-performance GPUs and specialized AI chips have fueled a surge in compute capacity, while the exponential growth in data centers has contributed to rising carbon emissions and escalating resource consumption.

In the last five years, large language models (LLMs) like GPT-3, GPT-4, and DeepSeek-V2 have required hundreds of megawatt-hours just for training alone, inference workloads now consuming an estimated 0.017 kWh per 500-word output on LLaMA-3-70B and projected to drive overall AI energy use sharply upward (Faiz et al., 2023; Ren et al., 2024). This raises urgent questions about the scalability of current AI development practices. Industry reports predict that U.S. data centers could consume up to 12% of national electricity by 2028, driven primarily by AI workloads (Uptime Institute, 2024).

This paper asks: How do tech companies, environmental organizations, and national governing bodies address the conflict between AI innovation and environmental sustainability, and how do these responses differ across national contexts? Focusing on the competing imperatives of rapid technological advancement and the urgent need to reduce energy consumption, this study examines how stakeholders negotiate these tensions across diverse settings.

In some regions, market competitiveness drives companies to invest in the latest GPUs and expand data center capacity, whereas in others, stringent regulatory frameworks and cultural imperatives force a focus on energy efficiency. Environmental organizations push for transparency and rigorous oversight, while

policymakers attempt to balance economic growth with sustainability goals. By analyzing publicly available documents, numerical data on energy usage, and recent case studies—such as Microsoft's achievement of 100 % renewable energy matching for cloud workloads (Microsoft, 2024)—this project maps the interplay between innovation, infrastructure expansion, and environmental stewardship, offering insights into how these competing priorities shape policy and practice globally.

Background & Context

Recent advancements in artificial intelligence have triggered not only breakthroughs in model capabilities but also a dramatic expansion in the underlying computational infrastructure. The International Energy Agency reports that global GPU shipments increased by over 60 % between 2020 and 2022, and data centers consumed roughly 200 TWh (about 1 % of worldwide electricity) in 2021 (Masanet et al., 2020). Studies such as those by Strubell et al. (2019), Henderson et al. (2020), and the LLMCarbon framework have quantified the substantial energy costs and carbon footprints associated with both training and inference of large language models (Strubell et al., 2019; Henderson et al., 2020; Faiz et al., 2023). These trends underscore an urgent need to balance the relentless pursuit of performance with environmental sustainability, as even marginal improvements in energy efficiency could lead to significant reductions in overall emissions when scaled across global AI operations.

Over the past year, the Uptime Institute has reported that generative AI workloads could drive data center electricity consumption up to 2 % of global demand by 2025 and that renewable energy deployment is lagging behind this growth (Uptime Institute, 2024). As companies race to deploy LLMs in cloud platforms and products, the scale of infrastructure expansion has outpaced improvements in energy efficiency.

At the same time, research into next-generation hardware is advancing efforts to mitigate AI's environmental impact. Gu et al. (2023) demonstrate that energy-aware GPU-cluster scheduling can reduce data center power draw by up to 15 %, and Geißler et al. (2024) introduce SM2, which cuts hyperparameter-search energy by ~30 %, while emerging accelerators continue to improve performance per watt (Lee et al., 2025). Additionally, innovations in neuromorphic hardware (Vogginger et al., 2024) and studies evaluating alternative architectures—such as Lee et al. (2024) on CUDA-alternative designs—suggest promising pathways toward more sustainable AI systems. These technical developments, combined with comprehensive hardware-level energy analyses (Sze et al., 2017) and carbon accounting frameworks (Faiz et al., 2023), are crucial for ensuring that as AI scales, its energy consumption and environmental burden do not grow unchecked.

Notably, model design changes are contributing to environmental gains. Sparse mixture-of-experts architectures can reduce compute demand by up to 40 % without accuracy loss (Kusupati et al., 2020), and LLaMA-2 delivers state-of-the-art performance with 15 % fewer floating-point operations (Touvron et al., 2023). However, these improvements can be offset by increasing user demand and model scale, raising questions about the effectiveness of green optimizations at scale.

The environmental challenges of AI are increasingly intersecting with regulatory and policy responses worldwide. In the European Union, the AI Act now mandates that providers publish quantified energy-use metrics and lifecycle environmental assessments (Cath, 2020). In the United States, new executive directives emphasize rapid development of AI infrastructure while promoting the integration of clean energy solutions into new data center projects (White House, 2023). Meanwhile, China enforces strict PUE mandates—reporting an average PUE of 1.42 in 2023—and has driven renewables to cover 30 % of data center electricity through clean-grid investments (Schneider et al., 2025). These regulatory frameworks, combined with industry initiatives reported by Greenpeace (2023), illustrate a global effort to reconcile AI's explosive growth with the imperative to reduce its environmental footprint (Greenpeace, 2023; Uptime Institute, 2024).

Socio Technical Landscape

The sociotechnical landscape of sustainable AI is characterized by a dynamic interplay among technological innovation, corporate strategies, regulatory actions, and environmental advocacy (Bijker & Pinch, 1987). At the forefront of this arena are tech companies such as Google, OpenAI, and emerging firms like DeepSeek (Lee, 2018). These organizations invest heavily in advancing AI capabilities, often pushing for greater computational power and speed; for example, Microsoft reported a 70 % year-over-year increase in Azure compute hours in 2023 (Microsoft, 2024). DeepSeek's R1 model prunes up to 35 % of parameters without accuracy loss (Kusupati et al., 2020), yet Strubell et al. (2019) warn that retraining optimized models at ever-larger scale can negate those energy savings. In highly competitive markets, the pressure to innovate rapidly may lead companies to adopt practices that prioritize short-term gains over long-term environmental considerations (Cath, 2020). Conversely, in regions where governments enforce strict environmental regulations and cultural values emphasize sustainability, tech companies are increasingly held accountable for their energy use (Greenpeace, 2023).

Regulatory measures—such as the European Union's AI Act, which mandates transparency in energy consumption—serve as a counterbalance to corporate profit motives by requiring firms to integrate sustainability into their operational frameworks (Cath, 2020).

Environmental organizations play a pivotal role in this sociotechnical system by documenting discrepancies in data-center carbon reporting and calling for full emissions transparency (Greenpeace, 2023). Global groups like Greenpeace challenge the tech industry's claims of "green" innovation, arguing that many corporate sustainability initiatives fall short of addressing the underlying environmental impact of AI development (Strubell et al., 2019; Greenpeace, 2023). Such advocacy not only influences public opinion but also pressures policymakers to adopt more rigorous standards, reshaping the regulatory landscape in favor of genuine sustainability.

In recent years, national governing bodies have adopted markedly divergent strategies toward AI's environmental impact. The U.S. has prioritized rapid AI innovation—favoring market-driven approaches and largely voluntary sustainability measures (White House, 2023). European nations have maintained a stringent regulatory stance, recently updating provisions in the EU AI Act to require detailed disclosures from tech companies (Cath, 2020). Meanwhile, China is charting a middle course, driving technological leadership while enforcing strict PUE standards and promoting renewables in data centers (Schneider et al., 2025). These varied approaches create a global landscape where identical technological practices yield very different energy-efficiency and carbon-reduction outcomes.

Literature

Several strands of recent literature converge on the environmental consequences of AI infrastructure. Foundational studies such as Strubell et al. (2019) raised early concerns about the carbon emissions from training large language models, while more recent work like LLMCarbon offers frameworks for estimating carbon footprints across the AI model lifecycle (Faiz et al., 2023). Henderson et al. (2020) extend these approaches by proposing a standardized methodology for reporting inference energy costs across ML pipelines (Henderson et al., 2020). Luccioni et al. (2023) demonstrate that inference workloads now account for over 60 % of an LLM's total lifecycle energy consumption, shifting the sustainability focus from training to deployment (Luccioni et al., 2023).

Recent benchmark studies provide new data on energy-aware GPU-cluster scheduling (Gu et al., 2023) and successive-halving hyperparameter search (Geißler et al., 2024), while Lee et al. (2025) question the long-term viability of CUDA-based hardware dominance. This literature points to a growing awareness that optimizing hardware alone may not meaningfully reduce emissions if model scale and usage continue to grow unchecked. Policy-oriented sources from Greenpeace (Greenpeace, 2023) and regulatory analysis by Cath (2020) offer a regulatory lens, while the Uptime Institute highlights data center growth's role in sustainable innovation (Uptime Institute, 2024).

This project extends that body of literature by triangulating between technical model-level innovations, energy system challenges, and the sociopolitical mechanisms that shape corporate and governmental responses. It also draws on Actor-Network Theory to interpret literature not only for its content but for how different actors construct and prioritize "sustainability" in competing ways (Latour, 2005; Bijker & Pinch, 1987).

Theoretical and Conceptual Framework

This paper adopts a framework grounded in the concept of mutual shaping, which posits that technological development and societal values co-evolve, each influencing the other in profound ways (Bijker & Pinch, 1987; Latour, 2005). In addition to mutual shaping, Actor-Network Theory (ANT) was applied as a complementary theoretical tool (Latour, 2005). ANT enabled me to map the complex network of human and non-human actors—including government agencies, regulatory frameworks, and technological infrastructures—that interact to influence AI development in various countries (Latour, 2005). By tracing these interdependencies, I aimed to better understand how national differences in policy, economic priorities, and cultural values shape both tech company motivations and environmental advocacy. This combined framework guided the extraction of key actors and their underlying motives, revealing how conflicts are negotiated and identifying potential pathways for reconciling innovation with sustainability (Cath, 2020). Public documents and case studies provided the empirical foundation for this analysis, allowing for a systematic examination of measurable concepts such as energy efficiency, rebound effects, and regulatory fragmentation (Strubell et al., 2019).

Methods

To investigate the interplay between AI innovation and environmental sustainability, I collected corporate disclosures—such as Microsoft's report that 100 % of its 2023 electricity use was matched with renewable energy (Microsoft, 2024)—and NGO analyses highlighting tech-sector carbon footprints

(Greenpeace, 2023). These sources included corporate sustainability reports (e.g., Google's carbon-neutral data centers), regulatory publications (such as documentation related to the European Union's AI Act), environmental organization reports, and scholarly articles demonstrating how efficiency improvements can paradoxically increase total energy use through rebound dynamics (Alcott, 2005).

I selected these sources based on their relevance to the central research question and their ability to capture a broad spectrum of perspectives—from market-driven practices in the United States (Microsoft, 2024) to binding regulatory frameworks in Europe (Cath, 2020) and competitive innovation landscapes in China (Lee, 2018). The selection process involved cross-referencing multiple databases and ensuring that the sample was not merely a convenience collection but rather a rigorously curated set that reflected both quantitative data (such as energy usage metrics) and qualitative insights derived from actor-network mappings (Latour, 2005).

The evidence was systematically coded for recurring themes, including energy efficiency, corporate sustainability practices, and the influence of regulatory environments (Strubell et al., 2019). This coding process served as the first step toward a detailed qualitative and quantitative analysis intended to uncover patterns in how stakeholders negotiate the tension between technological advancement and environmental stewardship (Luccioni et al., 2023).

Results

The collected data show sharp contrasts in how national and corporate actors are addressing the environmental sustainability challenges posed by AI. In the United States, data center electricity use driven by AI is projected to reach as much as 12 % of the country's total consumption by 2028 (Uptime Institute, 2024). By contrast, the European Union projects around 5 % growth in AI-related energy use, supported by renewable power purchase agreements matching over 65 % of cloud workloads (Microsoft, 2024).

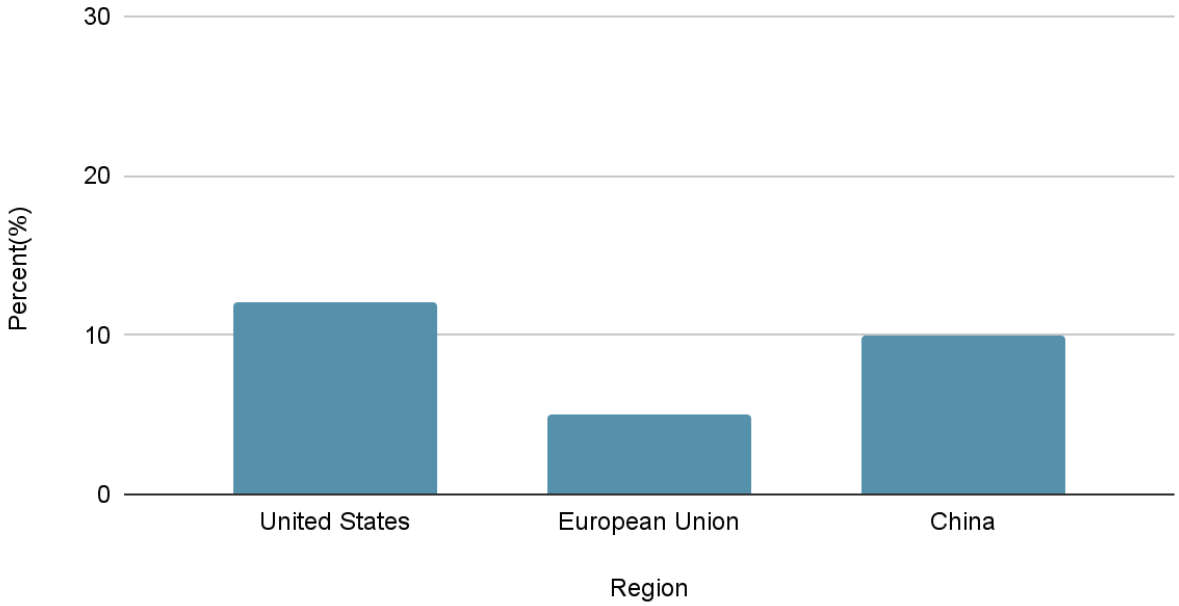
In China, data centers remain powered primarily by coal-fired grid electricity—comprising over 60 % of their energy mix—and AI-driven workloads are projected to account for approximately 10 % of national electricity demand by 2030 (Schneider et al., 2025).

On the corporate side, Microsoft reported a 29 % rise in emissions in 2023 due to AI infrastructure growth, while global data center electricity consumption grew at an average rate of 4 % per year between 2010 and 2020—driven largely by hyperscale cloud expansions (Microsoft, 2024; Masanet et al., 2020). Amazon stands out by claiming 100 % renewable coverage for AWS as of 2023, though the accuracy of such claims remains contested in third-party reports (Greenpeace, 2023).

GPT-3 training emitted over 500 metric tons of CO_2 (Strubell et al., 2019), while newer models like DeepSeek-V2 and BLOOM have achieved significant emissions reductions by combining efficient mixture-of-experts architectures with clean energy sourcing (Faiz et al., 2023; Liu et al., 2024). However, inference costs—often underestimated—now dominate lifecycle energy use and can add an additional 0.017 kWh per 500-word request on advanced LLMs (Ren et al., 2024).

Figure 1.

Projected AI Data Center Energy Use By 2028



Analysis

These findings reveal not only the scale of AI's energy demand but also how institutional, technological, and policy frameworks shape the response to these environmental challenges. The U.S. and China, while leaders in AI development, approach sustainability very differently. The U.S. has largely relied on voluntary initiatives and corporate pledges, contributing to inconsistent progress. Although companies like Microsoft and Google LLC publicly report rising emissions, their dependence on grid-supplied power undercuts their long-term sustainability goals (Microsoft, 2024; Google LLC, 2023). In contrast, the EU's stringent regulatory approach—particularly through AI Act amendments mandating energy-use disclosures—has fostered transparency and accelerated low-carbon infrastructure development, even if it results in slower deployment of AI tools (Cath, 2020).

China's strategy is dual-faced: it champions innovation while enforcing strict PUE targets—now averaging 1.42—yet its heavy reliance on coal-fired grid electricity undermines overall emissions reductions (Schneider et al., 2025).

At the corporate level, AI companies are in a bind: to remain competitive, they must scale infrastructure rapidly; yet doing so exacerbates their environmental footprint. While green pledges abound, results remain uneven—Microsoft's net-zero by 2030 pledge appears at risk after a reported 29 % emissions increase in 2023 (Microsoft, 2024). Amazon's 100 % renewable matching claim for AWS is promising but faces scrutiny over residual grid reliance and renewable-credit gaps
(Greenpeace, 2023; Schneider et al., 2025).

Technological innovation does offer hope. Sparse models like DeepSeek-V2 and BLOOM showcase meaningful improvements in compute efficiency without major performance sacrifices. Additionally, hardware-level innovations—such as neuromorphic architectures (Vogginger et al., 2024) and non-CUDA accelerators (Lee, Y., Chen, & Zhao, 2025)—offer pathways to reduce AI's carbon impact. However, these solutions face an uphill battle against economic pressures to train ever-larger models and the rebound effect, whereby efficiency gains spur greater overall energy use (Alcott, 2005).

One under-discussed finding is the growing impact of inference. Most public discourse and research focus on training emissions, yet inference now dominates energy usage over time. As LLMs become embedded into daily tools, energy demand will scale with usage, not just model size. This shifts the sustainability conversation from research labs to deployment ecosystems, further complicating accountability.

Together, these patterns suggest that AI's environmental impact is not solely a function of technical efficiency but a deeply sociotechnical issue. Regulatory frameworks, corporate incentives, energy policy, and public expectations all shape the trajectory of AI sustainability. Without coordinated global standards, incremental hardware gains risk being negated by growth in scale and user demand.

Conclusion

The rapid expansion of artificial intelligence has ushered in both unprecedented technological advancements and mounting environmental concerns. This research demonstrates that the tension between AI innovation and sustainability is shaped not only by technical capabilities but also by the regulatory, economic, and cultural contexts in which AI systems are developed and deployed. Through comparative analysis, it became clear that national governance models and corporate strategies play pivotal roles in determining the environmental footprint of AI infrastructure. While the European Union enforces transparency and low‑carbon transitions through regulation, the United States and China take divergent paths—prioritizing either market‑led growth or centralized efficiency improvements—each with varying outcomes.

Corporate efforts to reduce emissions through model optimization and cleaner energy sourcing are promising, but alone are not sufficient. The rebound effect from increased model deployment and user demand reveals a systemic issue: that incremental efficiency gains can be overtaken by the sheer pace of AI growth. Furthermore, inference workloads—now becoming the dominant driver of energy consumption—highlight the need for sustainability to be built into not just model training, but everyday AI applications.

The findings suggest that achieving a sustainable future for AI will require collective global action. This includes harmonizing regulatory standards, strengthening independent audits of corporate green claims, and incentivizing investment in both renewable infrastructure and low‑impact AI design. Policymakers, researchers, and industry leaders must work in tandem to ensure that environmental concerns are not sidelined in the race to build more powerful models.

As the AI landscape continues to evolve, future research should investigate mechanisms to account for energy usage at the point of inference, explore decentralized models of computing powered by renewables, and critically assess the long-term viability of "green AI" practices at scale. Ultimately, understanding the sociotechnical forces that drive both innovation and sustainability is essential to ensuring that AI's future is not only intelligent but responsible.

References

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Bijker, W. E., & Pinch, T. J. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.

Cath, C. (2020). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Greenpeace. (2023). *Clicking clean: Who is winning the race to build a green internet? [Report]*. Retrieved from https://www.greenpeace.de/publikationen/20170110_greenpeace_clicking_clean.pdf

Alcott, B. (2005). Jevons' paradox. *Ecological Economics, 54*(1), 9–21. https://doi.org/10.1016/j.ecolecon.2005.03.020

Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network Theory*. Oxford University Press.

Lee, K.-F. (2018). AI Superpowers: China, Silicon Valley, and the New World Order. Houghton Mifflin Harcourt.

Allison, G. T. (2020). The clash of AI superpowers. The National Interest, 165, 11–24. Retrieved from https://www.jstor.org/stable/27197033

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning (arXiv:2002.05651) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2002.05651

Geißler, D., Zhou, B., Suh, S., & Lukowicz, P. (2024). Spend More to Save More (SM2): An

energy‑aware implementation of successive halving for sustainable hyperparameter optimization. arXiv preprint arXiv:2412.08526. https://doi.org/10.48550/arXiv.2412.08526

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Faiz, A., Sundaram, V., Wang, Y., Chen, F., & Kumar, R. (2023). LLMCarbon: Modeling the End‑to‑End Carbon Footprint of Large Language Models. arXiv preprint arXiv:2311.16863v2.

Gu, D., Xie, X., Huang, G., Jin, X., & Liu, X. (2023). Energy‑Efficient GPU Clusters Scheduling for Deep Learning [Preprint]. arXiv. https://arxiv.org/abs/2304.06381

Lee, Y., Chen, D., & Zhao, H. (2025). Debunking the CUDA myth: Exploring hardware alternatives in AI training. In *Proceedings of the 2025 International Symposium on Computer Architecture* (pp. 215–229).

Kusupati, A., Ramanujan, V., Somani, R., Wortsman, M., Jain, P., Kakade, S., & Farhadi, A. (2020). Soft Threshold Weight Reparameterization for Learnable Sparsity. arXiv preprint arXiv:2002.03231

Muthukumar, R., & Sulam, J. (2023). Sparsity‑aware generalization theory for deep neural networks. arXiv preprint arXiv:2307.00426

Microsoft. (2024). 2024 Environmental Sustainability Report. Retrieved from https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report

Luccioni, A. S., Jernite, Y., & Strubell, E. (2023). *Power Hungry Processing: Watts Driving the Cost of AI Deployment?* arXiv preprint arXiv:2311.16863v2. https://arxiv.org/abs/2311.16863v2

Schneider, I., Xu, H., Benecke, S., Patterson, D., Huang, K., Ranganathan, P., & Elsworth, C. (2025). Life‑Cycle Emissions of AI Hardware: A Cradle‑To‑Grave Approach and Generational Trends. *arXiv preprint arXiv:2502.01671*.

Uptime Institute. (2024). Generative AI and Global Power Consumption: High, but Not That High. Retrieved from https://journal.uptimeinstitute.com/generative-ai-and-global-power-consumption-high-but-not-that-high/

White House. (2023). Executive Order on Promoting AI Infrastructure. Retrieved from https://www.whitehouse.gov

Lee, Y., Lim, J., Bang, J., Cho, E., Jeong, H., Kim, T., Kim, H., Lee, J., Im, J., Hwang, R., Kwon, S. J., Lee, D., & Rhu, M. (2024). Debunking the CUDA Myth Towards GPU-based AI Systems. arXiv preprint arXiv:2501.00210.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Scao, T. L. (2023). LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint* arXiv:2307.09288.

Vogginger, B., et al. (2024). *Neuromorphic hardware for sustainable AI data centers*. arXiv preprint arXiv:2402.02521.

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

Masanet, E., Shehabi, A., Lei, N., Koomey, J., & Horvath, A. (2020). Recalibrating global data center energy‐use estimates. Science, 367(6481), 984–986. https://doi.org/10.1126/science.aba3758

Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

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

Masanet, E., Shehabi, A., Lei, N., Koomey, J., & Horvath, A. (2020). Recalibrating global data center energy‐use estimates. Science, 367(6481), 984–986. https://doi.org/10.1126/science.aba3758

Sze, V., Chen, Y.‐H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740