

Navigating the AI Energy Challenge: A Sociotechnical Framework and Strategic Solutions for Sustainable Artificial Intelligence

Xiaoyang Liu^{1*}

¹ School of Business, Macau University of Science and Technology, Macau, China

Abstract. Artificial intelligence is at the intersection of innovation and escalating energy demands. This paper addresses the AI energy paradox through an integrated sociotechnical framework that combines technological architectures, organizational practices, and adaptive governance. Comprehensive case analyses reveal critical leverage points where targeted interventions boost performance while significantly reducing energy consumption. Our findings challenge conventional views of inherent efficiency–performance trade-offs, showing that these limitations largely stem from outdated design choices. We propose a balanced strategy: deploy mid-scale models for routine, high-efficiency tasks (e.g., dataset processing and rapid document summarization) and reserve high-capacity models with advanced reasoning for complex challenges. By aligning optimized hardware architectures with strategic policy measures, our approach offers considerable economic, operational, and environmental benefits. Furthermore, our analysis highlights an urgent need for innovative, energy-conscious AI development strategies. This roadmap empowers researchers, practitioners, and policymakers to harness AI's transformative potential while ensuring ethical and sustainable development for current and future generations.

1 Introduction

Artificial intelligence stands at a critical juncture, simultaneously functioning as both a significant energy consumer and a potential energy optimizer. Large language models like GPT-4 consume substantial computational resources, exponentially increasing training energy requirements as models grow [1]. Industry analyses indicate that AI training energy consumption has grown approximately 300,000 times between 2012 and 2024, with the most significant models now requiring energy equivalent to the annual consumption of thousands of households. This trend is expected to persist as the development of larger and more capable models accelerates.

This phenomenon exemplifies the "AI energy paradox"—where energy-intensive AI systems are developed partly to create tools that optimize energy consumption across other domains. For example, AI systems consuming megawatt-hours during training may

* Author: 1230005631@student.must.edu.mo

eventually optimize HVAC systems, power grids, or transportation networks, potentially saving gigawatt-hours during their operational lifetimes [2]. This complex relationship between consumption and conservation requires nuanced consideration that defies simplistic analysis. The paradox necessitates a shift from static energy consumption measures to a dynamic "value return" analysis framework that balances direct energy costs against broader potential energy savings across systems.

Considering these trends, it becomes clear that conventional design choices often impose the seemingly inherent trade-offs between efficiency and performance rather than dictated by physical limits. By re-envisioning system architectures, recent empirical studies demonstrate that paradigm shifts can simultaneously enhance performance and markedly reduce energy consumption [3]. Consequently, integrating innovative technical and sociotechnical approaches drives operational improvements and lays the groundwork for a sustainable AI ecosystem that delivers economic, operational, and environmental benefits.

Building on this perspective, our proposed framework emphasizes coordinated interventions across technical architecture, user and organizational behaviours, and adaptive governance. In doing so, it contributes to performance improvements smoothly along every stage of development while preserving the overall integrity of the system's design. This holistic approach ensures that each component—from hardware optimizations to behavioural nudges-reinforces the others in delivering significant energy savings.

Ultimately, the framework offers a clear path forward by challenging the notion of fixed efficiency-performance trade-offs. It establishes that, with thoughtful integration, AI systems can be competent and sustainably efficient, meeting immediate operational demands while safeguarding resources for future generations.

2 Challenges

2.1 Technical dimensions and zero-sum paradigm

The relationship between AI performance and energy efficiency represents a complex optimization landscape with multiple local optima. Contrary to simplistic linear trade-off assumptions, this relationship manifests as a Pareto frontier with opportunities for simultaneous improvement in both dimensions—the current emphasis on parameter count as the primary scaling metric has created artificial performance-efficiency trade-off.

Pareto optimization theory and game theory provide complementary frameworks for understanding this challenge. Pareto theory delineates the technical possibility boundary—identifying optimal trade-off points between performance and energy efficiency within current technological paradigms. This technical feasibility boundary is not a simple linear trade-off but a complex landscape with multiple locally optimal solutions. At the Pareto frontier, improving one dimension typically requires sacrificing the other [4]. From a game theory perspective, the AI industry faces a classic "prisoner's dilemma." While collective energy conservation benefits the industry, individual companies are incentivized to pursue energy-intensive models for competitive advantage [5-6]. This dynamic creates an "arms race" effect that continually drives energy consumption. The competition between leading AI labs has resulted in exponential increases in computational resources devoted to frontier model development, with computational requirements for state-of-the-art models doubling approximately every 6-10 months since 2018. This competitive dynamic creates substantial market pressure against energy optimization unless specifically incentivized through policy or market mechanisms.

The relationship between AI performance and energy efficiency extends beyond algorithm design to infrastructure deployment decisions. Strategic data centre locations near

renewable energy sources can dramatically alter the efficiency-performance equation, minimizing transmission losses while decreasing operational carbon footprints. For example, through natural cooling, Microsoft's underwater data centre initiative has realized modest efficiency improvements—8–11%. Similarly, Google's strategic deployment of AI computing clusters in Nordic regions effectively harnesses renewable energy and natural cooling, substantially lowering carbon intensity by roughly 60% compared to typical data centre operations.

Recent technical breakthroughs demonstrate that performance-efficiency trade-off can be transcended. The CoMP framework offers a non-intrusive approach that intelligently allocates precision at the operator level based on convergence awareness, optimizing the balance between accuracy and computational efficiency [7]. Similarly, sparse activation models like Mixture of Experts selectively activate only relevant network parts for each input, dramatically reducing computation during inference [8]. Knowledge distillation techniques effectively compress model knowledge into more compact architectures. Recent implementations have demonstrated that these methods can preserve approximately 90–95% of the original model's capabilities while reducing its size by roughly 65–80%. In parallel, quantization techniques—such as 8-bit quantization—can significantly lower memory usage and computational requirements, typically cutting inference costs by around 50–60% with minimal impact on performance on overall performance [9].

2.2 Sociotechnical factors in AI energy consumption

While technical optimization remains crucial, significant breakthroughs in AI energy consumption require expanding our perspective to broader sociotechnical systems. User behaviour, organizational decisions, and cultural factors collectively shape AI's actual energy footprint. Research indicates that much of AI energy consumption stems from user interaction patterns and deployment decisions rather than fundamental technical limitations [10]. This perspective focuses on the holistic energy efficiency of "human-machine systems," explaining why technically efficient AI systems might fail to achieve expected energy savings in real-world deployments. For example, a technically optimized AI recommendation system that perpetually engages users may consume more energy through increased server utilization and client-side processing than a less optimized system with more thoughtful engagement patterns. Similarly, default settings that continuously run resource-intensive background processes may negate algorithmic efficiency gains if not aligned with actual user needs.

Beyond technical innovation, sociotechnical factors play a pivotal role in shaping the actual energy footprint of AI systems. User behaviours, interface designs, and organizational cultures all contribute substantially to overall energy consumption [11]. In practice, optimizing interface design and adjusting internal structures have helped many applications achieve energy savings while either maintaining or improving user experience. Additionally, organizations that have restructured their processes to prioritize energy efficiency report noticeable improvements without compromising product performance. Moreover, medium-scale models have proven highly effective for routine content creation tasks—such as writing blogs, managing social media, processing datasets, generating quick document summaries, etc.—delivering sufficient performance at a much lower energy cost. In contrast, large-scale models, which offer advanced reasoning and deeper contextual understanding, excel in handling complex tasks that require creative problem-solving and specialized expertise [12]. Therefore, deploying them selectively and complementing their use with specialized optimization techniques and renewable energy strategies is advisable for functions requiring high-capacity models.

Collectively, these findings underscore that when sociotechnical strategies are effectively integrated with technical optimizations, the apparent trade-offs vanish. The benefits achieved across energy, performance, and user satisfaction confirm that system-level redesigns can yield synergistic advantages.

2.3 Governance frameworks and ethical considerations

Turning to governance, adaptive policy frameworks are critical in steering AI development toward sustainable practices. For example, carbon pricing mechanisms, dynamic regulations, and unified sustainability indices create tangible incentives that align technological progress with environmental responsibility. Evidence from model comparison studies reinforces the idea that efficiency improvements need not come at the cost of performance, thereby supporting a balanced deployment strategy that accounts for immediate energy savings and long-term ethical imperatives.

In this context, intergenerational justice introduces an additional ethical dimension to AI energy decisions. The current trajectory of AI development risks disadvantaging future generations by depleting shared environmental resources for short-term gains [13]. This challenge compels us to critically balance our present capabilities with the imperative to preserve resources for the future. Moreover, the potential “sustainability divide” faced by developing countries—with their limited AI infrastructure—underscores the necessity for international cooperation, technology transfer initiatives, and targeted funding mechanisms to ensure an inclusive and equitable transition toward sustainable AI.

Addressing AI energy challenges also demands that policy interventions be tailored to the unique lifecycles and deployment characteristics. Traditional policy tools often fall short in this regard, necessitating more flexible, adaptive approaches. For instance, carbon pricing can influence AI development in multiple ways: imposing carbon taxes that raise the cost of high-carbon energy, adjusting electricity prices to increase operational costs, and offering location incentives that encourage data center migration to regions with low-carbon energy sources [14]. Research indicates that carbon pricing within the range of \$40–60 per ton provides robust economic incentives to drive AI development toward more energy-efficient architectures—while even higher prices may further accelerate research into energy-efficient computing paradigms.

Further reinforcing these policy measures is “the Principle of Sufficiency,” which offers an ethical baseline by questioning the unlimited growth logic often inherent in technological development [15]. Applied to AI, this principle advocates carefully evaluating model scale against actual task value. For example, medium-scale models have demonstrated strong efficiency in routine content creation tasks such as dataset processing and rapid document summarization. These models are competent in handling high-efficiency tasks while substantially reducing energy consumption. In contrast, large-scale models, with their superior reasoning capabilities and nuanced contextual understanding, are better suited for complex problems demanding high precision. This delineation suggests that, rather than unthinkingly scaling up, AI deployment should prioritize sufficiency—choosing the appropriate model scale for the specific task. Selectively deploying high-capacity models only where necessary and complementing their use with targeted optimization strategies can lead to more sustainable AI practices.

Yet beyond the technical and economic considerations, the Principle of Sufficiency raises a more profound ethical challenge: what does it mean for something to be sufficient? At its core, sufficiency is not merely about meeting a threshold of adequacy but understanding and embracing responsible restraint. The difficulty lies in defining this balance—how much is enough, and who decides? Are we optimizing for energy efficiency, human benefit, or long-term planetary sustainability? The challenge is to regulate AI development responsibly and

cultivate a mindset that resists excess for excess's sake. Thus, sufficiency demands more than a rigid set of constraints; it requires a harmonious interplay between technological advancement and ethical foresight. In navigating this challenge, we must move beyond static definitions and continuously reflect on what it means to develop AI systems that serve both present and future needs without overstepping the limits of sustainability.

3 Solutions

3.1 Paradigm-shifting technical architectures

Addressing the technical dimensions of the AI energy challenge requires moving beyond incremental improvements to implement paradigm-shifting architectures that fundamentally alter the energy-performance relationship. The following approaches offer actionable pathways for both immediate implementation and long-term transformation.

Neuromorphic computing represents a promising pathway for AI energy optimization. By mimicking the brain's architecture with spike-based processing and co-located memory and computation, neuromorphic systems demonstrate energy efficiency improvements of 1,000x over conventional architectures for specific tasks [16]. Intel's Loihi 2 and IBM's TrueNorth chips prove that spike-based processing with co-located memory and computation can be implemented at scale [17]. Quantitative benchmarks indicate that for pattern recognition tasks, these architectures achieve comparable accuracy to conventional deep learning while consuming less than 0.1% of the energy.

Though still in its early developmental stages, quantum computing promises exponential speedups for specific AI workloads, potentially reducing energy requirements dramatically for certain classes of problems [18]. Recent demonstrations by IBM, Google, and other leading research organizations have illustrated a clear quantum advantage for specific computational tasks. Ongoing algorithmic improvements further suggest that quantum techniques may soon apply to machine learning optimization challenges within a relatively short timeframe.

Additionally, biologically inspired systems offer additional efficiency optimization pathways. The brain's remarkable energy efficiency stems partly from its adaptive energy allocation and tolerance for imprecision [19]. AI systems could implement similar principles through dynamic precision adjustment, selectively applying high-precision computation only when necessary. Research has shown that many models can operate with 4-bit or even binary weights with minimal accuracy loss after appropriate retraining [20, 21].

To accelerate industry-wide adoption, organizations should implement standardized energy efficiency benchmarks alongside performance metrics, develop open-source reference implementations for energy-efficient architectures, create cross-industry working groups focused on efficiency standard implementation, and establish clear migration pathways from current architectures to more efficient alternatives.

3.2 Human-machine system optimization

Addressing the sociotechnical dimensions requires designing systems that optimize human-machine interaction rather than focusing on technical efficiency.

Organizations can achieve significant energy efficiency gains by adopting an integrated strategy that addresses both system design and organizational culture, thereby optimizing the entire human-machine ecosystem. At the system level, practical design solutions can significantly enhance overall efficiency. For example, integrating energy monitoring tools into system dashboards offers operators real-time, actionable insights, while adaptive

resource management techniques dynamically adjust computing power according to actual workload demands. In mature AI platforms, a more realistic approach emphasizes backend optimizations and adaptive deployment strategies rather than rapid, disruptive changes to user interfaces. Rather than incorporating overt energy metrics or rigid energy-saving defaults that might alter the user experience, organizations can deploy strategies that automatically scale down processing during periods of low activity and minimize unnecessary background operations. Furthermore, a targeted deployment strategy may involve employing mid-capacity models for routine tasks—such as dataset processing and rapid document summarization—while reserving high-capacity models exclusively for complex challenges requiring advanced reasoning and nuanced contextual understanding. This selective allocation of computational resources effectively addresses deployment inefficiencies that contribute significantly to overall energy consumption. At the organizational level, companies can transform their development culture by realigning incentive structures to balance efficiency with performance. This might include incorporating energy efficiency targets into engineering performance evaluations, establishing dedicated phases in the development cycle focused on energy conservation, forming cross-functional “Green Teams” that unite machine learning engineers with sustainability experts, and implementing internal carbon pricing mechanisms. Such initiatives have led to considerable improvements in overall efficiency within a relatively short period, all without compromising product performance.

Additionally, further optimizing interface design—through methods like progressive loading, prominently featuring eco-mode options, redesigning notification systems, and refining data retrieval processes—can result in meaningful reductions in energy consumption without sacrificing functionality. Deploying multi-tiered AI systems that dynamically allocate computational resources based on actual demand also plays a crucial role in lowering overall energy requirements. Together, these technical and cultural measures form a practical and sustainable pathway for the future of AI development.

In practice, organizations can achieve significant energy savings while maintaining optimal system performance and user satisfaction by integrating energy-aware algorithms with user-centric strategies—emphasizing seamless backend improvements, transparent efficiency modes, and user-controlled options.

3.3 Adaptive governance for AI sustainability

Effective AI governance frameworks must balance innovation with environmental responsibility through practical implementation mechanisms. Specific governance approaches can address sustainability challenges while maintaining technological advancement.

A Unified AI Sustainability Index represents a foundational governance tool, combining energy efficiency metrics, carbon emissions, hardware utilization, and model reuse statistics. This standardized measurement framework enables consistent assessment across diverse applications and facilitates evidence-based policy development. Dynamic carbon pricing offers a sophisticated approach tailored to the application type and development stage. Initial training phases could receive temporary allowances if the resulting model demonstrably reduces system-wide energy consumption. This mechanism could redirect development resources toward more efficient architectures, creating substantial sustainability improvements without compromising innovation. For smaller organizations facing resource constraints, solutions include leveraging cloud services optimized for energy efficiency, implementing open-source monitoring tools, and participating in industry collaborations. These efficiency improvements typically achieve ROI within 1-2 years, creating clear business incentives beyond compliance requirements. Public-private partnerships provide

additional opportunities by aligning market incentives with sustainability objectives. Government funding can prioritize fundamental research, while industry partners focus on commercialization and scale. As noted, this multifaceted approach addresses implementation challenges across organizational contexts, particularly for mature organizations facing entrenched systems and cultural resistance, while providing concrete pathways toward sustainable AI development [22].

4 Conclusion

Achieving sustainable AI energy consumption requires an integrated approach that simultaneously addresses technical, sociotechnical, and governance dimensions. Rather than viewing these as separate domains, organizations should develop coordinated strategies that leverage synergies between them. Our research demonstrates that perceived efficiency-performance trade-offs often represent artificial constraints that can be transcended through systemic innovation rather than fundamental physical limitations.

First, technical innovations must extend beyond incremental improvements to paradigm-shifting architectures that fundamentally alter the energy-performance relationship. This requires developing standardized reference architectures, creating clear migration pathways from current systems to efficient alternatives, and establishing industry-wide sustainability benchmarks. The neuromorphic, quantum, and biologically inspired approaches discussed provide concrete pathways toward order-of-magnitude efficiency improvements. Second, sociotechnical interventions must address the human dimensions of AI energy use through integrated energy considerations in development workflows, user interfaces that naturally guide efficient usage patterns, and organizational cultures that value sustainability. The interface design principles, organizational structures, and incentive mechanisms outlined provide actionable implementation pathways that can deliver significant efficiency improvements without compromising functionality. Third, policy frameworks should balance innovation with environmental protection through collaborative public-private partnerships and incentive structures that reward system-wide optimization. The dynamic carbon pricing mechanisms, self-governance frameworks, and research partnerships discussed create complementary approaches that can collectively redirect AI development toward more sustainable trajectories while maintaining innovation capacity. Overall, organizations embracing this integrated approach will not only fulfil ethical imperatives but also gain competitive advantages through reduced operational costs, enhanced reputation, and increased resilience against energy price volatility.

While our framework introduces a novel approach to sustainable AI, we recognize our limited exploration of emerging hardware technologies and potential rebound effects where efficiency gains might paradoxically increase usage. Future research will address comprehensive lifecycle analysis, ethical access considerations, and industry-specific adaptations across diverse computational environments. AI's path lies not in choosing between performance and efficiency but in reimagining both through thoughtfully designed systems that deliver maximal value with minimal resource consumption—benefiting current and future generations. By moving beyond artificial constraints toward a more sustainable AI ecosystem, we can ensure that artificial intelligence fulfils its promise as a transformative technology that enhances human capability while respecting planetary boundaries.

References

1. V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, et al., A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 128096 (2024)

2. S. Moveh, E. A. Merchán-Cruz, A. O. Ibrahim, et al., Thermodynamic Optimization of Building HVAC Systems Through Dynamic Modeling and Advanced Machine Learning. *Sustainability*, **17**(5), 1955 (2025)
3. X. Bi, D. Chen, G. Chen, et al., Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954* (2024)
4. K. Deb, A. Pratap, S. Agarwal, et al., A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans Evol Comput*, **6**(2), 182-197 (2002)
5. A. Younesi, M. Ansari, A. Ejlali, et al., GAP: Game Theory-Based Approach for Reliability and Power Management in Emerging Fog Computing. *arXiv preprint arXiv:2412.11310* (2024)
6. T. Hazra, K. Anjaria, Applications of game theory in deep learning: a survey. *Multimed Tools Appl*, **81**(6), 8963-8994 (2022)
7. W. Dai, Z. Jia, Y. Bai, et al., Convergence-aware operator-wise mixed-precision training. *CCF Trans High Perform Comput*, **7**(1), 43-57 (2025)
8. D. Dai, C. Deng, C. Zhao, et al., Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066* (2024)
9. P. Liang, J. Chen, Y. Wu, et al., Data-free knowledge distillation with feature synthesis and spatial consistency for image analysis. *Sci Rep*, **14**(1), 27557 (2024)
10. W. König, S. Löbbe, S. Büttner, et al., Establishing energy efficiency-drivers for energy efficiency in German manufacturing small and medium-sized enterprises. *Energies*, **13**(19), 5144 (2020)
11. M. A. Craciun, Behavioral Economics and Technology Innovation: Using Choice Architecture to Build and Scale Products. In *Proc Int Conf Bus Excell*, **17**(1), 904-913 (2023).
12. P. Tominc, D. Oreški, V. Čančer, et al., Statistically significant differences in AI support levels for project management between SMEs and large enterprises. *AI*, **5**(1), 136-157 (2024)
13. A. Halsband, Sustainable AI and intergenerational justice. *Sustainability*, **14**(7), 3922 (2022)
14. X. Yang, Z. Zhang, H. Chen, et al., Assessing the carbon emission driven by the consumption of carbohydrate-rich foods: The case of China. *Sustainability*, **11**(7), 1875 (2019)
15. D. S. Watson, L. Gultchin, A. Taly, et al., Local explanations via necessity and sufficiency: Unifying theory and practice. *Uncertainty in Artificial Intelligence*, 1382-1392 (2021)
16. T. Luo, W. F. Wong, R. S. M. Goh, et al., Achieving green ai with energy-efficient deep learning using neuromorphic computing. *Commun ACM*, **66**(7), 52-57 (2023)
17. A. Pal, Z. Chai, J. Jiang, et al., An ultra energy-efficient hardware platform for neuromorphic computing enabled by 2D-TMD tunnel-FETs. *Nat Commun*, **15**(1), 3392 (2024)
18. A. Ajagekar, F. You, Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. *Renew Sustain Energy Rev*, **165**, 112493 (2022)
19. J. I. Okonkwo, M. S. Abdelfattah, P. Mirtaheri, et al., Energy-aware bio-inspired spiking reinforcement learning system architecture for real-time autonomous edge applications. *Front in Neurosci*, **18**, 1431222 (2024)

20. M. Vandersteegen, K. Van Beeck, T. Goedemé, Integer-only cnns with 4-bit weights and bit-shift quantization scales at full-precision accuracy. *Electronics*, **10**(22), 2823 (2021)
21. L. Liu, Z. Zheng, C. Wang, et al., Binary Neural Networks for Large Language Model: A Survey. *arXiv preprint arXiv:2502.19008* (2025)
22. I. Kulkov, J. Kulkova, R. Rohrbeck, et al., Artificial intelligence - driven sustainable development: Examining organizational, technical, and processing approaches to achieving global goals. *Sustain Dev*, **32**(3), 2253-2267 (2024)