Frontiers in Economics and Management Volume 7 Issue 3, 2026
ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

Study on the Full Life-Cycle Environmental Impact and Systematic Optimization of Large Models

Yajie Zhu

Guangzhou College of Commerce, Guangzhou 511363, China

Abstract

While large-scale artificial intelligence models drive technological innovation, their entire lifecycle is accompanied by significant energy consumption and carbon emissions, posing a real risk of sliding from the vision of "Green AI" toward the reality of a "high-carbon technology". This paper constructs a three-tier analytical framework encompassing the training, inference, and hardware infrastructure stages to systematically assess the environmental impact of large models. Furthermore, it analyzes the underlying causes of the high-carbon dilemma from three dimensions: technological dependency, economic incentives, and governance gaps. The research indicates that a single technological optimization is insufficient for achieving sustainable transformation. It is necessary to collaboratively advance algorithm lightweighting, system scheduling optimization, industry standard establishment, and an ethical paradigm shift to construct a multi-level, systematic development pathway for Green AI. Only through the co-evolution of technology, policy, and culture can artificial intelligence be genuinely guided to become an enabler for addressing environmental challenges.

Keywords

Green AI; Carbon Footprint; Environmental Impact Assessment; Optimization Pathways; AI Ethics.

1. Introduction

1.1. Research Background and Problem Statement

Currently, large-scale artificial intelligence models represented by the GPT series, Gemini, LLaMA, etc., are propelling various sectors of society into a new stage of intelligence. These models demonstrate exceptional capabilities in areas like natural language processing and multimodal understanding, significantly enhancing production efficiency and service experience. Their development follows the "scaling law", where model performance improves continuously with the exponential growth of parameter scale, data volume, and computational resources. Driven by this principle, global academia and industry are engaged in intense competition centered on model scale, pursuing a development path of "bigger is better".

However, behind this competition lies an increasingly severe cost in energy and environmental terms. Training large models requires immense computational power, with energy consumption reaching millions of kilowatt-hours and carbon emissions amounting to hundreds of tons, equivalent to the total annual emissions of hundreds of households. Furthermore, continuously operating inference services, rapid hardware iteration, and the expansion of data centers collectively exacerbate the systemic carbon footprint. While artificial intelligence is positioned as an "enabler" for climate solutions, it is itself gradually evolving into a significant high-carbon emitting sector[6].

In response to this challenge, the concept of "Green AI" has emerged, advocating for energy efficiency and carbon footprint to be placed alongside performance as core evaluation metrics, thereby steering AI development towards sustainability[3]. However, current industrial

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

practices still largely prioritize optimal performance, often overlooking computational costs and environmental impacts. This has led to a stark contrast between the "green vision" and the "high-carbon reality". A significant disconnect exists between academic calls for ethics and industrial competitive practices, with environmental costs largely treated as externalities and not yet integrated into the core decision-making processes of AI R&D and deployment.

Within this context, this study focuses on exploring the following three core questions:

What is the extent of the environmental impact of large models throughout their entire lifecycle? What are the limitations of existing assessment frameworks?

What are the deep-seated causes of the high-carbon attributes of large models? Do they involve the combined effect of multidimensional systems including technology, economy, and governance?

How can an integrated pathway spanning technology, systems, policy, and ethics be constructed to drive the paradigm shift of AI from a "high-carbon technology" to "Green AI"?

1.2. Research Significance

This research holds significant theoretical and practical value:

In terms of theoretical significance, firstly, this study integrates environmental sustainability into the core dimensions of AI ethics, expanding AI ethics research from fairness, transparency, and privacy protection to include environmental impact, thereby enriching its theoretical scope. Secondly, it constructs an environmental impact assessment framework covering the entire lifecycle-training, inference, and hardware-overcoming the current research limitation that predominantly focuses on the training stage, and provides a theoretical tool for systematically evaluating the carbon footprint of large models. Finally, through a critical analysis of the technology development paradigm dominated by the "scaling law", it encourages the academic community to reflect on the feasibility of continuously expanding technological paths under resource constraints, providing a basis for constructing a theory of sustainable AI.

In terms of practical significance, it can serve as a reference for multiple stakeholders: For AI developers and enterprises, it provides specific emission reduction strategies from algorithm optimization to system deployment, helping them balance performance competition with environmental responsibility and reduce long-term compliance and operational risks. For policymakers and regulatory agencies, it offers a basis for designing carbon accounting methods, energy efficiency standards, and green incentive policies, promoting the formation of industry norms. For investors and the public, it reveals the environmental risks and transformation opportunities within the AI industry, guides capital towards greener innovation, and enhances societal awareness and supervisory oversight of the hidden environmental costs of digital technology.

1.3. Literature Review

Relevant research can be primarily categorized into the following two threads, from which current research gaps and shortcomings can be identified:

1.3.1. Early Research on AI Energy Consumption and Carbon Emissions

Research in this field has evolved from case studies to trend warnings. Early research (Strubell et al., 2019) quantified the training energy consumption of specific natural language processing models, first revealing the significant potential carbon footprint of AI models and drawing academic attention to the environmental costs of computation. Subsequent research further tracked the exponential growth in energy consumption driven by increasing model scale, noting that its growth rate far exceeds the compensatory capacity of hardware efficiency improvements (Lacoste et al., 2019; Schwartz et al., 2020). These findings laid the groundwork for understanding the severity of the issue. However, most focused on isolated analysis of the training stage, and methodologies were not yet unified, making systematic comparison difficult.

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

1.3.2. The Research Thread of "Green AI"

In response to the above problems, "Green AI" research has primarily unfolded along the path of improving technical efficiency, which can be summarized into three aspects: First, efficient algorithm and model architecture innovation, such as developing sparse architectures like mixture of experts and improving attention mechanisms to reduce computational complexity; second, model compression and simplification techniques, including methods like knowledge distillation, quantization, and pruning, aimed at reducing model size and inference overhead; third, carbon-aware computing and system optimization, researching how to dynamically schedule computing tasks to cleaner times and regions based on the spatiotemporal variation of grid carbon intensity, and optimizing data center energy use efficiency. This thread has provided numerous technical tools for reducing the environmental cost per unit of computation.

1.3.3. Deficiencies in Existing Research

Current research still exhibits three significant gaps: First, a lack of a full lifecycle perspective. Most studies treat the training or inference stages in isolation, failing to integrate analysis across the entire chain from chip manufacturing and data center construction to model deployment and hardware disposal, thus making it difficult to reflect the true total environmental cost of AI models. Second, severe under-assessment of the inference stage. While model training represents a concentrated energy "peak", it is a one-time event; the inference energy consumption generated by massive user calls is continuous and large-scale, and its cumulative impact may far exceed that of the training stage. Yet, related quantitative research and impact assessment remain extremely scarce. Third, weak socio-technical systems analysis. Existing research often stems from an engineering and technical perspective, lacking sufficient analysis of the deeper socio-institutional factors driving the high-carbon race-such as economic logic, governance gaps, and industry culture-leading to solutions that are mostly confined to the technical sphere and lacking a research vision that promotes systemic change. This study aims to address these gaps by constructing a full lifecycle assessment framework, strengthening inference stage analysis, and incorporating a multidimensional perspective of technology-economy-governance to systematically explain the causes of the high-carbon dilemma in large models and propose comprehensive transformation pathways.

2. Assessment Framework and Current Analysis of the Environmental Impact of Large Models

To systematically assess the environmental impact of large models, it is essential first to clarify the assessment boundaries, indicators, and framework. Current research predominantly concentrates on the direct energy consumption of the training stage. To fully reveal the true environmental footprint of large models, this chapter constructs a full lifecycle assessment framework and conducts an in-depth analysis of three key stages: training, inference, and hardware infrastructure.

2.1. Definition of Core Concepts and Construction of the Assessment Framework

2.1.1. "Carbon Footprint" Boundary in a Full Lifecycle Perspective

This study employs the Lifecycle Assessment (LCA) framework to define the carbon footprint of large models, covering all stages from "cradle to grave". Specifically, this includes: Upstream, covering the energy and resource consumption during the raw material extraction, manufacturing, and transportation of dedicated computing chips, servers, network equipment, and data center buildings. Midstream, including the direct energy consumption of the model training and inference deployment stages, as well as the operational energy consumption of data centers supporting these computing activities. Downstream, involving the environmental

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

impact generated during the recycling, disposal, or landfilling of hardware after it becomes obsolete due to technological iteration or physical wear and tear.

Focusing solely on midstream operational energy consumption would significantly underestimate the true environmental impact of large models. Research indicates that the manufacturing stage carbon footprint of high-performance computing hardware may constitute a substantial proportion of its full lifecycle footprint, representing an "implicit cost" that cannot be ignored.

2.1.2. Key Assessment Indicators

To comprehensively measure the environmental impact of large models, this study focuses on the following core indicators: Energy consumption, measured in megawatt-hours (MWh), with particular attention to data center Power Usage Effectiveness (PUE). A lower PUE value indicates less energy used for auxiliary functions like cooling. Carbon emissions, calculated as greenhouse gas emission equivalents based on the energy consumption structure and grid carbon intensity, expressed in metric tons of CO_2 equivalent (tCO_2e), serving as the key indicator for assessing climate impact. Water usage, encompassing both direct water use and indirect water use in data center cooling processes, measured in cubic meters (m^3). Electronic waste, including the solid waste generated from discarded hardware and the environmental risks posed by contained heavy metals and other hazardous substances.

2.1.3. Three-Tier Analytical Framework

Based on the aforementioned boundaries and indicators, this study constructs a three-tier analytical framework to systematically deconstruct the environmental impact of large models: Tier 1: Training stage – high-intensity, centralized, one-time energy and emission "peak". Tier 2: Inference stage – low-intensity, distributed, continuously accumulating energy and emission "baseload". Tier 3: Hardware infrastructure – the physical support system and its implicit environmental costs that persist throughout the model's entire lifecycle.

2.2. Training Stage: The One-Time "Energy Behemoth"

Model training is the first and most scrutinized environmental impact phase in a large model's lifecycle, requiring the investment of massive computational resources before the model becomes operational.

2.2.1. Exponential Growth in Computing Demand

Research indicates that since 2012, the computing power required to train the largest AI models has doubled approximately every 3.4 months on average, a growth rate far exceeding the speed of hardware efficiency improvements. Taking GPT-3 as an example, its training computational consumption is estimated at about 1,300 petaFLOP/s-day. Subsequent larger models generally exhibit orders-of-magnitude growth in training compute, driven primarily by parameter scale expansion, training dataset growth, and increased complexity of training strategies.

2.2.2. Carbon Emission Estimates and Case Studies

Strubell et al. (2019) first quantified the training carbon footprint of natural language processing models: Training a Transformer model with 213 million parameters could result in emissions up to approximately 0.63 tons of CO_2e. The study further extrapolated that training a super-large model combined with extensive neural architecture search could result in emissions as high as 284 tons, equivalent to the lifecycle emissions of five gasoline-powered cars. Subsequent estimates for models like GPT-3 (Lacoste et al., 2019) show training emissions ranging from hundreds to over a thousand tons, with specific values highly dependent on the energy mix of the data center used for training.

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

2.2.3. The "Multiplier Effect" of Hyperparameter Search and Repeated Training

In actual R&D processes, to find the optimal model architecture and hyperparameter combinations, extensive exploratory experiments are often required. The energy consumption from these can be tens of times that of the final successful training run. Furthermore, model retraining due to software debugging, hardware failures, or continuous optimization further amplifies the environmental impact. This "multiplier effect" means the real environmental cost of the training stage far exceeds the theoretical minimum.

2.3. Inference Stage: The Persistent "Undercurrent of Emissions"

The process of deploying a trained model online to provide real-time services to a vast number of users is called inference. Although the energy consumption per request is low, its cumulative scale effect is significant.

2.3.1. The Underestimated Cost of Inference

Schwartz et al. (2020) point out that the total inference energy consumption over the lifecycle of a widely used model is likely to far exceed its training energy consumption. Inference energy consumption is a function of model complexity and user scale. As large models are integrated into diverse scenarios like search engines and office software, their cumulative emissions will form a non-negligible "undercurrent".

2.3.2. Energy Efficiency Differences Across Service Modes

The environmental efficiency of inference is significantly influenced by the deployment mode: Cloud-based centralized API calls, provided by major cloud service providers, can achieve higher energy efficiency through economies of scale, advanced cooling technologies, and load balancing. If connected to a high proportion of renewable energy, carbon emissions can be significantly reduced. Local or edge deployment, where models are deployed on user devices or private servers, reduces transmission losses but typically suffers from lower hardware efficiency, inadequate thermal management, and difficulty leveraging green energy. This mode is more suitable for smaller, specialized models optimized via distillation. Overall, professional, centralized cloud services generally hold more potential for energy and carbon efficiency, but actual performance depends on the service provider's energy mix and management practices.

2.4. Implicit Environmental Costs of Hardware and Infrastructure

The operation of large models relies on extensive physical infrastructure, the environmental cost of which is often hidden behind digital services.

2.4.1. High Resource Consumption in Chip Manufacturing

The manufacturing of high-end GPUs/TPUs is a resource-intensive process involving hundreds of precise steps, consuming large amounts of high-purity chemicals, water, and electricity. Studies suggest that the carbon emissions from manufacturing a single high-end graphics card may be equivalent to its operational emissions over several years. Furthermore, chip manufacturing relies on critical minerals like rare earths, whose extraction and refining are accompanied by significant ecological damage and pollution.

2.4.2. Energy and Water Pressure from Data Centers

Data centers are the core facilities supporting computational power, with a significant portion of their energy consumption used for cooling. Although efficient cooling technologies like liquid cooling are gradually being adopted, their operation still relies on continuous electricity and substantial water resources. Particularly in arid regions, large data centers using evaporative cooling can consume millions of liters of water daily, equivalent to the daily water use of a city with tens of thousands of people, exacerbating regional water stress.

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

2.4.3. Rapid Iteration and E-Waste

Driven by the pursuit of higher performance, the refresh cycle for AI hardware has shortened to 1-2 years, leading to the premature disposal of large quantities of still-functional equipment. This equipment contains toxic substances like lead and mercury, which can cause long-term soil and water pollution if not properly handled. Moreover, the shortened hardware usage cycle diminishes the efficiency of amortizing its manufacturing-stage environmental costs, thereby increasing the environmental footprint per unit of computation over its full lifecycle.

3. Analysis of the Roots of the High-Carbon Dilemma: Technological, Economic, and Governance Perspectives

The high-carbon characteristics exhibited by large models are systemic results deeply embedded in their development paradigms, market mechanisms, and governance frameworks. This chapter systematically analyzes the formation logic of the high-carbon dilemma from three interrelated perspectives-technological path, economic incentives, and governance gaps-revealing the structural tension between the current trajectory of AI development and sustainable development goals.

3.1. Technological Path Dependence and the "Performance-First" Evaluation Culture

Current artificial intelligence, especially in the field of deep learning, has long been dominated by the "scaling law"[1]. This law posits a predictable power-law relationship between model performance and model scale, data volume, and computational power. This empirical rule has objectively propelled a technology evolution path reliant on computational expansion, making the increase in model scale a primary benchmark for measuring technological progress.

3.1.1. Unidirectional Development Mode Driven by the Scaling Law

The "bigger is better" development logic has led research institutions and companies into a competition centered on parameter scale[4]. The focus of technological innovation has been excessively concentrated on achieving the superposition of computing power and data scale, rather than exploring fundamentally more efficient algorithms or alternative intelligence paradigms that might break the scaling law. The result is that technological evolution is locked onto a high-energy-consumption path, where environmental costs increase exponentially with marginal performance gains.

3.1.2. Marginalization of Environmental Dimensions in Evaluation Systems

Accompanying the above technological path is a "performance-first" evaluation culture within the AI community. In mainstream academic conferences and industry benchmarks, the core standard for evaluating model quality is typically its accuracy on benchmark datasets, while environmental sustainability indicators like energy efficiency and carbon efficiency have long remained on the margins. This bias in value prioritization leaves researchers with little intrinsic motivation to develop green algorithms, thereby suppressing the R&D and adoption of high-efficiency, lightweight technologies.

3.2. Economic Incentives and Market Competition Structure

Behind the choice of technological path lie significant economic rationales and market structural factors. The current large model industry exhibits clear platform economy characteristics, where the market expectation of "winner-takes-all" further intensifies the competition for resource investment.

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

3.2.1. Computing Power as a Strategic Asset and Competitive Barrier

In the industrial landscape of "model-as-a-service", companies possessing leading computing power and model performance have the potential to build ecosystem advantages and achieve market dominance. Therefore, leading enterprises view large-scale computing power reserves as core strategic assets and competitive barriers. Over-investing in training resources is not only aimed at optimizing models but also serves the strategic intent of demonstrating technical strength, attracting capital and talent, and building industry moats. In this logic, computing power consumption has transformed from an R&D cost into a tool for market competition.

3.2.2. Environmental Externalities and Market Failure

The environmental costs-such as carbon emissions and water consumption-arising from enterprise model training and operations are not fully internalized through effective mechanisms. Instead, they are borne by society as "negative externalities". In the absence of policy interventions like carbon pricing or resource taxes, companies do not pay the corresponding cost for this portion of environmental impact, leading to distorted market price signals. The market mechanism thus fails to guide resources towards greener technologies. A structural conflict arises between the competitive strategy of individual firms pursuing optimal performance and the collective goal of low-carbon sustainable development for society as a whole.

3.3. Governance Gaps and Lack of Transparency

When technological paths become entrenched and market incentives become distorted, a sound governance system and transparency mechanisms are key to correcting developmental biases. However, the governance framework for the environmental impact of large models remains in its preliminary stages.

3.3.1. Lack of Unified Accounting Standards and Assessment Methods

Currently, there is a lack of widely accepted, comparable standards for carbon footprint accounting and reporting within the industry. When disclosing environmental data, companies often adopt their own approaches regarding system boundaries, emission factor selection, and allocation methods. This makes data difficult to compare and verify, reducing the visibility and credibility of the industry's overall environmental performance.

3.3.2. Insufficient Corporate Environmental Information Disclosure

Mirroring the "black box" nature of the models themselves, major companies also generally lack transparency in disclosing environmental data. Key environmental information such as training locations, energy mix, actual PUE, and water consumption is often treated as commercial secrets and not made public. This state of information opacity hinders external oversight, academic research, and public awareness, making accurate environmental impact assessment and societal discussion difficult.

3.3.3. Gaps in Policy and Regulation

Compared to the increasingly developed legislation and regulation in areas like algorithmic fairness and data security, specialized policies targeting the carbon footprint of AI remain relatively scarce globally. Mandatory energy efficiency standards, carbon disclosure requirements, or industry access mechanisms based on environmental impact have not been systematically established. Regulatory lag effectively acquiesces to the continuation of the current development model, allowing enterprises to persist on the high-carbon competitive path in the absence of clear rules and expectations.

The high-carbon dilemma of large models stems from the systemic dysfunction arising from the coupling and mutual reinforcement of technological path, economic incentives, and governance systems. The scaling law and performance culture shape a preference for high-energy-

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

consumption technology; market competition structures drive strategies of computing power expansion, while environmental externalities are not internalized; the absence of governance standards, insufficient information disclosure, and gaps in policy regulation leave this path lacking external correction mechanisms. Therefore, cracking the high-carbon dilemma must go beyond single-dimensional technological or policy measures, requiring systematic, multidimensional, cross-stakeholder collaborative intervention and institutional restructuring.

4. Towards Green AI: A Multi-Level Optimization Pathway

Addressing the high-carbon dilemma of large models cannot rely on singular technological breakthroughs or isolated policy tools. It requires a coordinated transformation spanning technology R&D, system operations, industry norms, and core values. This chapter constructs a systemic Green AI transition pathway from four mutually supportive levels-algorithm/model, system/operations, industry/policy, and culture/ethics-aiming to deeply integrate environmental sustainability into the DNA of AI development.

4.1. Algorithm and Model Level: "Slimming and Enhancing Efficiency"

Designing more efficient and streamlined models from the outset is the most fundamental way to reduce environmental footprints. This necessitates a shift from pursuing "absolute performance" to seeking the Pareto optimum of "performance-efficiency".

4.1.1. Model Architecture Innovation

Future model design must break the dependence on mere parameter stacking and move towards more intelligent architectures.

Sparsification and Conditional Computation: For example, mixture of experts models, where the core idea is to have different "expert" sub-networks specialize in processing different types of input, activating only the relevant experts during inference, significantly reducing computation.

More Efficient Attention Mechanisms: The self-attention mechanism in Transformers has a computational complexity that grows quadratically with sequence length. Researching variants like linear attention and sliding window attention can substantially reduce computation for long sequence processing while maintaining performance[5].

Neural Architecture Search (NAS) and AutoML: Utilizing NAS technology to automatically search for the optimal model structure under given computational budgets or energy efficiency constraints.

4.1.2. Model Compression and Simplification

For already-trained large models, post-processing techniques can make them more suitable for efficient deployment.

Knowledge Distillation: Using a large, high-performance "teacher model" to guide the training of a lightweight "student model", enabling the student to achieve performance close to the teacher with a much smaller size.

Quantization: Converting model weights and activations from high-precision floating-point formats to lower precision formats, reducing memory usage and computational energy consumption, particularly effective for inference acceleration.

Pruning: Identifying and removing redundant weights, neurons, or even entire layers from the model, resulting in a sparser, more compact network with minimal performance loss.

4.1.3. Efficient Training Strategies

There is significant room for improving the energy efficiency of the training process itself.

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

Dynamic Curriculum Learning and Early Stopping: Designing smarter data sampling strategies and using validation set monitoring to stop training promptly when performance plateaus can accelerate convergence.

Carbon-Aware Hyperparameter Optimization: Incorporating the predicted carbon emissions of the training process as one of the optimization objectives, alongside model performance metrics, into hyperparameter search algorithms to automatically find "greener" configurations.

Advanced Optimization Algorithms: Developing optimization algorithms with faster convergence and employing better initialization and normalization techniques to reduce the number of training iterations required.

4.2. System and Operations Level

Once a model is established, fine-grained infrastructure and operational management can significantly reduce its runtime environmental impact.

4.2.1. Carbon-Aware Scheduling

Leveraging the spatiotemporal variability of grid carbon intensity to dynamically schedule computing tasks.

Time-Shifting: Scheduling non-real-time, large-scale training tasks for nights or weekends when grid load is low and the proportion of wind, solar, and other renewables may be higher.

Geographical Load Balancing: Within globally distributed data center networks, directing computing loads in real-time to data center regions currently using the highest proportion of renewable energy. Companies like Google and Microsoft are already practicing this.

4.2.2. Mixed-Precision Computing and Hardware-Software Co-Optimization

Mixed-Precision Training and Inference: Mixing low-precision formats (e.g., FP16, BF16) with high-precision formats (e.g., FP32) during training and inference. This ensures numerical stability while fully utilizing modern AI accelerators' high efficiency for low-precision computation, improving throughput and reducing energy consumption.

Hardware-Software Co-Design: Designing specialized hardware optimized for specific efficient algorithms or optimizing system software stacks to maximize hardware energy efficiency.

4.2.3. Model Sharing and Reuse Ecosystem

Promoting Pre-trained Model Libraries and Open-Source Culture: Encouraging the development of platforms like Hugging Face, enabling research institutions and SMEs to directly fine-tune and use open-source large-scale pre-trained models, avoiding the enormous environmental cost of training from scratch.

Establishing Model Lifecycle Registration and Traceability Mechanisms: Recording information such as a model's training energy consumption, carbon emissions, and performance, facilitating evaluation and selection by subsequent users and promoting the circulation and reuse of efficient models.

4.3. Industry and Policy Level

Market failures and governance gaps require strong industry consensus and policy intervention to correct, creating a fair and sustainable competitive environment for Green AI.

4.3.1. Establishing Standards and Certification Systems

Developing AI Carbon Footprint Accounting Standards: Established by international standards organizations (e.g., ISO), industry associations, and academic institutions to clarify system boundaries, accounting methodologies, emission factors, and data disclosure formats, ensuring comparability and credibility of data.

Introducing AI Energy/Carbon Efficiency Labels and Benchmarks: Similar to appliance energy labels, establishing energy efficiency benchmarks for different categories and scales of AI

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

models and services (e.g., the energy efficiency track in MLPerf) and introducing certification labels to guide market choice.

4.3.2. Applying Economic and Policy Instruments

Carbon Pricing and Differentiated Electricity Pricing: Extending carbon tax or emissions trading systems to cover data centers and large computing facilities, or linking data center electricity consumption to real-time grid carbon intensity with floating tariffs, directly internalizing environmental costs.

Green Public Procurement and Subsidies: Making carbon footprint a significant evaluation criterion in government and corporate procurement of AI services. Establishing R&D subsidies or tax incentives specifically supporting research in Green AI algorithms, systems, and hardware.

4.3.3. Strengthening Information Disclosure and Auditing

Mandatory Environmental Impact Disclosure: Requiring AI projects exceeding certain computational scale or energy consumption thresholds to simultaneously disclose key environmental indicators (energy consumption, carbon emissions, water use) for both training and estimated inference stages upon release, subject to third-party auditing.

4.4. Culture and Ethics Level

The deepest transformation lies in reshaping values and paradigms, internalizing sustainability from an external constraint into the core ethics of the discipline and industry.

4.4.1. Advocating a "Comprehensive Evaluation" Academic and Industrial Culture

Reforming Paper Review and Competition Evaluation Standards: Top-tier conferences and journals should require submitted papers to report model training energy consumption and carbon emissions, considering these as important references in assessing contribution. Competitions should establish "efficiency tracks", rewarding solutions that achieve the best performance under strict computational budgets.

Promoting "Responsible AI Research" Charters: Elevating environmental responsibility to stand alongside research integrity, fairness, and safety as core ethical principles for AI researchers.

4.4.2. Developing the Principle of "AI Appropriateness"

Advocating Task-Matched Model Selection: Not all applications require trillion-parameter models. Developers should be encouraged to first assess task requirements, prioritizing lightweight models, traditional machine learning methods, or even rule-based systems. "Small models, big intelligence" should become a valued engineering aesthetic[2].

Critically Examining Demand: For certain high-energy-consumption applications with questionable social value or extremely low marginal benefit, the industry and society should engage in ethical reflection and self-restraint.

4.4.3. Strengthening Whole-Chain Social Responsibility

Researcher Responsibility: Considering environmental costs at the experimental design stage, choosing more efficient baselines and methods.

Developer and Engineer Responsibility: Treating energy efficiency optimization as a key performance indicator in system design and implementing best practices in operations.

Corporate Leadership Responsibility: Integrating environmental sustainability into corporate strategy, setting public emission reduction targets, and linking executive compensation partly to environmental performance.

The path towards Green AI is one of co-evolution, progressing from micro-level technological innovation, to meso-level system optimization, to macro-level rule restructuring, and ultimately reaching a deep paradigm revolution. This requires a concerted effort from technical

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

experts, business leaders, policymakers, the academic community, and the public to collectively calibrate AI's development trajectory in a direction compatible with a sustainable future. Only by completing this systemic transformation can artificial intelligence truly become a tool for addressing environmental challenges, rather than a perpetually expanding new problem.

5. Conclusion and Outlook

5.1. Research Conclusion

Through systematic assessment of the full lifecycle environmental impact of large models, root-cause analysis of their high-carbon dilemma, and exploration of comprehensive optimization pathways, this study arrives at the following core conclusions:

5.1.1. The Development of Large Models Currently Exhibits Significant "High-Carbon Technology" Characteristics, and Their Environmental Impact is a Systemic Issue Present Throughout Their Lifecycle.

The carbon footprint of large models extends far beyond the one-time energy consumption of training. It encompasses a complete chain from the energy-intensive manufacturing of hardware, to the concentrated burst of model training, to the long-term, sustained, and massive-scale inference services, and finally to hardware disposal. Among these, the inference stage, due to its distributed and normalization nature, has had its cumulative environmental cost severely underestimated for a long time, becoming a potential "undercurrent of emissions". Without imposing systematic constraints on this full lifecycle impact, the expansion of the AI industry will fundamentally conflict with global carbon reduction goals.

5.1.2. This High-Carbon Dilemma is Rooted in the Deep Coupling and Mutual Reinforcement of a Triple Logic.

It is not a singular technological defect but a systemic issue. At the technological level, the "scaling law" and "performance-first" culture jointly lock development onto a path of extensive reliance on computing power expansion. At the economic level, a "winner-takes-all" market structure incentivizes enterprises to view computing power reserves as a strategic barrier, while environmental costs, as externalities, fail to be effectively priced by the market, leading to market failure. At the governance level, unified accounting standards, mandatory transparency in disclosure, and targeted policy regulation are all in a state of vacancy, allowing high-carbon practices to proceed without external constraints or corrective mechanisms. These three layers are nested together, collectively constituting the deep-seated structure that makes the current development paradigm resistant to autonomous change.

5.1.3. Moving Towards "Green AI" is a Complex Socio-Technical Systems Engineering Endeavor.

No single-point breakthrough will suffice; it must rely on the coordinated evolution of multiple levels and multiple actors. A genuine transformation requires the synergy of four pillars: At the algorithm/model level, achieving "slimming and efficiency enhancement" through architectural innovation and compression/simplification; at the system/operations level, achieving "meticulous calculation" through carbon-aware scheduling and ecosystem sharing; at the industry/policy level, conducting "guidance and regulation" through standard-setting, carbon pricing, and green procurement; and ultimately, at the culture/ethics level, completing the paradigm shift from "bigger and stronger" to "better and greener" by advocating comprehensive evaluation and practicing the principle of appropriateness. Only when technological breakthroughs, economic adjustments, policy interventions, and ethical reflection converge can environmental sustainability be embedded into the core DNA of AI development, moving from a peripheral topic to a central principle.

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

5.2. Future Outlook

Based on this study, future work can delve deeper in the following directions:

5.2.1. Research Frontiers

Develop more precise dynamic environmental impact tracking tools. Future research needs to develop standardized, automated software tools capable of real-time tracking and predicting the energy consumption, carbon emissions, and water footprint of AI workloads from chips to the cloud, with dynamic visualization to support fine-grained management and decision-making.

Expand assessment boundaries to broader environmental dimensions. Beyond carbon emissions, future assessments need to incorporate more environmental dimensions such as biodiversity impact, rare resource consumption, and full lifecycle toxicity and pollution, forming a more comprehensive "planetary health" impact assessment framework.

Construct benchmarks and theory for model "environmental efficiency". There is a need to establish an "environmental efficiency" assessment theory analogous to the "performance-efficiency" Pareto frontier, defining and measuring the comprehensive intelligence benefits gained per unit of environmental cost, promoting a fundamental shift in evaluation paradigms.

5.2.2. Practical Challenges

The trade-off dilemma between green performance and model capability. In practice, techniques like model compression and quantization often come with slight performance losses. Defining acceptable "performance-efficiency" trade-off points for different application scenarios and convincing users to accept a certain performance premium or latency for "greenness" will be key challenges for market acceptance.

The complexity of global collaborative governance. The AI industry chain and computing power networks are globally distributed, while environmental policies are regional. Coordinating carbon accounting standards, cross-border data rules, and regulatory requirements across different jurisdictions to avoid "carbon leakage" and regulatory arbitrage, and establishing effective global governance coordination mechanisms, are extremely complex political-economic issues.

Conflict between short-term commercial interests and long-term sustainable development. Transitioning to Green AI may mean increased R&D investment and adjusted competitive rhythms in the short term. Designing transitional policies and incentive mechanisms to balance enterprise survival pressures with long-term social responsibility is a severe reality that policy design must confront.

5.2.3. Ultimate Vision

The ultimate outlook of this study is to call for and promote a profound reflection on the philosophy of AI development. Future AI should not merely be a passive object of assessment and optimization but should become an exemplary model actively practicing sustainability principles. This means:

From Tool to Exemplar: AI is not only a tool for optimizing energy grids and aiding climate science; its entire process of R&D, deployment, and operation should itself become a demonstration of the circular economy, resource minimization, and net-zero emissions.

From Add-on to Embedded: Environmental sustainability should not be an "add-on" constraint applied after the fact but should become an "embedded" core principle from algorithm conception, architecture design, and system development to business model innovation.

From Problem Source to Solution: Ultimately, through this systemic transformation, artificial intelligence has the potential to completely shed the label of "high-carbon technology"[7], accomplishing a fundamental identity shift from a contributor to environmental problems to a

ISSN: 2692-7608 DOI: 10.6981/FEM.202603_7(3).0015

provider and practitioner of systemic solutions, thereby truly matching its promised potential to create a better future.

References

[1] Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 3645-3650).

[2] Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.

[3] Osondu Joshua. (2025). Red AI vs. Green AI in Education: How Educational Institutions and Students Can Lead Environmentally Sustainable Artificial Intelligence Practices. Doi: 10.13140/RG.2.2.27929.12644.

[4] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120), 1-39.

[5] Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020, November). Transformers are rnns: Fast autoregressive transformers with linear attention. In International conference on machine learning (pp. 5156-5165). PMLR.

[6] Mesarčík, Matúš & Solarova, Sara & Podroužek, Juraj & Bielikova, Maria. (2022). Stance on The Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence – Artificial Intelligence Act. Doi: 10.31235/osf.io/yzfg8.

[7] Yigitcanlar, T. (2021). Greening the artificial intelligence for a sustainable planet: An editorial commentary. Sustainability, 13(24), 13508.