Green AI Benchmarking of Domestic Accelerator–Driven Foundation Models: Performance, Efficiency, and Carbon Footprint Analysis

Finn D. Rebarts
Department of Computer Science, George Mason University, Fairfax, VA, USA.
finnwork@gmu.edu

Jack Jenas
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
jackjones97@uab.edu

Congling Hou
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
conglinghou357@uc.edu

Abstract

The rapid proliferation of large-scale foundation models has brought the environmental sustainability of artificial intelligence into sharp focus, prompting the emergence of Green AI as a rigorous paradigm for evaluating performance jointly with energy consumption and carbon emissions. This paper presents a comprehensive benchmarking framework for domestic accelerator–driven foundation models, with an emphasis on edge-native architectures trained entirely on non–GPU hardware such as China-developed AI accelerators. The study systematically examines the interplay among model throughput, inference latency, energy efficiency, training dynamics, and cradle-to-grave carbon footprint across heterogeneous hardware ecosystems. We adopt a system-level perspective that integrates infrastructure considerations, software–hardware co-design, data center energy mix, and supply chain factors into a unified analytical lens. By comparing domestic accelerator–based systems against conventional GPU–centric deployments, the paper reveals structural trade-offs that transcend simple floating-point operations per second comparisons, highlighting the importance of memory bandwidth utilization, compiler maturity, and workload scheduling. The analysis extends to governance and policy dimensions, discussing how national AI sovereignty, semiconductor supply chain resilience, and carbon accounting standards interact with system design choices. Through this multi-layered evaluation, the paper argues for an expanded definition of computational efficiency that incorporates geopolitical, infrastructural, and lifecycle carbon costs, while proposing benchmarking methodologies that align with emerging regulatory frameworks for sustainable computing. The findings indicate that domestic accelerator ecosystems, despite immature software stacks, can yield significant carbon efficiency advantages when coupled with regionally optimized energy grids and tailored model architectures, underscoring the need for context-aware Green AI metrics.

Keywords

Green AI, foundation models, domestic AI accelerators, carbon footprint, edge-cloud computing, benchmarking, sustainability.

1. Introduction

The emergence of foundation models with billions of parameters has reshaped contemporary artificial intelligence, delivering unprecedented capabilities in natural language understanding, computer vision, and multimodal generation. Yet this capability comes at a steep environmental cost. The energy consumed during training and inference of state-of-the-art models has surged by orders of magnitude over the past decade, attracting intense scrutiny from researchers, policymakers, and the public. Early landmark studies documented that training a single large transformer model can emit as much carbon dioxide as several transatlantic flights, forcing the community to reckon with the climate implications of ever-larger architectures [1]. These revelations catalyzed the Green AI movement, which advocates for making energy efficiency and carbon transparency first-class metrics alongside accuracy [2]. While initial Green AI research focused predominantly on model architecture innovations and training optimizations within GPU-dominated data centers, a new dimension has emerged with the rise of domestically produced AI accelerators, particularly those developed in China and deployed in edge-cloud configurations. These hardware ecosystems, built on non-NVIDIA architectures such as Huawei's Ascend series or other China-specific neural processing units, promise to reshape not only the geopolitics of AI infrastructure but also the energy profiles of foundation model deployment.

Understanding the sustainability implications of these accelerator-driven systems demands a holistic benchmarking methodology that moves beyond simple energy-per-operation measurements. The unique characteristics of domestic accelerators—including their memory hierarchies, quantization support, compiler toolchains, and integration with regional energy grids—require a system-level analytical lens. Moreover, foundation models deployed on such hardware often target edge-native scenarios where inference occurs close to the data source, reducing data transmission energy but introducing new performance constraints. This paper undertakes a conceptual and methodological examination of Green AI benchmarking for foundation models trained and served on domestic accelerators. We propose an integrated framework that simultaneously captures performance, efficiency, and carbon footprint while accounting for the structural particularities of emerging hardware ecosystems. In doing so, we address the following overarching questions: What system-level factors differentiate domestic accelerator–based foundation models from conventional GPU deployments in terms of energy proportionality and carbon efficiency? How should benchmarking methodologies adapt to edge-native architectures with heterogeneous hardware backends? And what policy and governance mechanisms can ensure that the pursuit of AI sovereignty aligns with global climate objectives?

To ground this inquiry, we draw on a diverse body of prior work spanning carbon accounting for machine learning, hardware-software co-design for AI accelerators, and infrastructure-level sustainability metrics. We also examine a concrete case: the JuZhou 1.0 edge-native text-to-image model, which represents the first foundation model trained entirely on China-developed AI accelerators [11]. By situating our analysis at the intersection of systems engineering, environmental science, and technology policy, the paper contributes a blueprint for interdisciplinary Green AI assessment that is acutely sensitive to the infrastructural and geopolitical realities of the 2020s.

2. Evolving Landscape of AI Carbon Accounting and Hardware Heterogeneity

The quantification of the environmental impact of AI systems has progressed from back-of-the-envelope estimates to sophisticated lifecycle assessment tools. Initial investigations into

the carbon footprint of deep learning established the central role of electricity consumption during GPU-hours of computation, typically multiplied by a fixed carbon intensity factor [3]. Subsequent work refined this approach by incorporating regional differences in grid carbon intensity, renewable energy procurement, and hardware manufacturing emissions. The BLOOM project, for example, provided a detailed public accounting of the training emissions of a 176-billion-parameter multilingual language model, revealing that location and time-of-day scheduling could alter the carbon impact by more than an order of magnitude [12]. Such studies have underscored the importance of temporal and spatial granularity in carbon accounting, yet they have largely focused on GPU-centric compute clusters located in a small number of cloud regions.

The rise of domestic AI accelerators introduces several new variables that complicate existing accounting frameworks. These accelerators often exhibit fundamentally different power scaling curves than GPUs due to their customized arithmetic units and on-chip memory designs. For instance, many domestic neural processing units emphasize integer or low-precision floating-point operations for inference, which can dramatically reduce energy per token relative to FP32 or FP16 GPU execution, provided the model is appropriately quantized with minimal accuracy loss. However, the energy savings at the chip level may be partially offset by inefficiencies in the software stack. Compiler maturity, operator fusion capabilities, and the availability of optimized libraries directly influence achieved hardware utilization, which can vary by as much as forty percent across different accelerator generations. None of these system-level interactions are captured by conventional metrics such as TOPS (tera operations per second) or floating-point efficiency, demanding new benchmarking constructs.

Lifecycle assessment must also incorporate the embodied carbon of the accelerators themselves. The semiconductor manufacturing process for advanced nodes below seven nanometers is exceptionally energy-intensive, and the carbon cost of producing a single accelerator chip can rival the operational emissions incurred over years of inference workloads, depending on utilization rates and grid mix. Domestic accelerator supply chains, which often rely on regionally sourced silicon, packaging, and assembly, may have distinct embodied carbon profiles compared to the globally distributed supply chains of leading GPU vendors. A comprehensive Green AI benchmark must therefore integrate embodied carbon models tailored to specific foundry locations and manufacturing processes. Furthermore, edge-native deployments reduce the need for large-scale data center cooling infrastructures, shifting the balance further toward operational efficiency if local energy grids are sufficiently clean. These intertwined hardware-software-infrastructure dynamics form the substrate for the following sections.

3. System Architecture and Infrastructure Configuration

A foundational premise of this work is that performance, efficiency, and carbon footprint cannot be meaningfully benchmarked without a detailed characterization of the underlying system architecture. Domestic accelerator–driven foundation models typically operate within a heterogeneous compute fabric that spans edge devices, on-premise server clusters, and regional cloud nodes. The edge-native paradigm exemplified by certain text-to-image models relies on a tiered architecture where a lightweight inference engine on the accelerator handles time-sensitive generation tasks, while a heavier training or fine-tuning backend resides in a centralized facility connected via high-speed low-latency links. This architectural split introduces novel trade-offs: pushing computation to the edge reduces network energy and

latency but may increase total energy if inference must be performed on less efficient device-level chips due to thermal and power constraints.

The memory subsystem plays a particularly critical role. Domestic accelerators frequently incorporate high-bandwidth memory (HBM) or proprietary on-chip static RAM caches that differ markedly from the GDDR or HBM configurations found on consumer and data center GPUs. For large foundation models, memory bandwidth often becomes the primary bottleneck during autoregressive generation, where each token prediction requires a full traversal of the model weights. An accelerator with superior peak TOPS but lower memory bandwidth will underperform in tokens-per-second and joules-per-token when serving a transformer with billions of parameters. Moreover, the ability to keep the model weights resident in accelerator memory without offloading to host DDR memory, a capability influenced by memory capacity and compression techniques, significantly influences energy efficiency. Hence, any benchmark must report memory bandwidth utilization and weight residency ratios alongside traditional throughput metrics.

On the software side, the compiler and runtime stack is equally decisive. The gap between theoretical and achieved performance on domestic accelerators is often wider than on mature GPU ecosystems, primarily due to the relative immaturity of kernel libraries for attention mechanisms, layer normalization, and fused activation functions that are essential to transformer models. Progress in graph-level optimizers and automated parallelization frameworks has begun to close this gap [8]. Nevertheless, current benchmarking practices must explicitly document the software versions, operator schedules, and precision configurations used, as small changes in these parameters can lead to large swings in energy efficiency. Infrastructure orchestration also matters: containerized deployment on Kubernetes clusters with dynamic resource scaling can improve overall cluster utilization and power usage effectiveness (PUE) compared to static provisioning, but may introduce scheduling overhead that harms tail latency. These infrastructure design choices interact deeply with carbon accounting, as higher PUE values amplify the carbon cost of every joule consumed by the accelerator.

4. Performance and Efficiency Analysis Under Heterogeneous Constraints

Benchmarking performance and efficiency of foundation models on domestic accelerators requires moving beyond the single-metric FLOPS-based comparisons that have historically dominated accelerator evaluations. In a holistic Green AI framework, we argue for a multidimensional metric space that simultaneously reports throughput (tokens per second or images per second), energy consumption per unit of useful work (joules per token or joules per image), tail latency at specified percentiles, and accuracy retention after quantization or pruning. The appropriate weighting of these dimensions depends on the deployment context: an interactive chatbot service may prioritize tail latency and tokens-per-watt, while a batch offline image generation pipeline may tolerate higher latency in exchange for maximal throughput per carbon unit.

A critical aspect of efficiency analysis in domestic accelerator ecosystems is sensitivity to batch size and concurrency. Many custom neural processing units are optimized for high throughput at large batch sizes by exploiting extensive data-level parallelism, yet real-world edge serving often encounters bursty, low-concurrency workloads. In such regimes, fixed overheads from kernel launch, memory transfer, and pipeline bubbles can dominate energy consumption, driving down effective efficiency even when peak throughput metrics appear competitive. Modeling this behavior demands a parametric performance model that captures

the accelerator's latency–throughput Pareto frontier under varying load. Recent work on serving efficiency for large language models has shown that continuous batching and iteration-level scheduling can dramatically improve hardware utilization [9]. Adapting these techniques to domestic accelerators, with their unique warp or tile sizes and synchronization primitives, is an open systems challenge with profound implications for Green AI performance benchmarking.

Temperature-dependent clock throttling and power capping introduce further non-linearities. Domestic accelerators deployed in edge environments often lack the sophisticated cooling infrastructures of hyperscale data centers, leading to thermal hotspots that trigger frequency scaling under sustained load. Consequently, steady-state efficiency measured after thermal saturation may differ substantially from burst-mode efficiency. A responsible benchmark must emulate realistic workload duty cycles and report efficiency metrics after thermal equilibrium is reached. At the system level, composability with host processors also matters: if the accelerator must frequently interrupt the host CPU for memory management or token sampling, inter-chip communication power can erode the energy benefits of the accelerator's specialized compute. All these considerations point to a need for open, reproducible profiling suites that are portable across domestic and GPU-based backends, enabling apples-to-apples efficiency comparisons grounded in end-to-end application workloads rather than microbenchmarks.

5. Carbon Footprint Assessment Across Lifecycle Stages

The carbon footprint of a foundation model is the cumulative result of emissions generated during hardware manufacturing, model training, inference serving, and eventual hardware disposal. While many prior studies have focused on training emissions, the operational carbon footprint of inference over the model's full deployment lifetime often dominates, especially for models with broad user reach [7]. Domestic accelerator–based edge architectures can shift this balance in important ways. By employing quantized models with low-bitwidth arithmetic optimized for specific inference chips, the total energy per inference can be reduced by a factor of two to five compared to unquantized GPU baselines, assuming careful accuracy calibration. However, these savings at the inference phase may be contingent on a more expensive and energy-intensive quantization-aware training process, which itself consumes accelerator-hours and emits carbon. A truly lifecycle-oriented assessment must integrate the training carbon cost amortized over the expected total inference volume, yielding a metric of grams of CO_2 equivalent per million inference tokens or images.

The carbon intensity of the electrical grid that powers both training and inference is the single largest modulating factor. A model trained on a grid with an average intensity of 800 grams of CO_2 per kilowatt-hour will have a footprint roughly four times larger than one trained on a cleaner 200 grams per kilowatt-hour grid, even if all hardware and software configurations are identical. Domestic accelerator clusters are often geographically collocated with specific regional grids, reflecting national cloud infrastructure strategies. This opens the possibility of aligning training schedules with periods of high renewable penetration through carbon-aware workload orchestration, a practice that has been demonstrated on GPU clouds [10] but remains unexplored for domestically produced accelerators. By co-designing the workload scheduler with real-time grid signals, these systems could achieve substantial carbon reductions without any hardware changes.

Embodied carbon from chip manufacturing is a relatively neglected factor that becomes especially relevant when assessing newly established domestic accelerator supply chains.

Recent analyses have revealed that the carbon emissions from fabricating a single leading-edge wafer can exceed several kilograms of CO_2 per square centimeter, and that total embodied carbon of a server-grade accelerator can range from tens to hundreds of kilograms of CO_2 equivalent [6]. These upfront emissions must be amortized across the accelerator's total lifetime inference work; thus, the carbon payback period becomes a crucial metric. If an energy-efficient domestic accelerator replaces a less efficient legacy inference cluster, the operational carbon savings must be weighed against the embodied carbon debt created by manufacturing new hardware. In sustainable computing policy, principles of life extension, reuse, and cascading deployment of older accelerators for less demanding tasks become instrumentally important. The JuZhou 1.0 technical report documented an edge-native model trained entirely on China-developed accelerators, highlighting a low operational energy profile but also underscoring the need for comprehensive embodied carbon accounting as accelerator manufacturing scales [11].

6. Governance, Policy, and the Geopolitics of Green AI Infrastructure

Green AI benchmarking of domestic accelerator–driven foundation models cannot be divorced from the broader governance frameworks that shape technology development trajectories. National strategies for AI sovereignty have emerged as powerful drivers of domestic accelerator programs, motivated by concerns over semiconductor supply chain dependencies and export controls. These geopolitical dynamics create a dual imperative: to achieve self-sufficiency in AI hardware while adhering to increasingly stringent climate commitments. The tension between rapid domestic capacity building and rigorous environmental accounting presents a novel policy challenge. Policymakers may face incentives to prioritize performance metrics that demonstrate parity with GPU incumbents, potentially sidelining energy efficiency considerations unless carbon transparency mandates are encoded into procurement standards.

Carbon reporting standards for AI are still nascent but evolving rapidly. Initiatives such as the Machine Learning Emissions Calculator and the broader Green Software Foundation's Software Carbon Intensity specification aim to standardize how operational and embodied emissions are allocated to computational workloads [13]. Adapting these standards to heterogeneous accelerator environments requires extensions that explicitly model accelerator-specific power telemetry, manufacturing carbon databases, and country-specific grid carbon intensity factors. The existence of a domestically controlled accelerator ecosystem may actually facilitate more granular carbon accounting, as a single national entity could mandate consistent onboard power monitoring and public carbon disclosure for all foundation model training runs conducted on its hardware. This contrasts with the fragmented reporting landscape among global GPU cloud providers, where proprietary cloud services often obscure the energy mix and utilization details behind contractual walls. By embedding carbon telemetry into the accelerator driver stack and making it available through standardized APIs, domestic ecosystems could set a precedent for transparent and auditable Green AI reporting.

Equity and fairness considerations also intersect with Green AI policy. The carbon benefits of energy-efficient edge inference must be distributed equitably across different communities and regions. If low-carbon edge models are primarily deployed in affluent urban environments with access to renewable-powered micro-data centers, while marginalized regions continue to rely on carbon-intensive centralized inference, the overall climate burden may be regressively distributed. Moreover, the geopolitical imperative to build domestic accelerator industries may lead to the construction of new semiconductor fabrication facilities

that impose local environmental burdens on water and chemical usage. A just transition framework for sustainable AI must account for these localized impacts and incorporate community-level environmental justice metrics into Green AI benchmarking [14]. Without such multidimensional governance, the pursuit of low-carbon AI through domestic hardware acceleration risks externalizing ecological costs to the very communities it purports to benefit.

7. System-Level Trade-Offs and Architecting for Sustainability

Synthesizing the previous dimensions reveals a rich space of system-level trade-offs that must be navigated by architects of domestic accelerator–driven foundation models. The decision to pursue edge-native deployment, for instance, entails a deliberate sacrifice of peak floating-point throughput in favor of reduced data movement energy and latency. This trade-off is often justified for interactive applications where user-perceived responsiveness is paramount, but it requires careful benchmarking to ensure that the total system carbon per user request is indeed lower than a centralized cloud-based alternative once all infrastructure overheads are accounted for. The emergence of model cascades, where a small edge model handles common queries and escalates difficult cases to a larger cloud model, introduces a workload-dependent partitioning that has no single optimal configuration; instead, the greenest architecture varies with the query distribution, requiring adaptive orchestration informed by real-time carbon intensity signals.

Model compression techniques are integral to architecting for sustainability on domestic accelerators. Quantization not only reduces memory footprint but also enables the use of specialized low-bitwidth multiply–accumulate arrays that are far more energy efficient than their full-precision counterparts. However, aggressive compression may degrade model fairness across different demographic subgroups, a phenomenon that has been documented in compressed language and vision models [14]. Green AI benchmarking must therefore incorporate fairness metric regressions as a dimension of evaluation, creating a multi-objective optimization landscape where carbon minimization must not come at the cost of discriminatory performance degradation. This requirement dovetails with the governance themes discussed previously, as procurement guidelines for public-sector AI systems are beginning to mandate both energy efficiency and non-discrimination criteria, effectively compelling the integration of fairness audits into the benchmarking suite.

The software-hardware co-design space for sustainable foundation models extends into training methodology. Low-precision training, selective activation recomputation, and communication-efficient distributed algorithms can reduce training carbon by factors of several times while preserving model quality [5]. Yet these techniques require hardware support for dynamic mixed-precision operations and efficient all-reduce collectives, which may be at different levels of maturity on domestic accelerator interconnects compared to NVLink or InfiniBand fabrics. Benchmarking the training efficiency therefore demands careful measurement of interconnect bandwidth, communication computation overlap, and scaling efficiency as the number of accelerators increases. Poor scaling can erase carbon savings if the training run must consume many more aggregate accelerator-hours due to diminishing returns from parallelism. System-level benchmarking frameworks must thus probe these distributed system dynamics, not just single-chip metrics.

8. Conclusion

This paper has argued that Green AI benchmarking of domestic accelerator–driven foundation models necessitates a deeply interdisciplinary and system-level approach that transcends

conventional hardware performance comparisons. By weaving together the threads of energy accounting, hardware architecture, software stack maturity, lifecycle carbon modeling, and geopolitical governance, we have outlined a holistic framework for evaluating the sustainability of next-generation AI infrastructures. The analysis reveals that domestic accelerator ecosystems present both opportunities and risks for green computing. On the one hand, their co-designed nature, edge-native deployment potential, and alignment with nationally controlled energy grids can unlock order-of-magnitude improvements in carbon efficiency relative to GPU baselines when measured across the full lifecycle. On the other hand, software immaturity, opaque supply chain emissions, and the potential for governance capture by narrow sovereignty objectives threaten to undermine these environmental gains without deliberate and transparent benchmarking.

Our discussion has highlighted the importance of expanding the metric space of Green AI beyond FLOPs and joules to include fairness impacts, embodied carbon payback periods, thermal resilience, and distributed scaling behavior. The JuZhou 1.0 model serves as a concrete exemplar of the promise and complexity of this new paradigm, demonstrating that frontier foundation models can be constructed entirely outside the dominant GPU ecosystem while maintaining competitive quality [11]. However, the broader community must resist the temptation to treat such results as proof of universal eco-efficiency; context-aware evaluation that accounts for regional grid carbon intensity, workload patterns, and hardware refresh cycles remains essential. As AI systems become increasingly embedded in critical societal infrastructures, from healthcare to energy management, the methodologies for assessing their environmental footprint must evolve to match the sophistication of the technologies themselves.

Looking forward, we advocate for the establishment of an open, multi-stakeholder Green AI benchmarking consortium that includes domestic accelerator vendors, cloud operators, environmental scientists, and civil society representatives. Such a body could develop and maintain standardized benchmarking suites that are portable across hardware backends, mandate transparency in carbon reporting in both training and inference phases, and continuously update reference carbon intensity models to reflect real-world grid transitions. Only through such collaborative and transparent efforts can the global community harness the sustainability potential of foundation models while responsibly managing the environmental consequences of the accelerating AI hardware race.

References

1. 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.

2. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

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

4. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

5. Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., ... Hazelwood, K. (2022). Sustainable AI: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems, 4, 795–813.

6. Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv preprint arXiv:2007.03051.

7. Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Jiang, L. (2023). LLMCarbon: Modeling the end-to-end carbon footprint of large language models. arXiv preprint arXiv:2309.14393.

8. Li, Z., Zhuang, S., Guo, S., Zhuo, D., Zhang, H., Ng, A. Y., & Stoica, I. (2023). Alpa: Automating inter- and intra-operator parallelism for distributed deep learning. Proceedings of the 17th USENIX Symposium on Operating Systems Design and Implementation, 559–578.

9. Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Wong, R. Y. Y., Chen, Z., Arfeen, D., Abhyankar, R., & Jia, Z. (2022). Towards efficient generative large language model serving: A survey from algorithms to systems. arXiv preprint arXiv:2312.15234.

10. Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A., Smith, N. A., DeCario, N., & Buchanan, W. (2022). Measuring the carbon intensity of AI in cloud instances. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1877–1894.

11. Chen, C., Wang, C., Li, Y., Wan, Z., Geng, M., Xiao, J., ... & Peng, Y. (2026). JuZhou 1.0 Technical Report: The First Edge-Native Text-to-Image Foundation Model Trained Entirely on China-Developed AI Accelerators. arXiv preprint arXiv:2606.28421.

12. Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. Journal of Machine Learning Research, 24(253), 1–15.

13. 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. Journal of Machine Learning Research, 21(248), 1–43.

14. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

15. Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.-J., Anderson, B., Breughe, M., Charlebois, M., Chou, W., Chukka, R., Coleman, C., Davis, S., Deng, P., Diamos, G., Duke, J., Fick, D., Gardner, J. S., Hubara, I., ... Zhou, Y. (2020). MLPerf inference benchmark. Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture, 446–459.