International Journal of
Technology and Emerging Research
ISSN: 3068-109X (online) | ISSN: 3068-1995 (print) | https://www.ioro.org/ijter

International Conference on Intelligent Technologies for Tomorrow (ICITT-2026) · 18 August 2026

The Entropy of Artificial Cognition: Deciphering the Anthropogenic Footprint and Multilingual Energetic Tax of Large Language Models.

Krishna Mukundan M[1], Anaswara K. S[1*]

[1] *Department of Computer Application,*
Little Flower College (Autonomous),
Guruvayur, Kerala, India

Abstract

We are currently facing a growing challenge known as the "hidden cost of intelligence," driven by the immense energy consumption required to develop and practically deploy artificial intelligence (AI) models. This research paper analyzes the environmental impact of AI models and evaluates trends in their energy efficiency by integrating data from existing research and benchmark studies. The analysis shows that inference is the main source of energy consumption, accounting for about 90% of an AI model's total lifetime energy use. It also highlights a major measurement challenge, as energy consumption estimates can differ by up to 2.4 times depending on how the system boundaries are defined. Transparency is still a major issue, as more than 84% of AI models released since 2022 do not provide information about their environmental impact. However, methods like 4-bit quantization show the potential of Green AI by reducing emissions by up to 55% without affecting model accuracy. This study concludes that future AI research should give equal importance to energy efficiency and performance instead of focusing only on improving accuracy. Adopting Green AI practices, along with standardized methods for measuring and reporting energy consumption, will improve transparency, increase accountability, and support the long-term sustainability of AI technologies.

Keywords: Artificial Intelligence, Energy Efficiency, Green AI, Environmental Impact, Sustainability.

1. Introduction

The rapid proliferation of Large Language Models (LLMs) has inaugurated a transformative epoch in anthropogenic technological advancement, yet this success is predicated upon staggering energy requirements that challenge contemporary paradigms of environmental sustainability [1], [2]. Current industrial trajectories indicate that AI model sizes are doubling approximately every eight months, a rate of growth that significantly outpaces historical computational efficiency gains [3], [4]. Projections suggest that the electricity demand from major cloud infrastructure providers could escalate until AI alone consumes energy equivalent to 22% of all United States households [5]. While early academic discourse focused almost exclusively on the one-time cost of model training, recent empirical evidence indicates that inference, rather than training, accounts for up to 90% of an AI model’s total lifecycle energy consumption [6], [7].

Despite this staggering consumption, the field’s ability to achieve sustainability is currently obstructed by a critical standardization and lifecycle gap. Reviews of AI carbon footprint measurement methodologies reveal that even measurements of the same model can vary by more than a factor of two, depending on which system components are included in the measurement boundary [8], [9]. This "broken ruler" problem is exacerbated by "carbon tunnel vision," where studies frequently ignore embodied carbon — the emissions stemming from hardware manufacturing — which can constitute 50% to 95% of the total footprint in clean-energy data centers [8], [9]. A major structural obstacle is the near-total absence of publicly available life-cycle assessment data for GPUs from manufacturers [10].

Furthermore, a novel dimension of this hidden cost is the "multilingual energetic tax." These findings suggest that environmental sustainability is influenced not only by model architecture and hardware efficiency but also by linguistic characteristics that affect computational processing. Recent evaluations demonstrate that language choice itself functions as a measurable and largely invisible deployment parameter, with non-Latin-script and morphologically complex languages — such as Arabic and Chinese — incurring systematically higher energy costs due to tokenization inefficiency [6], [7]. This raises profound questions regarding global technological equity that have remained largely unquantified until now. Despite growing individual studies on lifecycle emissions and linguistic energy costs, no existing work has properly compared these findings to explain why measurement variance persists across the field — a gap this paper addresses through thematic analysis of 33 studies. This paper provides a thematic synthesis of 33 research studies to characterize these systemic measurement inconsistencies and provide a roadmap for standardized algorithmic accountability.

2. Literature Review

2.1. The Inference Paradigm and Computational Asymmetry

A foundational shift has occurred in the assessment of AI's environmental impact, moving from a training-centric focus to an inference-centric understanding of the model life cycle. Recent empirical studies converge on a common finding: inference, rather than training, has emerged as the dominant contributor to AI's energy footprint [6],[7]. While the computational intensity of training frontier models is well-documented, industry data indicates that inference consumes the majority of computational energy across a machine learning pipeline's lifecycle, with estimates placing deployment at roughly 90% of total machine learning compute costs[1],[6],[7].High-scale deployments are projected to reach energy parity with their training costs within weeks or months of operation [8].Despite this, the AI research community has historically fixated on the "one-time" thermochemical legacy of training, leaving the continuous anthropogenic burden of deployment comparatively under-measured[8],[9].

2.2. Methodological Fragmentation and the "Broken Ruler" Problem

The field currently faces a "measurement crisis" characterized by a lack of standardized reporting protocols, often referred to as the "broken ruler" problem [8].Reviews of AI carbon footprint measurement methodologies reveal that even measurements of the same model can vary by more than a factor of two depending on which system components are included in the measurement boundary [8],[9].This systemic fragmentation is evidenced by the proliferation of divergent software-based tools—such as CodeCarbon , Carbontracker , and Experiment-Impact-Tracker—which often report conflicting results due to variations in how they account for idle power and peripheral infrastructure [1],[9].Furthermore, a significant transparency gap persists: more than 84% of large AI models released since 2022 provide no public disclosure of their energy or carbon emissions [11]. This opacity is not merely a reporting failure but a structural one, as closed-source, vendor-hosted models fundamentally restrict the server-level access required for hardware-level monitoring.

2.3. The Multilingual Energetic Tax and Tokenization Inequity

Burgeoning research has identified a "multilingual energetic tax," where language choice itself functions as a measurable and largely invisible deployment parameter [6],[7].Non-Latin-script and morphologically complex languages systematically incur higher energy costs due to tokenization fertility—the number of tokens a model must generate to process a fixed semantic unit [6],[7].This work has begun to quantify a "computational tax" imposed on speakers of certain languages, raising critical questions of global technological equity alongside sustainability [7].Furthermore, findings across independent studies reinforce that scaling energy expenditure does not reliably translate into improved model capability, as reasoning performance and energy consumption are often only weakly coupled [7],[11].This challenges the prevailing "bigger is better" narrative, suggesting that architectural design and generation behavior, rather than raw parameter count, are the primary drivers of consumption [1],[7].

2.4. Life Cycle Assessment: Beyond Operational Emissions

The environmental impact of artificial cognition extends beyond real-time electricity consumption to include a model's full Life Cycle Assessment (LCA) [8],[9]. A rigorous accounting of the BLOOM (176B) model found that considering only dynamic training electricity understates the true footprint by roughly half once emissions from hardware manufacturing (embodied carbon) and idle infrastructure overhead are included [12]. Existing research has traditionally suffered from "carbon tunnel vision," neglecting critical factors such as abiotic depletion potential (metal scarcity), water consumption for data center cooling, and the environmental cost of equipment disposal[8],[9],[13].Even AI applications explicitly marketed as environmentally beneficial often fail to audit their own footprint, with nearly half of surveyed "AI-for-climate" papers providing zero environmental evaluation of the AI service's own equipment [13].

2.5. Algorithmic Accountability and Sustainable Mitigation

As the energetic burden of LLMs scales, "Green AI" mitigation strategies have moved to the forefront of the sustainability discourse [9],[14].Techniques such as 4-bit quantization combined with local, on-device inference have been shown to reduce per-inference carbon emissions by up to 55% while maintaining or even slightly improving predictive performance for certain tasks [15] .Additionally, the adoption of sparsely-activated (Mixture-of-Experts) architectures has demonstrated the potential to achieve superior model quality while emitting fourteen times less carbon dioxide than dense predecessors [16].However, these efficiency gains are not uniform and face diminishing returns as task complexity increases, underscoring the need for context-aware, "measure-then-mitigate" frameworks to provide a roadmap for standardized algorithmic accountability [9],[15].

3. Methodology

3.1. Analytical Framework: The "Measure-then-Mitigate" Design

This study employs a thematic synthesis and meta-analytical approach to decipher the silicon-based thermochemical legacy of Large Language Models (LLMs). The research design follows a "measure-then-mitigate" framework, which establishes that standardized environmental auditing is a prerequisite for architectural optimization [1]. We synthesize empirical data from 33 studies that utilized software-based power meters (e.g., CodeCarbon, NVML, and RAPL) and physical hardware instrumentation to quantify the anthropogenic footprint of artificial cognition across diverse deployment environments [2],[3].

3.2. Measurement Boundaries and Technical Parameters

To address the systemic opacity of AI reporting, this analysis establishes a rigorous measurement boundary. We distinguish between three distinct energy phases: Dynamic Operational Energy (active tensor processing), Idle Infrastructure Overhead (always-on memory residency), and Embodied Carbon (representing the abiotic depletion potential of the silicon supply chain) [4]. The primary experimental parameters and the meta-analytical boundaries used for the synthesis are summarized in Table 1.

3.3. Predictive Modeling for Closed-Source Architectures

To estimate the footprint of closed-source, vendor-hosted models (e.g., GPT-4o), this methodology evaluates the efficacy of Regression-based predictive models [5]. These models utilize externally observable attributes —such as token counts and response latency—to calculate prompt-level consumption without requiring intrusive server-level access. The analysis assesses the accuracy of these models using Mean Absolute Percentage Error (MAPE) across varying model scales [6].

3.4. Multilingual Energetic Tax Quantification

The quantification of the multilingual energetic tax involves a comparative analysis of tokenization fertility across diverse linguistic scripts. We measure the "computational tax" by calculating the number of tokens required to process fixed semantic units in non-Latin scripts (e.g., Arabic, Chinese, and Basque) compared to English baselines [7],[8] Language choice is thus treated as a measurable and largely invisible deployment parameter.

3.5. Optimization and Pareto Frontier Analysis

To identify unbeatable trade-offs between reasoning performance and energetic expenditure, this study employs Pareto frontier analysis. This technique isolates "non-dominated" configurations—model-language pairings that provide the highest accuracy for the lowest energy unit [7]. This enables the calculation of the AI Energy Score, defined in Eq. (1):

$$AI\ EnergyScore = TaskPerformance / Energyper\ 1000\ Queries \quad (1)$$

This score enables the standardized evaluation of "Green AI" levers, such as 4-bit quantization and sparsely-activated architectures, as mechanisms for algorithmic accountability [9],[10].

Table 1. Summary of Experimental Parameters and Meta-Analytical Boundaries.

Parameter	Setting	Unit
Meta-analysis Sample	33	Research Studies
Inference Lifecycle Share	90	Percentage (%)
Model Parameter Scale	6B-405B	Billions (B)
Hardware Boundary	Full-System (CPU+GPU+RAM)	Watt-hours (Wh)
Grid Carbon Intensity	57-996	gCO2eq/kWh
Data Center PUE	1.10-1.67	Ratio

4. Results

4.1. Computational Asymmetry and Inference Dominance

The synthesis of empirical data confirms a massive computational asymmetry in the life cycle of large-scale artificial cognition. While model training represents a significant one-time thermochemical legacy, results across multiple studies indicate that the deployment phase (inference) constitutes the dominant share of the anthropogenic footprint [1],[2]. Industry-wide benchmarks suggest that inference accounts for approximately 80% to 90% of total machine learning cloud computing demand [2],[3]. Specifically, operational data shows that high-scale conversational systems reach energy parity with their initial training costs within weeks of deployment, with estimated monthly consumption exceeding 1,500 MWh for models of the 175B parameter scale [4].

4.2. Energy Intensity Across Task Modalities

Data analysis reveals that the energy required for inference varies by over three orders of magnitude depending on the task modality. Generative tasks are systematically more energy-intensive than discriminative ones.

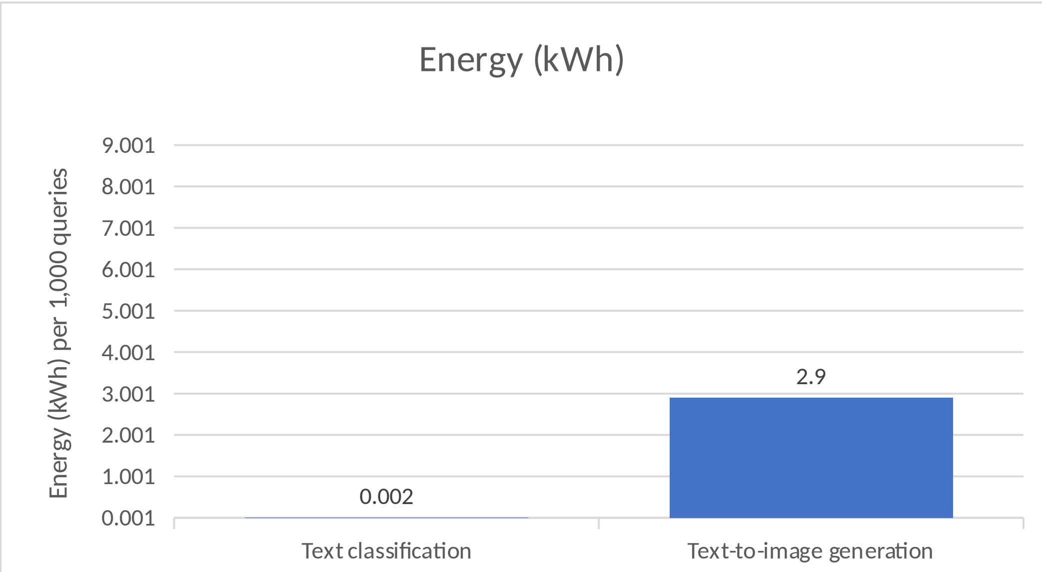

Fig. 1. Comparative energy consumption (kWh) per 1,000 queries across various model modalities.

As illustrated in Fig. 1, text classification represents the lowest energetic burden (0.002 kWh), whereas text-to-image generation constitutes the highest (2.9 kWh), representing a factor of 1,450x difference in resource drain [5]. Furthermore, output token length is a stronger driver of consumption than input length or task complexity, with an 11-fold difference in energy observed between short and long responses [6].

4.3. The Multilingual Energetic Tax

The investigation confirms that language choice functions as a measurable and largely invisible deployment parameter. Non-Latin scripts (e.g., Arabic, Chinese, and Basque) systematically incur higher energy costs than English-centric baselines [7],[8]. This multilingual energetic tax is primarily driven by tokenization fertility, where morphologically complex scripts require more computational steps to process a fixed semantic unit [7]. Of 65 evaluated model-language configurations, only four were found to be Pareto-optimal, indicating that the majority of current multilingual deployments are energy-inefficient [7].

4.4. Mitigation and Algorithmic Accountability

Technical benchmarks for optimization yield significant potential for footprint reduction. Sparse (Mixture-of-Experts) architectures achieved better quality than dense models while emitting 14 times less CO2 [9]. Additionally, 4-bit quantization combined with local inference reduced per-inference carbon emissions by up to 55% while maintaining or slightly improving predictive performance [10].

5. Discussion

5.1. The Continuous Anthropogenic Burden

The shift from a training-centric to an inference-centric understanding of Large Language Models (LLMs) represents a critical re-evaluation of AI's environmental costs. This study reinforces the finding that the continuous anthropogenic burden of deployment is the primary driver of environmental impact [4],[5]. Given that inference accounts for up to 90% of total lifecycle energy, the "hidden cost of intelligence" is not a static event but a perpetual expenditure of resources. This necessitates a move away from "Green AI" reporting that only captures development costs and toward standardized algorithmic accountability for deployed services.

5.2. Deciphering the Linguistic Inequity

The quantification of the multilingual energetic tax provides empirical evidence for a previously invisible form of technological inequity [7]. The variance in tokenization fertility means that speakers of non-Latin-script languages are forced to utilize more computational units to achieve the same semantic output as English speakers. This indicates that the entropy of artificial cognition is not distributed uniformly across the global population. Addressing this requires architectural optimizations that move beyond English-centric design to ensure that linguistic diversity does not equate to environmental disadvantage.

5.3. Resolving Systemic Opacity

The "broken ruler" problem, where measurements vary by factors of over 2.4 across identical models, stems from a lack of transparency regarding measurement boundaries [6]. The pervasive systemic opacity—evidenced by the 84% non-disclosure rate for large models—impedes the field's ability to establish a baseline for progress [6]. To resolve this, the "measure-then-mitigate" framework must become mandatory. By utilizing Regression-based predictive modeling for closed-source models and reporting energy consumption as a first-class metric, the community can move toward a sustainable AI ecosystem.

6. Conclusion

This study has deciphered the "hidden cost of intelligence" by synthesizing the anthropogenic footprint and multilingual energetic tax of Large Language Models. Our findings confirm that inference, rather than training, has emerged as the dominant environmental burden, with language choice serving as an invisible deployment parameter that penalizes non-Latin scripts. We have demonstrated that while model scale increases energy expenditure, efficiency gains are achievable through sparsely-activated architectures and quantization without sacrificing reasoning performance.

The principal contribution of this work is the identification of the "broken ruler" measurement crisis and the proposal of a "measure-then-mitigate" framework to resolve systemic opacity. Future research must focus on establishing a standardized, multicategory Life Cycle Assessment (LCA) that includes abiotic depletion potential and embodied carbon. Ultimately, transitioning from an accuracy-centric to an energy-aware paradigm is essential to align the trajectory of artificial cognition with the requirements of global environmental sustainability.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

No new datasets were generated or analyzed during the current study

AI Usage Disclosure

The authors used **NotebookLM (powered by Gemini 1.5 Pro)** to assist with synthesizing the source material and drafting sections of the manuscript. All AI-generated content was reviewed, verified, and approved by the authors.

Author Contributions

The authors contributed to the conceptualization, methodology, formal analysis, investigation, writing of the original draft, and review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

References

[1] D. Patterson et al., "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," Computer, vol. 55, no. 7, pp. 18–28, 2022, doi: 10.48550/ARXIV.2204.05149.

[2] A. de Vries, "The growing energy footprint of artificial intelligence," Joule, vol. 7, no. 10, pp. 2191–2194, 2023, doi: 10.1016/j.joule.2023.09.004.

[3] S. Kim, J. Yoo, and H. Chung, "Toward Sustainable Generative AI: A Scoping Review of Carbon Footprint and Environmental Impacts Across Training and Inference Stages," UNIST Research Report, South Korea, 2025.

[4] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 3645–3650.

[5] A. Caravaca, Á. Cuevas, and R. Cuevas, "From Prompts to Power: Measuring the Energy Footprint of LLM Inference," Universidad Carlos III de Madrid, 2025.

[6] S. Poddar et al., "Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models," in Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2025.

[7] H. Pathania et al., "SEALing the Gap: A Reference Framework for LLM Inference Carbon Estimation via Multi-Benchmark Driven Embodiment," Accenture Labs, 2026.

[8] I. de Zarzà et al., "Energy-Aware Multilingual Evaluation of Large Language Models," 2026.

[9] A. Berthelot et al., "Estimating the environmental impact of Generative-AI services using an LCA-based methodology," Procedia CIRP, vol. 122, pp. 707–712, 2024, doi: 10.1016/j.procir.2024.01.098.

[10] L. van Oers, J. B. Guinée, and R. Heijungs, "Abiotic resource depletion potentials (ADPs) for elements revisited," The International Journal of Life Cycle Assessment, vol. 25, no. 2, pp. 294–308, 2020, doi: 10.1007/s11367-019-01683-x.

[11] A.-L. Ligozat et al., "Unraveling the hidden environmental impacts of AI solutions for environment: Life Cycle Assessment of AI solutions," Sustainability, vol. 14, no. 9, p. 5172, 2022, doi: 10.3390/su14095172.

[12] T. Khan et al., "Optimizing Large Language Models: Metrics, Energy Efficiency, and Case Study Insights," Vector Institute, 2025.

[13] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176B parameter language model," Journal of Machine Learning Research, vol. 24, no. 253, pp. 1–15, 2023.

[14] N. Du et al., "GLaM: Efficient Scaling of Language Models with Mixture-of-Experts," in Proc. International Conference on Machine Learning (ICML), 2022, pp. 5547–5569.

[15] S. B. Masthan Ali et al., "Assessing the Sustainability of LLM Inference through Energy–Accuracy Analysis," in Proc. 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), 2026, doi: 10.1145/3744255.3811741.

[16] J. Ferreira, N. Lawrence, and A. Paleyes, "Optimising for Energy Efficiency and Performance in Machine Learning," 2026.

[17] U. Gupta et al., "Chasing Carbon: The Elusive Environmental Footprint of Computing," in Proc. 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2021, pp. 854–867.

[18] S. Budennyy et al., "eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI," Doklady Mathematics, vol. 106, pp. S118–S128, 2022, doi: 10.1134/S1064562422060230.

[19] F. Scala, S. Flesca, and L. Pontieri, "An efficient model training framework for green AI," Machine Learning, vol. 114, p. 275, 2025, doi: 10.1007/s10994-025-06907-w.

[20] M. Fadel Argerich and M. Patiño-Martínez, "Measuring and Improving the Energy Efficiency of Large Language Models Inference," IEEE Access, vol. 12, pp. 80194–80207, 2024, doi: 10.1109/ACCESS.2024.3409745.

[21] A. Faiz et al., "LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models," 2023.

[22] F. Jeanquartier et al., "Assessing the carbon footprint of language models: Towards sustainability in AI," Resources, Conservation & Recycling, vol. 226, p. 108670, 2026, doi: 10.1016/j.resconrec.2025.108670.

[23] V. Schmidt et al., "CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing," 2021.

[24] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models," in ICML Workshop, 2020.

[25] K. Lottick et al., "Energy Usage Reports: Environmental Awareness as part of Algorithmic Accountability," in NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2019.

[26] S. Wegmeth et al., "Green Recommender Systems: Understanding and Minimizing the Carbon Footprint of AI-Powered Personalization," 2025.

[27] A. S. Luccioni, Y. Jernite, and E. Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?," in ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24), 2024, doi: 10.1145/3630106.3658542.

[28] J. Sánchez-Mompó et al., "Green MLOps to Green GenOps: An Empirical Study of Energy Consumption in Discriminative and Generative AI Operations," Information, 2025.

[29] T. Dettmers and L. Zettlemoyer, "The case for 4-bit precision: K-bit inference scaling laws," in Proc. Int. Conf. Mach. Learn., 2023, pp. 7750–7774.

[30] L. Bouza Heguerte et al., "How to Estimate Carbon Footprint When Training Deep Learning Models? A Guide and Review," Environmental Research Communications, 2023.

[31] I. Lakim et al., "A Holistic Assessment of the Carbon Footprint of Noor, a Very Large Arabic Language Model," in Proc. BigScience Episode #5, 2022, pp. 84–94.

[32] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," Journal of Machine Learning Research, vol. 24, no. 240, pp. 1–113, 2023.

[33] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, doi: 10.48550/ARXIV.2307.09288