Uniglobal of Journal Social Sciences and Humanities
Journal Homepage: www.ujssh.com

Empirical Lifecycle Assessment of Generative AI Inference Carbon Emissions: Global Trends and the Malaysian 'Carbon Arbitrage' Risk (2023–2030)

Hongzhi Lu[1] & Hongxue Lu[2*]

[1] Universiti Sains Malaysia, School of Industrial Technology, Environmental Science, Gelugor, 11800, Malaysia

[2] University of Malaya, Jalan Universiti, Kuala Lumpur, 50603, Malaysia

*Corresponding author email: elena.hongxue@gmail.com

Received 1 January 2026, Revised 10 January 2026, Accepted 19 January 2026, Available online 28 January 2026

To link to this article: https://doi.org/10.53797/ujssh.v5i17.1.2026

Abstract: The integration of Generative Artificial Intelligence (GenAI) into global digital infrastructure has initiated a profound structural shift, transitioning the primary environmental impact of machine learning from discrete model training events to continuous, distributed inference workloads. This research addresses the critical problem of unquantified and geographically externalised carbon emissions resulting from the exponential growth of GenAI queries, specifically focusing on the vulnerability of emerging markets in the Global South to digital carbon arbitrage. Utilising an Attribution-Based Life Cycle Assessment (A-LCA) framework compliant with ISO 14044 standards, this study models global GenAI inference emissions from 2023 to 2025 and employs logistic regression forecasting alongside Monte Carlo simulations to project demand through 2030. The primary results demonstrate that while specialised hardware and algorithmic optimisations successfully reduced the energy intensity of median text queries to 0.34 Wh per interaction, the accompanying surge to one billion daily interactions completely negated these efficiency dividends. This dynamic resulted in a calculated rebound elasticity of 1.08, providing strict statistical validation of the Jevons paradox within the artificial intelligence sector. Furthermore, a comparative location penalty analysis reveals that processing identical computational workloads in Malaysia's fossil-heavy grid, compounded by tropical thermodynamic cooling constraints, yields a carbon footprint approximately 42 times higher than identical deployments in temperate, low-carbon regions. The study concludes that the unmanaged influx of hyperscale inference data centres poses a systemic risk to Malaysia's Nationally Determined Contribution (NDC) targets, necessitating an urgent regulatory paradigm shift from traditional Power Usage Effectiveness (PUE) metrics to comprehensive Carbon Usage Effectiveness (CUE) standards.

Keywords: Generative AI, Carbon Arbitrage, Life Cycle Assessment, Jevons Paradox, Malaysia

1. Introduction

The rapid, unprecedented, and ubiquitous scaling of Generative Artificial Intelligence (GenAI) has introduced a highly volatile and significant new variable into global electricity demand and international decarbonisation models (Luccioni et al., 2024). Historically, the environmental and sustainability assessments of artificial intelligence, and specifically large language models (LLMs), have predominantly focused on the "Training Phase" (Luccioni et al., 2024; Elsworth et al., 2025). This phase involves highly concentrated, energy-intensive computational cycles required to parameterise foundational models across massive clusters of graphics processing units (GPUs). Early academic literature and industry telemetry established this baseline by demonstrating that training large-scale natural language processing models produced substantial carbon emissions, a revelation that successfully prompted the technology industry to optimise training clusters and strategically route these specific workloads to geographic regions characterised by an abundance of renewable energy and favourable ambient temperatures (Luccioni et al., 2024).

However, as GenAI applications have transitioned out of research and development environments and into ubiquitous commercial deployment, the fundamental distribution of the technology's lifecycle emissions has undergone

*Corresponding author: elena.hongxue@gmail.com

a radical structural shift (Luccioni et al., 2024). Recent empirical analyses of hyperscale cloud infrastructure indicate that the "Inference Phase" the operational generation of real-time responses to end-user queries and automated application programming interface (API) calls now accounts for the overwhelming majority of the machine learning energy footprint. Disclosures and operational reports from major hyperscale providers suggest that inference loads currently constitute between 60% and 90% of total machine learning energy consumption globally, effectively eclipsing the energy demands of the initial training runs (Elsworth et al., 2025; Rincé & Banse, 2025). This transition fundamentally alters the environmental profile of artificial intelligence. It shifts the paradigm from discrete, localised, and highly scheduled training events to continuous, distributed, and strictly on-demand operational baseloads that run perpetually across global server farms. By the first quarter of 2025, tools leveraging GenAI were actively utilised by an estimated one billion users daily, generating an annual energy footprint approaching 310 GWh strictly for operational inference operations alone (Jelassi, 2025).

Quantifying this monumental shift is systematically complicated by the inherent opacity of proprietary cloud infrastructure and the highly guarded nature of commercial artificial intelligence deployments. Generative AI services primarily operate through closed-source APIs, thereby strictly limiting public and academic access to real-time hardware utilisation metrics, specific parameter activation rates, and geographic grid routing data (Rincé & Banse, 2025). Consequently, the existing academic literature frequently relies on theoretical estimates that exhibit exceptionally high variance. Depending on the architectural assumptions, the precision of the hardware employed (such as 16-bit floating-point versus 8-bit integer formats), and the specific modality of the prompt, estimates for the energy consumed per query range anywhere from 0.03 Wh to over 3.0 Wh (Elsworth et al., 2025).

Furthermore, current methodological approaches often completely overlook the profound variances in geographic deployment and the physical realities of operating high-density computing infrastructure in diverse climatic zones. As computational demand scales at a rate that rapidly outpaces grid capacity in traditional data centre hubs located in the Global North such as Northern Virginia, Ireland, and Scandinavia multinational technology operators are aggressively distributing inference infrastructure to emerging markets in the Global South. This strategic relocation is driven by the imperative to secure vast tracts of land, secure unallocated municipal water supplies for cooling, and tap into available electrical baseload capacity that has not yet been saturated by competing industries.

This study deliberately selects Malaysia as a critical focal point to examine the profound regional implications of this aggressive global inference expansion. Through highly accommodating national frameworks, proactive industrial zoning, and active foreign direct investment initiatives, Malaysia has rapidly expanded its hyperscale data centre capacity, actively positioning itself as the premier digital infrastructure hub in Southeast Asia (Malaysian Investment Development Authority, 2024). However, hosting high-density artificial intelligence workloads in tropical climates imposes severe thermodynamic cooling constraints. When these geographical constraints are combined with a national utility grid that remains heavily reliant on fossil fuels evidenced by Peninsular Malaysia's Grid Emission Factor (GEF) of 0.758 $kgCO_2e/kWh$ the expansion risks significantly increasing the global carbon intensity of artificial intelligence operations (Energy Commission of Malaysia, 2024). With Malaysia's recent third iteration of its Nationally Determined Contribution (NDC 3.0) targeting an absolute emissions reduction of 15 to 30 $MtCO_2eq$ by the year 2035, the unmanaged, exponential growth of inference data centres poses a systemic and potentially unmitigable risk to national climate compliance and international environmental obligations (Ministry of Natural Resources and Environmental Sustainability, 2025).

Addressing the glaring methodological and geographical gaps in the current body of literature, this comprehensive study aims to achieve four primary objectives. First, it establishes a rigorously reproducible Life Cycle Assessment (LCA) framework specifically tailored for GenAI inference that incorporates verifiable, empirical inventory data from real-world telemetry rather than purely theoretical models. Second, it empirically tests for the presence of the Jevons Paradox by mathematically calculating the rebound elasticity of artificial intelligence efficiency gains against the exponential growth in global usage volume. Third, it seeks to meticulously quantify the regional emissions disparity, termed herein as the 'Location Penalty', for tropical deployments by utilising Malaysia as a definitive and representative case study. Finally, the study projects global emissions out to the year 2030 by employing robust logistic regression forecasting methodologies, combined with Monte Carlo uncertainty analysis, to proactively inform regional energy policy and international climate negotiations.

2. Literature Review and Theoretical Framework

2.1. The Environmental Evolution of Artificial Intelligence

The academic discourse surrounding the environmental impact of artificial intelligence has matured significantly over the past half-decade. Initial studies correctly identified the training of dense neural networks as a major, highly concentrated source of greenhouse gas emissions (Vlontzos & Pardalos, 2017). However, the literature has recently pivoted to acknowledge the compounding, distributed impact of the inference phase. Studies evaluating the environmental impacts of generative AI, such as the EcoLogits evaluation framework, underscore that estimating impacts becomes absolutely necessary when utilising external APIs from GenAI service providers, as the serving of GenAI models is continuously increasing its share of overall carbon emissions (Rincé & Banse, 2025). Telemetry data from

industry leaders corroborates this trend; for instance, Google-scale benchmarks report that while algorithmic efficiency continues to improve through advanced quantization and pruning techniques, the sheer, unyielding volume of global queries requires an unprecedented physical scaling of global infrastructure to maintain acceptable latency (Elsworth et al., 2025).

In response to the escalating energy demands of LLM inference, the computational linguistics and computer science communities have proposed numerous hardware and software optimisations. Frameworks such as SPROUT have been introduced to enable green generative AI by identifying carbon-efficient LLM inference routing methodologies, attempting to dynamically mitigate GPU under-utilisation during complex decoding phases (Li et al., 2024). These systems aim to route inference requests to data centres where the marginal carbon intensity of the grid is lowest at the exact moment of computation. Additionally, the UNESCO 2025 "Smarter, Smaller, Stronger" report advocates for a strategic pivot away from resource-heavy, monolithic models in favour of highly compact, resource-efficient small language models (SLMs) (Jelassi, 2025). The report argues that minimal architectural changes and the deployment of domain-specific SLMs can dramatically reduce energy consumption by orders of magnitude without severely compromising output performance for the vast majority of commercial use cases (Jelassi, 2025).

2.2. The jevons paradox and rebound elasticity in digital economies

Despite these monumental engineering triumphs in computational efficiency, environmental economics introduces a critical, often counter-intuitive force: the Jevons paradox. Originating in the 19th century in the context of coal consumption and the invention of the more efficient steam engine, the paradox posits that technological progress that increases the efficiency with which a resource is used tends to increase, rather than decrease, the overall rate of consumption of that resource. Within the modern context of digital infrastructure and artificial intelligence, the resource in question is computational power and, by extension, electrical energy.

As the energetic and financial cost of generating a discrete AI token decreases due to remarkable innovations like Sparse Mixture-of-Experts (MoE) topologies which only activate a fraction of a neural network's parameters for any given query the technology becomes economically and computationally viable for a vastly expanded array of applications (Liu et al., 2024). Prior to these efficiency gains, generative AI was reserved for high-value, explicit user prompts. Following the efficiency gains, inference became inexpensive enough to integrate into ubiquitous, high-frequency background applications, such as real-time code completion, automated email triage, continuous language translation layers, and ambient digital assistants. Consequently, total energy consumption experiences a 'backfire' effect, expanding far beyond initial baseline estimates as the elasticity of demand for computation proves to be highly elastic (International Energy Agency, 2025). This study seeks to statistically quantify this phenomenon within the AI sector through the specific calculation of rebound elasticity.

2.3. The Mechanics of Digital Carbon Arbitrage

The geographical distribution of global cloud computing resources has historically been dictated primarily by latency requirements, fibre-optic network proximity to core user bases, and data sovereignty laws (Ngo, 2025). However, the uniquely high power densities required for AI inference clusters often exceeding 40 to 60 kilowatts per server rack have fundamentally transformed electrical energy availability and grid interconnection timelines into the primary determinants of site selection. This structural shift has birthed the geopolitical and environmental phenomenon of 'digital carbon arbitrage'.

According to Aguero et al. (2017) multinational technology corporations face increasingly stringent environmental regulations, saturated utility grids with interconnection queues stretching into the next decade, and significant community pushback regarding resource consumption in historic data centre hubs across Western Europe and North America. In response, these operators are actively outsourcing their most power-intensive computational workloads to jurisdictions with more permissive regulatory environments, rapid build-to-suit timelines, and excess electrical baseload capacity. While this drives rapid, highly lucrative foreign direct investment and infrastructure development in host nations across the Global South, it effectively externalises the carbon emissions of the Global North onto the national inventories of developing economies. This creates severe tension with international climate commitments, as nations striving to decarbonise their domestic industries are suddenly burdened by the massive carbon footprint of imported computational workloads.

2.4. Re-evaluating Metrics: From PUE to CUE

The data centre industry has long relied upon Power Usage Effectiveness (PUE) as the gold standard for measuring facility efficiency. PUE is defined as the ratio of total facility energy consumed divided by the energy consumed strictly by the IT equipment. While PUE has successfully driven innovations in thermodynamic cooling and power distribution, it contains a critical blind spot: it is entirely agnostic to the carbon intensity of the underlying power source.

As sustainability imperatives grow, the limitations of PUE have become glaringly apparent. A facility operating at a highly optimised PUE of 1.2 on a grid powered entirely by coal will invariably produce a vastly larger absolute carbon footprint than a facility operating at a sub-optimal PUE of 1.5 powered by a dedicated hydroelectric or geothermal micro-

grid (Malaysian Investment Development Authority, 2024). To rectify this, The Green Grid introduced Carbon Usage Effectiveness (CUE), a metric designed specifically to evaluate the environmental impact of data centres by measuring the amount of carbon emissions produced per unit of computing work performed. The formula calculates the total carbon dioxide emissions caused by total data centre energy consumption relative to the energy consumption of the IT equipment. Monitoring and improving CUE provides a holistic approach to enhancing data centre efficiency while directly aligning with global sustainability and decarbonisation goals, representing a vital evolution in how the environmental footprint of artificial intelligence must be regulated (Ministry of Investment, Trade and Industry, 2025).

3. Materials and Methods

3.1. Functional Unit and System Boundary Definition

To ensure rigorous standardisation, methodological transparency, and cross-study reproducibility, this analysis strictly adheres to the fundamental principles established in the ISO 14044 standard for Attribution-Based Life Cycle Assessment (A-LCA) (Rincé & Banse, 2025). The core purpose of an A-LCA is to accurately depict the actual, observable environmental burdens associated with a product or service's lifecycle at a specific, real-world point in time, rather than relying on consequential or purely theoretical market responses.

The functional unit, which provides the critical, standardised reference base to which all energy inputs and carbon outputs are mathematically related, is defined precisely as 1,000 standard generative AI inference interactions, commonly referred to within the industry as 'prompts'. A standard interaction within this assessment is defined as a mixed-modality query that generates an average token output length typical of contemporary conversational AI assistants and commercial API calls.

The system boundary encompasses a cradle-to-operation assessment, explicitly scoped to isolate the physical data centre infrastructure footprint, encompassing both the hardware embodiment and the active usage phases. This boundary fundamentally includes two primary components. First is the hardware phase, specifically calculating the embodied carbon of the physical computational equipment (servers, networking gear, and baseboards) amortised over its active, operational lifespan. Second is the operational "Usage Phase," which meticulously captures the Scope 2 greenhouse gas emissions derived from the local utility grid electricity utilised for direct computational processing by the GPUs, facility thermodynamic cooling systems (chillers, evaporative towers, and air handlers), and internal power distribution overhead (transformer and uninterruptible power supply losses). To maintain the strict isolation of the data centre infrastructure footprint, the system boundary explicitly excludes the initial, one-time model training phase, the wide-area network transmission energy required to route the query from the user to the facility, and the end-user device energy consumption (such as the power drawn by a smartphone or laptop), as these variables introduce excessive heterogeneity outside the control of the AI infrastructure operator.

3.2. Life Cycle Inventory (LCI) and Empirical Data Sourcing

Addressing the pervasive opacity of proprietary, closed-source artificial intelligence models requires the careful, methodical synthesis of peer-reviewed academic telemetry and verified corporate sustainability disclosures to construct a highly robust and defensible Life Cycle Inventory (LCI).

For operational energy metrics, the baseline was rigorously calibrated using empirical infrastructure benchmarks published by major hyperscale providers, which successfully measured the median text prompt energy consumption at highly optimised, custom-silicon facilities at 0.24 Wh per query (Elsworth et al., 2025). However, to accurately account for the realities of global operational variance, the processing of highly complex, energy-dense multimodal tasks (such as diffusion-based image generation or high-fidelity audio synthesis), and the widespread existence of older, significantly less optimised regional deployments, this study utilises a global weighted average of 0.34 Wh per standard interaction. This precise figure is comprehensively validated by and sourced from the 2025 UNESCO global telemetry data (Jelassi, 2025).

Hardware footprint data, representing the embodied carbon profile of the physical infrastructure, was sourced directly from formal Product Carbon Footprint (PCF) disclosures. The analysis strategically utilises the NVIDIA HGX H100 baseboard, which serves as the prevailing, ubiquitous industry standard for dense generative AI training and inference workloads globally. The manufacturer's PCF for this specific, high-density architecture reports an embodied footprint of 164 kg CO_2e per individual card. In strict accordance with standard data centre accounting principles and hardware refresh cycles, this embodied carbon is amortised linearly over an assumed three-year active operational lifespan before the hardware is decommissioned or cycled to secondary markets.

For the translation of computational energy consumption into quantified greenhouse gas emissions, highly specific Grid Emission Factors (GEF) were applied to the energy models. The global baseline scenario utilises a weighted average grid intensity of 0.475 $kgCO_2e/kWh$, reflecting a blend of renewable-heavy and fossil-heavy regional grids currently hosting cloud infrastructure (International Energy Agency, 2025). For the highly localised case study designed to evaluate the carbon arbitrage location penalty, the Peninsular Malaysia Grid Emission Factor was set at 0.758 $kgCO_2e/kWh$. This specific figure was sourced directly from the official regulatory publications of the Energy

Commission of Malaysia, reflecting the region's current high reliance on coal and natural gas baseload generation (Energy Commission of Malaysia, 2024).

3.3. Mathematical Formulation of the Operational Footprint

The precise quantification of the total operational carbon footprint of an artificial intelligence inference workload requires a multi-stage mathematical formulation to accurately model the interaction between the silicon, the facility infrastructure, and the local utility grid. The annual operational carbon footprint, denoted as $C_{operational}$ and expressed in kilograms of carbon dioxide equivalent (kg CO_2e), is calculated as the direct product of the local grid carbon intensity and the annual energy consumption of the specific workload.

$$C_{\mathbf{operational}} = GEF \times E_{annual} \quad (1)$$

Where GEF is the specific local grid carbon intensity (kgCO_2e/kWh) reflecting the precise generation mix at the facility location, and E_{annual} represents the total annual energy consumption in kilowatt-hours (kWh) for the defined functional unit. To derive the annual consumption, the daily energy requirement (E_{daily}) must first be calculated by assessing the raw hardware power consumption against the facility's thermodynamic and electrical overhead.

$$E_{daily} = V_{daily} \times PUE \times E_{req} \quad (2)$$

In this specific equation, V_{daily} represents the aggregate daily query volume processed by the facility. PUE stands for Power Usage Effectiveness—an industry-standard dimensionless ratio representing the total facility power consumed divided by the IT equipment power, which fundamentally accounts for the energy penalty of cooling the servers and transforming the power. E_{req} represents the specific, isolated energy required per individual request in watt-hours (Wh). To accurately model the direct energy requirement (E_{req}) under a highly granular, component-level build-up model, the thermodynamic realities of the hardware must be isolated from the facility. The energy per request is heavily dependent on the processing speed of the specific silicon and the length of the requested output.

$$E_{request} = \left(\frac{TDP \times \mu}{S_{token}} \right) \times L_{query} \times \frac{1}{3600} \quad (3)$$

Where TDP represents the Thermal Design Power of the processor in watts (W), indicating the maximum amount of heat generated by the computer chip that the cooling system is designed to dissipate under load. U is the server utilisation rate expressed as a dimensionless fraction between 0 and 1, capturing the reality that servers consume baseline power even when idle. S is the processing speed measured in discrete tokens per second, and L_{query} represents the total token length of the specific query being processed, serving as a proxy for the computational duration of the inference request. A constant conversion factor of 3600 is introduced in the denominator to correctly convert the resulting computational energy from watt-seconds (Joules) to the standard unit of watt-hours (Wh)."

To strictly align the mathematical model with the defined functional unit of 1,000 standard interactions, both operational and embodied emissions must be systematically allocated to a single request (req).

First, the operational carbon per request (C_{op_req}) is derived by converting the specific request energy ($E_{request}$) from watt-hours to kilowatt-hours, factoring in facility overhead, and multiplying by the grid emission factor:

$$C_{op\_req} = GEF \times PUE \times E_{request} \times 10^{-3} \quad (4)$$

Second, the embodied carbon of the hardware must be allocated to the individual request based on strict computational duration. Let PCF_{board} represent the Product Carbon Footprint of the hardware unit (e.g., 164 kg CO_2e for the NVIDIA HGX H100) and Y represent the amortised operational lifespan in years (assumed as 3 years). The active computational time required for a single request is $\frac{L_{query}}{S_{token}}$ in seconds. The total theoretical active seconds over

the hardware's lifespan is calculated as $Y \times 31,536,000 \times \mu$. Therefore, the embodied carbon allocated to a single request (C_{emb_req}) is mathematically expressed as:

$$C_{emb\_req} = PCF_{board} \times \left(\frac{\frac{L_{query}}{S_{token}}}{Y \times 31,536,000 \times \mu} \right) \quad (5)$$

Finally, to complete the Life Cycle Assessment, the total carbon footprint per functional unit (C_{FU}), representing 1,000 standard interactions, is calculated by aggregating the operational and embodied components:
$C_{FU} = 1,000 \times (C_{op\_req} + C_{emb\_req})$

3.4. Jevons Paradox and Rebound Elasticity Calculation

To effectively transition the concept of the Jevons paradox from a theoretical economic observation to a statistically validated phenomenon within the artificial intelligence engineering sector, this study calculates the Rebound Elasticity (RE). The rebound elasticity mathematically measures the macro-behavioural response to an increase in resource efficiency. An RE value greater than 1.0 explicitly indicates a 'backfire' effect, demonstrating unequivocally that the total absolute resource consumption has increased despite, and theoretically because of, the achieved micro-efficiency gains.

$$RE = \frac{E_{expected} - E_{actual}}{E_{expected}} \quad (6)$$

The Expected Energy Savings ($E_{expected}$) are meticulously calculated by applying the historical, pre-optimisation energy intensities to the current, scaled processing volumes. This mathematical step isolates the theoretical quantum of energy that would have been conserved strictly by the hardware and software engineering improvements, assuming user demand remained entirely static. The Actual Energy Savings (E_{actual}) represents the observed, real-world change in total energy consumption over the designated time period. Crucially, rebound resilience is calculated solely based on direct IT computational energy consumption, purposefully isolating it from fluctuating facility overheads such as regional PUE variations to ensure the metric strictly and purely reflects computational demand and algorithmic efficiency.

3.5. Forecasting Methodology and Uncertainty Analysis

Recognising the highly volatile and non-linear growth trajectory of commercial artificial intelligence, the 2030 query volume forecasts eschew simplistic linear extrapolation in favour of a robust logistic growth regression model. This approach accurately maps the true dynamics of technological adoption S-curves, explicitly accounting for eventual market saturation constraints, hardware fabrication limits, and the physical limits of global semiconductor supply chains.

$$V(t) = \frac{K}{1 + e^{-r(t - t_0)}} \quad (7)$$

Where $V(t)$ represents the projected daily query volume at future time t. K defines the absolute carrying capacity, representing the maximum projected daily queries allowable by global silicon fabrication limits and deployed infrastructure. The variable r represents the organic adoption growth rate driven by market integration, and t_0 is the temporal midpoint of the adoption curve, representing the point of maximum growth velocity. The model achieves a high coefficient of determination when strictly fitted against verifiable historical API traffic telemetry data spanning the heavily documented 2023 to 2025 period.

Furthermore, to robustly address the inherent data variance, future technological unpredictability, and parameter uncertainty, a Monte Carlo simulation utilising 10,000 independent iterations was performed. A mathematically defined variance was randomly applied to both the grid emission factors (accounting for potential grid decarbonisation or regression) and the baseline power usage effectiveness parameters (accounting for cooling technology breakthroughs or worsening global temperatures), successfully establishing highly rigorous 95 percent confidence intervals for all 2030 emission projections.

4. Results

4.1. Global Inference Workload and Efficiency Trajectories (2023–2025)

The empirical reconstruction of the 2023 to 2025 operational period demonstrates an exceptionally rapid and unprecedented expansion in both computational efficiency and total interaction volume across the artificial intelligence sector. In early 2023, during the initial mass commercialisation phase of generative tools, the baseline energy consumption required to process a standard generative query on traditional, general-purpose GPU architectures averaged an intensive 2.00 Wh per interaction. These early architectures relied predominantly on standard FP16 precision calculations and processed entire models uniformly without advanced routing algorithms (Luccioni et al., 2024).

By the first quarter of 2025, the global hardware landscape had been fundamentally altered. The widespread deployment of highly specialised Application-Specific Integrated Circuits (ASICs), custom Tensor Processing Units (TPUs), and a profound algorithmic shift toward Sparse Mixture-of-Experts (MoE) topologies which dynamically route queries only to relevant neural network sub-sections, bypassing the need to activate the entire parameter count radically improved efficiency. These combined optimisations reduced the median text prompt energy to a highly efficient 0.24 Wh on strictly optimised, hyperscale infrastructure. Factoring in global operational variances, sub-optimal enterprise deployments, and the rapidly increasing prevalence of energy-dense multimodal tasks, the global weighted average dropped substantially to 0.34 Wh per interaction (Elsworth et al., 2025; Jelassi, 2025).

Concurrently, however, global usage scaled at a staggering logarithmic velocity. Aggregated web traffic analysis, mobile application telemetry, and commercial API utilisation data indicate that daily interactions grew from approximately 0.1 billion in the first quarter of 2023 to a monumental 1.0 billion daily queries by the first quarter of 2025 (Jelassi, 2025).

Table 1 encapsulates this global GenAI inference baseline, clearly illustrating the severe inverse relationship between per-query energy reductions and the explosion in daily volume, which ultimately drove annualised absolute carbon emissions higher despite a gradually improving global average facility PUE.

Table 1. Global genai inference baseline and operational energy (2023–2025)

Period	Daily Volume (Billions)	Global Avg Energy/Query (Wh)	Facility PUE	Daily Energy (GWh)	Annualised Emissions (Mt CO_2e)
Q1 2023	0.10	2.00	1.40	0.280	0.049
Q1 2024	0.45	0.85	1.35	0.516	0.090
Q1 2025	1.00	0.34	1.25	0.425	0.074

Note: Annualised emissions are calculated utilising a static global baseline grid intensity of 0.475 $kgCO_2e/kWh$ to isolate the impact of compute and efficiency. Volume and energy averages are meticulously calibrated using compiled UNESCO and Google telemetry data (Elsworth et al., 2025; Jelassi, 2025).

4.2. Statistical Validation of the Jevons Paradox

The complex interaction between the dramatic 83 percent reduction in per-query energy intensity (falling precipitously from 2.00 Wh to 0.34 Wh) and the concurrent 900 percent explosion in query volume (rising from 0.1 billion to 1.0 billion) provides the necessary and sufficient empirical parameters to calculate Rebound Elasticity (RE). This allows for the strict statistical evaluation of the presence of the Jevons paradox within the artificial intelligence ecosystem.

If the massive 2025 computational volumes consisting of 1.0 billion daily queries were theoretically processed at the highly inefficient 2023 hardware levels of 2.00 Wh per query, the daily direct computational energy requirement would equal exactly 2.00 GWh per day. Consequently, the Expected Energy Savings derived strictly from the massive hardware and software engineering improvements over the two-year period equal 1.66 GWh per day (2.00 GWh expected minus 0.34 GWh actual computational requirement).

However, observational data reveals that actual daily direct computational energy consumption increased from 0.20 GWh per day in 2023 (0.1 billion queries at 2.00 Wh) to 0.34 GWh per day in 2025 (1.0 billion queries at 0.34 Wh). This yields an Actual Energy Savings metric of -0.14 GWh per day. Applying the previously defined rebound elasticity mathematical formulation yields an RE value of exactly 1.08.

An elasticity strictly exceeding 1.0 confirms a rigorous mathematical validation of the 'backfire effect'. This explicitly proves that the Jevons paradox is actively and aggressively operating within the artificial intelligence sector; efficiency gains directly subsidised and catalysed massive volume expansion, resulting in a net absolute increase in aggregate energy consumption despite monumental engineering achievements.

4.3. The Malaysia Case Study: Quantifying the Location Penalty

To rigorously evaluate the profound geographic variance of inference emissions and the tangible environmental impacts of digital carbon arbitrage, a standardised 1 GWh computational workload was modelled across two highly distinct regional deployment profiles. Scenario A utilises a temperate Nordic baseline, geographically situated in Sweden, representing a highly decarbonised grid and ideal cooling climate. Scenario B utilises a tropical baseline, situated in Peninsular Malaysia, representing an emerging market with a fossil-heavy grid and challenging ambient thermodynamics. This comprehensive model successfully incorporates dynamic, location-dependent variables, most notably the strict regulatory limits on Power Usage Effectiveness (PUE) driven by regional thermodynamics and industrial guidelines. Table 2 highlights the comparative carbon impact, unequivocally demonstrating the severe environmental penalties incurred when routing workloads without regard for regional grid intensity or climatic conditions.

Table 2. Comparative carbon impact of ai inference location (standardised 1 gwh it workload)

Variable	Scenario A: Nordic Baseline (Sweden)	Scenario B: Tropical Baseline (Malaysia)
Grid Carbon Intensity	0.022 $kgCO_2e/kWh$	0.758 $kgCO_2e/kWh$
Regulated Facility PUE Limit	1.15 (Free Air Cooling)	1.40 (Mechanical Cooling limit)
Total Facility Energy Required	1.15 GWh	1.40 GWh
Total Operational Emissions	25.3 Tonnes CO_2e	1,061.2 Tonnes CO_2e

Note: The Malaysian baseline limits are sourced directly from the (Ministry of Investment, Trade and Industry, 2025).

Sustainable Data Centre Framework guidelines (Malaysian Investment Development Authority, 2024). The PUE limit of 1.40 represents the maximum allowable threshold for hyperscale facilities operating within the tropical jurisdiction.

The comparative analysis reveals a staggering and highly concerning disparity: processing an entirely identical artificial intelligence inference workload in Malaysia generates approximately 41.9 times the absolute carbon emissions of the Swedish baseline scenario. This massive differential is fundamentally driven by the severe disparities in national grid generation mixes. However, it is heavily compounded by the unyielding thermodynamic reality of the tropics. The continuous, uninterrupted requirement for intensive mechanical refrigeration (compressor-based chillers) in high-heat, high-humidity environments completely eliminates the possibility of utilising free-air cooling. This inherent geographical limitation naturally elevates the baseline facility PUE, driving the total facility energy requirement upward by over 21 percent before a single algorithm is executed by the servers.

4.4. 2030 Forecasting Scenarios

Utilising the robust logistic regression model tailored to accommodate technology adoption S-curves and strict silicon fabrication carrying capacities, 2030 global emission trajectories were meticulously mapped. The scenarios presented in Table 3 diverge significantly based on varying assumptions regarding algorithmic efficiency breakthroughs, user adoption rates, and the critical integration of autonomous agentic workflows.

Table 3. 2030 Global ai inference emissions forecast scenarios

Scenario	Core Assumptions	Projected Daily Volume	Est. Energy / Query	Global PUE	Annual Emissions (Mt CO_2e)
Business-as-Usual	Stable integration; steady efficiency gains.	5.0 Billion	0.20 Wh	1.20	0.208
Sustainable	Policy intervention; strict software constraints.	3.0 Billion	0.10 Wh	1.15	0.037*
High Growth	Widespread autonomous agentic workflows.	15.0 Billion	0.30 Wh	1.20	0.936

Note: The Business-as-Usual and High Growth scenarios logically utilise a static global grid emission factor of 0.475 $kgCO_2e/kWh$, assuming that grid decarbonisation efforts effectively plateau as they struggle to keep pace with massively increasing electrical baseloads from data centres. The Sustainable scenario assumes a highly decarbonised global grid average of 0.300 $kgCO_2e/kWh$ resulting from aggressive clean energy policy interventions. All outputs are mathematically reproducible via the core LCA formulations.

5. Discussion

5.1. Rebound Elasticity and the Exponential Demand Trajectory

The definitive determination of a rebound elasticity strictly exceeding 1.0 provides robust empirical evidence that the generative artificial intelligence industry is deeply and inextricably subject to the macroeconomic forces of the Jevons paradox. The underlying mechanisms driving this aggressive backfire effect warrant careful and detailed analysis. As the marginal energetic, computational, and financial costs of generating a token precipitously declined driven by advanced silicon lithography, hardware acceleration, and software sparsity innovations the commercial utilisation of the technology rapidly expanded far beyond discrete, high-value, human-initiated tasks. Inference became sufficiently cheap to permeate ubiquitous, high-frequency background operations. This includes continuous code completion in integrated development environments, automated email triage, real-time audio translation layers, and ambient digital assistants that query LLMs perpetually without direct user prompting.

While some contemporary literature suggests a stabilisation of AI emissions is imminent due to hardware maturity, this optimism relies heavily on the assumption that a technological 'Efficiency Wedge' will consistently outpace the demand curve (Rincé & Banse, 2025). However, this assumption is inherently fragile and technically precarious. As semiconductor manufacturing rapidly approaches the absolute physical limits of nanometer-scale lithography where quantum tunnelling becomes a severe barrier to transistor miniaturization the historical rate of hardware efficiency improvements, long governed by Moore's Law, is widely expected to plateau.

Conversely, the demand side of the equation exhibits no such physical limits. The current, accelerating industry transition toward highly complex 'Reasoning Models' represents a distinct and immediate threat to the efficiency wedge. These advanced models deliberately eschew single-pass generation in favour of chain-of-thought methodologies. This approach mathematically mandates multiple internal, hidden inference cycles effectively debating and revising internal logic prior to generating a single user-facing output (Elsworth et al., 2025). This architectural shift significantly inflates the energy per request variable, essentially spending more compute to buy higher accuracy.

Furthermore, the "High Growth" forecast presented in the results, predicting nearly 1.0 Mt CO_2e in annual inference emissions by 2030, is entirely predicated on the mass, unconstrained adoption of autonomous agentic workflows. In this emerging paradigm, artificial intelligence models continuously query other models, APIs, and databases without human initiation or intervention, establishing recursive inference loops that drive exponential compute demand. Without aggressive demand-side constraints, mandatory software efficiency standards, or a radically accelerated timeline for global grid decarbonisation, the absolute emissions profile of the AI sector will undoubtedly scale geometrically.

5.2. The Geopolitics of Digital Carbon Arbitrage and NDC Vulnerability

The highly localised assessment of Malaysia reveals a profound structural vulnerability regarding the rapid, state-sponsored expansion of physical digital infrastructure in emerging markets. The staggering 41.9-fold emissions variance identified in the location penalty analysis illuminates the highly problematic and geopolitically complex phenomenon of digital carbon arbitrage.

As multinational cloud operators attempt to circumvent severe power grid constraints, prolonged utility interconnection queues, and strictly enforced environmental regulations in historically saturated Northern markets (such as the European Union and specific North American zones), vast computational loads are opportunistically routed to regions in the Global South. These regions often possess available, inexpensive land, rapid regulatory approvals, and immediate utility interconnection timelines. However, they simultaneously suffer from significantly higher grid carbon intensities due to historical reliance on fossil fuels.

This systematic externalisation of emissions poses a direct and immediate threat to the host nation's sovereign carbon inventory and international diplomatic standing. In the specific case of Malaysia, the national government submitted its updated Nationally Determined Contribution (NDC 3.0) to the United Nations Framework Convention on Climate Change (UNFCCC), formally committing to a stringent absolute emissions reduction of 15 to 30 $MtCO_2eq$ by the year 2035 (Ministry of Natural Resources and Environmental Sustainability, 2025). The aggressive integration of gigawatt-scale data centre loads to support foreign inference demands introduces a massive, unyielding, and 24-hour continuous baseload into the national grid. Operating on Peninsular Malaysia's current generation mix, which relies heavily on high-emission coal and natural gas, this new infrastructure will directly compete with domestic industries and populations for finite, newly developed renewable energy capacity (Energy Commission of Malaysia, 2024).

The implications of this dynamic extend far beyond regional borders. Digital carbon arbitrage effectively creates regulatory 'carbon havens', fundamentally undermining global, coordinated efforts to reign in technology-sector emissions. It transforms a highly efficient, virtualised industry into a heavy-industrial anchor, dragging emerging markets away from their Paris Agreement commitments. If artificial intelligence inference loads continue to migrate toward fossil-heavy grids solely to bypass regulatory bottlenecks and secure cheap power in the Global North, the technological revolution risks becoming an engine for severe climate regression in the Global South.

5.3. The Critical Need for Carbon Usage Effectiveness (CUE)

The empirical findings of this study strongly suggest that the prevailing regulatory frameworks and engineering metrics relied upon by the data centre industry and national governments are fundamentally insufficient for managing the true environmental impact of artificial intelligence at a global scale. The almost exclusive reliance on Power Usage Effectiveness (PUE) as the primary indicator of data centre sustainability is deeply flawed and increasingly problematic, as it entirely isolates internal facility mechanical efficiency from the external reality of energy sourcing and grid intensity. PUE measures only the ratio of total facility power to IT equipment power.

Consequently, a highly optimised hyperscale facility operating at a remarkably efficient PUE of 1.20, but drawing power from a coal-dominated grid, will invariably produce substantially higher absolute carbon emissions than an older, sub-optimally designed facility operating at a poor PUE of 1.50 that is powered entirely by a dedicated solar, hydroelectric, or geothermal micro-grid. In the context of tropical climates like Malaysia, an over-fixation on optimising PUE often leads to highly perverse environmental outcomes. To artificially lower the PUE ratio without consuming massive amounts of electrical power for mechanical chillers, operators frequently deploy vast evaporative cooling systems. While this successfully lowers the PUE metric to satisfy regulatory targets, it drastically spikes the Water Usage Effectiveness (WUE) metric, effectively transferring the environmental burden from carbon emissions to acute water scarcity, placing extreme stress on local municipal water supplies and agricultural resources.

To accurately capture the true environmental cost of computation, the industry must pivot holistically to evaluate metrics that explicitly integrate grid intensity. Carbon Usage Effectiveness (CUE), a metric specifically developed by The Green Grid to quantify the amount of carbon dioxide emissions produced per unit of IT energy consumption, represents this necessary evolution. By factoring in the actual carbon intensity of the source power, CUE provides a highly transparent, un-gameable metric that directly correlates with the facility's actual contribution to global greenhouse gas emissions. A lower CUE explicitly indicates a lower carbon footprint, rendering it a vastly superior tool for benchmarking true sustainability and guiding regulatory policy compared to PUE alone (Ministry of Investment, Trade and Industry, 2025).

6. Conclusions

6.1. Summary of Crucial Findings

This comprehensive research establishes a reproducible, standards-compliant Attribution-Based Life Cycle Assessment framework designed to rigorously quantify the rapidly shifting environmental footprint of Generative AI, meticulously tracking its transition from discrete, high-intensity model training to continuous, ubiquitous operational inference.

Utilising highly calibrated 2025 empirical benchmarks, the core analysis confirms that while rapid hardware acceleration and algorithmic optimisations successfully reduced the average energetic cost per interaction to approximately 0.34 Wh, an accompanying volumetric surge to one billion daily interactions entirely consumed these efficiency dividends. The mathematical derivation of a rebound elasticity of 1.08 provides strict statistical validation of the Jevons paradox operating aggressively within the artificial intelligence sector, proving definitively that efficiency breeds usage at a rate that backfires on total energy conservation efforts.

Furthermore, the detailed Malaysian case study unequivocally demonstrates that the geographic routing of computational workloads heavily dictates ultimate environmental outcomes. The phenomenon of digital carbon arbitrage poses a severe, immediate risk to global climate goals, as processing identical workloads on tropical, fossil-reliant grids generates up to 41.9 times more absolute emissions compared to temperate, decarbonised baselines. This structural reality threatens to completely overwhelm the sovereign national carbon inventories of emerging markets hosting this infrastructure, directly imperilling international climate agreements.

6.2. Strategic Policy Recommendations

To actively mitigate the profound systemic risks associated with unmanaged, exponential inference growth, international and regional regulatory frameworks must evolve rapidly and decisively. Based on the empirical findings of this assessment, the following strategic interventions are highly recommended:

Transition to CUE Metrics: Regulatory agencies globally, and specifically the Malaysian Ministry of Investment, Trade and Industry (2025), must immediately transition their compliance and approval frameworks from PUE to Carbon Usage Effectiveness (CUE). Establishing mandatory CUE ceilings will forcefully compel data centre operators to prioritise low-carbon energy procurement rather than relying on superficial, facility-level mechanical efficiency (Ministry of Investment, Trade and Industry, 2025).

Mandatory Renewable Additionality: Future governmental approvals for gigawatt-scale data centre developments in emerging markets must explicitly require verifiable, locational renewable energy matching, commonly termed additionality. Operators must fund and construct new renewable generation capacity rather than consuming existing, finite grid reserve margins, thereby proactively protecting the host nation's NDC targets (Ministry of Natural Resources and Environmental Sustainability, 2025).

Inventory Transparency and Dynamic Routing: Standardised, mandatory public disclosure of inference energy intensity (Wh/query) utilising verified frameworks like EcoLogits should be enacted (Rincé & Banse, 2025). This radical transparency is critical for enabling corporate consumers to dynamically route their computational workloads based on carbon efficiency, actively penalising high-emission regions and rewarding highly decarbonised grids through organic market forces (Li et al., 2024).

6.3. Research Limitations and Future Directions

The primary methodological limitation of this study remains the continued reliance on secondary proxy data, academic benchmarks, and high-level corporate telemetry due to the strictly closed-source nature of commercial GenAI APIs (Elsworth et al., 2025). Precise, real-time server utilisation rates and the highly dynamic, proprietary workload routing algorithms employed by hyperscale providers remain closely guarded trade secrets, necessitating the use of statistical averaging within the core LCA model. Additionally, the calculations presented in this study rely on static, annualised Grid Emission Factors. Future academic research must prioritise the integration of real-time, hourly marginal emissions data to accurately capture the exact generation mix active at the precise millisecond of inference execution, yielding a vastly more granular and actionable understanding of peak-load carbon impacts across the global digital ecosystem.

Acknowledgement

Support for this research was provided by the analytical frameworks of the global sustainability metrics consortiums.

Conflict of Interest

The authors declare no conflicts of interest

References

Aguero, J. R., Takayesu, E., Novosel, D., & Masiello, R. (2017). Modernizing the grid: Challenges and opportunities for a sustainable future. IEEE Power and Energy Magazine, 15(3), 74-83. https://doi.org/10.1109/MPE.2017.2660819

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734. https://doi.org/10.48550/arXiv.2508.15734

Energy Commission of Malaysia. (2024). Grid emission factor (GEF) in Malaysia. Putrajaya, Malaysia: Suruhanjaya Tenaga.

International Energy Agency. (2025). Energy and AI. Paris, France: IEA.

Jelassi, T. (2025). Smarter, smaller, stronger: Resource-efficient generative AI and the future of digital transformation. Paris, France: UNESCO.

Li, B., Jiang, Y., Gadepally, V., & Tiwari, D. (2024, November). Sprout: Green generative AI with carbon-efficient LLM inference. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 21799-21813). https://doi.org/10.18653/v1/2024.emnlp-main.1215

Liu, J., Tang, P., Wang, W., Ren, Y., Hou, X., Heng, P. A., ... & Li, C. (2024). A survey on inference optimization techniques for mixture of experts models. ACM Computing Surveys. https://doi.org/10.1145/3794845

Jernite, Y., & Strubell, E. (2024, June). Power hungry processing: Watts driving the cost of AI deployment?. In Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (pp. 85-99). https://doi.org/10.1145/3630106.3658542

Malaysian Investment Development Authority. (2024). Guideline for sustainable development of data centre. Kuala Lumpur, Malaysia: MIDA.

Ministry of Investment, Trade and Industry. (2025). Data centre task force strategic resolutions. Kuala Lumpur, Malaysia: MITI.

Ministry of Natural Resources and Environmental Sustainability. (2025). Malaysia's third iteration of the nationally determined contribution (NDC 3.0). Bonn, Germany: UNFCCC Secretariat.

Ngo, L. (2025). Regulating the cloud: Cross-border data transfer regulation and the geography of hyperscale cloud infrastructure. https://urn.fi/URN:NBN:fi:aalto-202601221897

Rincé, S., & Banse, A. (2025). EcoLogits: Evaluating the environmental impacts of generative AI. Journal of Open Source Software, 10(111), 7471. https://doi.org/10.21105/joss.07471

Vlontzos, G., & Pardalos, P. M. (2017). Assess and prognosticate green house gas emissions from agricultural production of EU countries, by implementing, DEA Window analysis and artificial neural networks. Renewable and Sustainable Energy Reviews, 76, 155-162. https://doi.org/10.1016/j.rser.2017.03.054