CCWise: Carbon–Cost Aware Regional LLM Orchestration for Next-Gen Sustainable AI

Ratul Kishore Saha
TCS Research
Kolkata, India

Dheeraj Chahal
TCS Research
Pune, India

Rekha Singhal
TCS Research
New York, USA

Manoj Nambiar
TCS Research
Mumbai, India

Abstract

This paper presents a comprehensive orchestration for evaluating the sustainability of Large Language Models (LLMs) lifecycle by integrating carbon emissions, energy consumption, and cost-efficiency metrics across diverse geographic regions. We introduce two novel indices—Carbon-Cost Tradeoff Index (CCTI) and Green Cost Efficiency (GCE)—to quantify the environmental and economic trade-offs inherent in token generation of LLM deployment. Through extensive experimental analysis, including Pareto assessment of cost versus carbon footprint, we reveal the substantial impact of regional grid carbon intensity and model architecture on operational sustainability. Our findings highlight that smaller, region-optimized models consistently achieve superior carbon-cost performance, whereas deployments in carbon-intensive grids exhibit pronounced inefficiencies.

1 Introduction

Over the past decade, LLMs such as ChatGPT, LLaMA, and Mistral have achieved remarkable successes across a wide spectrum of natural language tasks—including text generation, comprehension, and dialogue systems. However, the deployment of recent LLMs poses profound sustainability challenges: their high parameter counts and substantial inference demands on GPU-equipped servers incur significant energy consumption and carbon emissions. While AI-powered services hold substantial promise—in one study, software development tools are projected to contribute over US$1.5 trillion to global GDP by 2030 (1)—such gains must be balanced against the environmental costs. For instance, even the frequent use of a relatively small model like CodeBERT—invoked thousands of times per day—may consume around 0.32 kWh, approaching the energy capacity of a typical consumer-grade laptop battery at approximately 70 Wh (3; 4). A laptop can only sustain CodeBERT for about 0.22 hours, insufficient for a typical workday, limiting developer mobility and flexibility. The 0.32 kWh energy use corresponds to roughly 0.14 kg CO2, comparable to driving 0.6 miles.

Therefore, this environmental concern is increasingly recognized in the research community under the banner of *Green AI*. Recent studies have begun quantifying both training- and inference-related emissions for LLMs—ranging from lifecycle assessments such as BLOOM's estimated 50 tCO_2 footprint (2) to simulation-based energy-and-carbon frameworks that evaluate inference under real-world GPU utilization and regional carbon grid intensities (6; 5). Innovative approaches such as SPROUT (7) have demonstrated over 40% reduction in inference-related carbon footprint using generation-directive strategies. Several simulation framework to quantify and optimize LLM inference energy use and carbon emissions under diverse deployment scenarios are illustrated in (5; 6). Shi et al. (8) introduces Avatar, a multi-objective optimization framework that compresses large code language models to 3MB to minimize energy consumption($184\times$) and carbon emissions ($157\times$) while preserving performance. Furthermore, each ChatGPT inference consumes approximately $10\times$ more energy, and LLM-generated code can far exceed the energy consumption of human-written code,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

as shown in (11) and (12), respectively. However, several studies on the energy, carbon emission are illustrated in (13; 9; 10).

Despite these advances, a key gap remains: the trade-off between *deployment cost* and *carbon footprint/emission* across geographic regions remains largely unexplored in the context of LLM orchestration and deployment where large number query to be executed. To address this, we make three primary contributions:

- **Benchmarking multi-region environmental impact:** We systematically quantify inference-related energy consumption and emissions of open-source LLMs across diverse geographic settings, accounting for infrastructure and energy grid carbon intensity.
- **Cost–carbon trade-off analysis:** We investigate the relationship between deployment overhead and carbon impact, examining how financial and ecological metrics diverge under different deployment strategies in LLM lifecycle and orchestration.
- **Novel cost–carbon efficiency metric:** We introduce a composite metric that jointly evaluates economic cost and carbon footprint across regions, serving as a decision-support tool for environmentally responsible LLM deployment.

2 Methodology

In modern deployments, LLMs are predominantly hosted on commercial cloud platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. These platforms offer identical hardware configurations across multiple geographic zones; however, both the operational cost of virtual machines (VMs) and the carbon intensity (C.I.) of the local electricity grid vary significantly between regions. This heterogeneity introduces substantial disparities in the economic and environmental impact of LLM inference. Sustainable deployment of LLMs, therefore, requires careful selection of deployment regions and hardware configurations to balance three competing objectives: monetary cost, throughput, and carbon emissions. Notably, the carbon emissions associated with LLM inference are primarily driven by the number of tokens generated, as each token requires a significant number of floating-point operations (FLOPs), leading to considerable energy consumption. We model the carbon emissions of LLM inference as a function of token generation rate. Let R denote the token generation rate (tokens/seconds(s)), P the average power consumption of the system in kilowatts (kW) on the target hardware while executing the LLM inference, and $C.I_{\text{region}}$ the regional carbon intensity in $\mathrm{gCO_2/kWh}$. The total carbon emissions (CE) for generating N tokens over a time period T seconds can be expressed as:

$$CE(\mathrm{gCO_2}) = \frac{P \times T}{3600} \times C.I_{\text{region}} \tag{1}$$

Since the inference time is $T = \frac{N}{R}$, the equation becomes:

$$CE(N) = \frac{P \times N}{3600 \times R} \times C.I_{\text{region}} \tag{2}$$

This formulation highlights that higher throughput (tokens/s) directly reduces the carbon footprint per token, whereas lower throughput results in increased emissions for the same workload.

Similarly, cloud costs are billed on an hourly basis. Let C_{region} denote the hourly VM cost in USD. The cost of generating N tokens is given by:

$$\mathrm{Cost}(N) = \frac{N \times C_{\text{region}}}{R \times 3600} \tag{3}$$

Based on these relationships, we propose two novel metrics for sustainable LLM deployment:

1. Carbon–Cost Tradeoff Index (CCTI)

$$\mathrm{CCTI} = \frac{CE(N)}{\mathrm{Cost}(N)} = \frac{P \times C.I_{\text{region}}}{C_{\text{region}}} \tag{4}$$

CCTI measures the grams of CO_2 emitted per dollar of cloud expenditure, providing a region-aware decarbonization efficiency indicator. A lower CCTI indicates that each dollar spent achieves lower emissions, making that region more environmentally efficient.

2. Green Cost Efficiency (GCE)

$$\mathrm{GCE} = CE(N) \times \mathrm{Cost}(N) = \frac{P \times C.I_{\text{region}} \times C_{\text{region}}}{3600^2} \left(\frac{N}{R}\right)^2 \tag{5}$$

GCE is a composite metric that combines both carbon emissions and monetary cost for a given workload. Lower GCE values correspond to more cost- and carbon-efficient deployments, while higher values penalize configurations that are both energy-intensive and economically inefficient. Together, these metrics enable geo-aware, cost-conscious, and environmentally sustainable LLM deployment planning, providing a quantitative basis for optimizing inference across regions and hardware types.

3 Results and Discussion

3.1 Dataset and LLM Selection

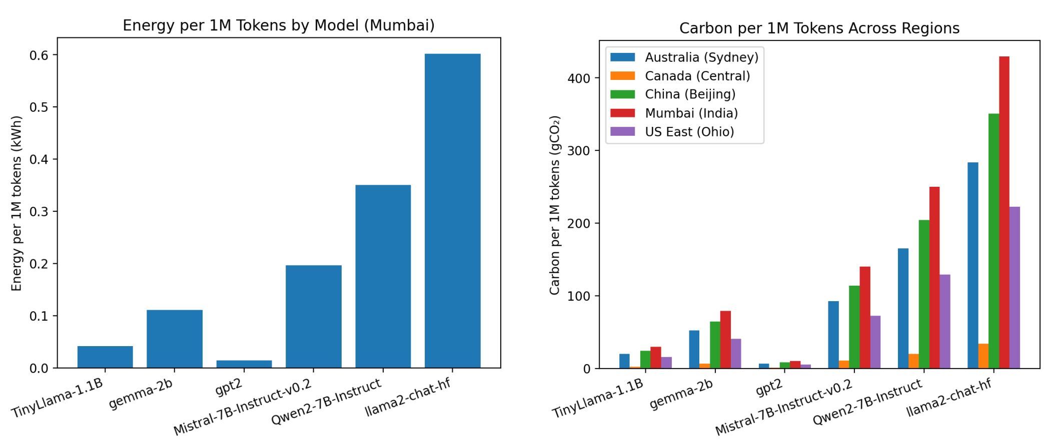
To evaluate the proposed sustainability aspects of LLMs, we conducted an extensive analysis assessing their performance across multiple geographic regions. The selected models include TinyLLaMA-1.1B(15), LLaMA2-7B-Chat-HF(14), Gemma-2B(17), GPT-2(16), Qwen2-7B(18), and Mistral-7B-Instruct-v0.2 (19), deployed in regions such as Mumbai, China, US East, Australia, and Canada. The experiments focused on text generation tasks using the Wikitext dataset (21). All experiments were executed on an NVIDIA A10G GPU hosted on AWS, with deployment configurations varying by region. Table 2 presents a qualitative overview of the deployment cost and associated C.I factors for each location. The results clearly indicate significant heterogeneity across geographic regions in both cost and carbon intensity.

Table 1: LLM Inference Metrics: Tokens Generated, Total Time, Throughput, and Energy Consumption

Model	Tokens	Time (s)	Tokens/s	Energy (kWh)
TinyLlama-1.1B	12,973	9.06	1,432.20	0.000541
LLaMA2-Chat-HF-7B	7,158	62.60	114.35	0.004311
gemma-2b	15,407	26.16	588.89	0.001705
GPT-2	16,976	4.55	3,728.15	0.000238
Qwen2-7B-Instruct	7,557	38.84	194.59	0.002648
Mistral-7B-Instruct-v0.2	14,428	40.42	356.96	0.002828

Table 2: Cost of deployment in AWS (22) and Carbon Intensity (20)by Location (gCO2/kWh)

Location	USD/Hr	C.I(gCO2/kWh)
Mumbai (India)	1.208	713.441
China (Beijing)	9.514	582.317
US East (Ohio)	1.006	369.473
Australia (Sydney)	1.308	470.783
Canada (Central)	1.117	56.039



(a) (b)

Figure 1: (a) Energy per 1M tokens of LLMs, (b) Carbon emission of LLMs various geo-regions

3.2 Multi Geo-Regional Sustainable Analysis of LLMs Deployment

The presented analyses collectively illuminate the intricate trade-offs between computational cost, energy consumption, and carbon emissions across geographic regions and language model scales, providing actionable insights for sustainable AI deployment. Table 1 illustrates the quantitative performance of LLMs on the considered GPU. The carbon intensity Figure 1(b) reveals stark regional disparities, with Mumbai and Beijing exhibiting the highest emissions per million tokens, while Canada (Central) and US East (Ohio) demonstrate significantly lower carbon footprints due to cleaner energy mixes. The energy consumption Figure 1(a) further highlights that energy use scales non-linearly with model size, where larger models (e.g.,LLaMA2-Chat-HF) consume up to 0.6 kWh per million tokens, amplifying emissions particularly in carbon-intensive grids, whereas smaller models (e.g., GPT-2, TinyLlama-1.1B) exhibit more stable and environmentally resilient performance. The GCE heatmap (Figure 2(a)) indicates that lower values correspond to more sustainable and cost-effective deployments; however, high-capacity models in emission-heavy regions show GCE values that are orders of magnitude worse, especially in China (Beijing) and Mumbai. Similarly, the CCTI heatmap (Figure 2(b)) quantifies the carbon penalty per dollar spent, revealing that deployments in high-emission regions can exceed 140 gCO2/USD, while low-carbon grids achieve values below 15 gCO2/USD.

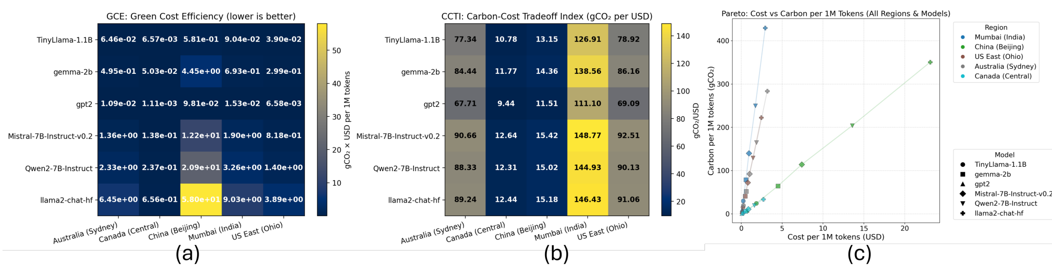


Figure 2: (a), (b) shows the CCTI and GCE heatmap of LLMs across multiple geo-region, (c) shows the Pareto plot of carbon-cost trade-off

3.3 Pareto Analysis: Cost vs Carbon foot print Trade-off of LLMs

The Pareto analysis presented in Figure 2 (c) delineates the trade-off frontier between cost per one million tokens (USD) and carbon emissions per one million tokens (gCO2) across diverse geographic regions and language model architectures. The plot demonstrates that low-cost deployments often coincide with lower carbon emissions, particularly in regions with cleaner energy grids, such as Canada (Central) and US East (Ohio). Conversely, deployments in China (Beijing) and Mumbai (India) exhibit pronounced inefficiencies, occupying the upper-right quadrant with both elevated costs and disproportionately high emissions. Larger-scale models, such as LLaMA2-Chat-HF and Mistral-7B-Instruct-v0.2, display wide variability across regions, underscoring the significance of grid carbon intensity in determining sustainability outcomes. For the LLaMA2-Chat-HF model, the carbon emissions between Mumbai and China exhibit only a marginal difference, whereas a significant disparity exists in deployment cost. This pattern is consistent across other LLMs as well. Notably, the Pareto-efficient frontier aligns with the lower-left envelope, where models such as TinyLLaMA-1.1B and GPT-2 strike an optimal balance between economic and environmental impact. This underscores a pronounced trade-off between carbon footprint and deployment cost across geographic regions.

4 Conclusion

This study presents a comprehensive analysis of sustainable LLM deployment life cycle across diverse geographic regions. By integrating Pareto-based optimization for both carbon footprint and operational expenditure, the findings demonstrate that strategic model selection and deployment localization can substantially reduce environmental impact without compromising performance. The results highlight a practical framework for guiding industry stakeholders and policymakers toward greener AI operations while ensuring cost efficiency and scalability.

References

[1] Dohmke, Thomas, Marco Iansiti, and Greg Richards. "Sea change in software development: Economic and productivity analysis of the ai-powered developer lifecycle." arXiv preprint arXiv:2306.15033 (2023).

[2] Luccioni, Alexandra Sasha, Sylvain Viguier, and Anne-Laure Ligozat. "Estimating the carbon footprint of bloom, a 176b parameter language model." Journal of machine learning research 24.253 (2023): 1-15.

[3] Hellendoorn, Vincent J., et al. "When code completion fails: A case study on real-world completions." 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). IEEE, 2019.

[4] Karampatsis, Rafael-Michael, et al. "Big code!= big vocabulary: Open-vocabulary models for source code." Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. 2020.

[5] Jegham, Nidhal, et al. "How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference." arXiv preprint arXiv:2505.09598 (2025).

[6] Özcan, Miray, et al. "Quantifying the Energy Consumption and Carbon Emissions of LLM Inference via Simulations." arXiv preprint arXiv:2507.11417 (2025).

[7] Li, Baolin, et al. "Sprout: Green generative AI with carbon-efficient LLM inference." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

[8] Shi, Jieke, et al. "Greening large language models of code." Proceedings of the 46th international conference on software engineering: software engineering in society. 2024.

[9] Shi, Jieke, Zhou Yang, and David Lo. "Efficient and Green Large Language Models for Software Engineering: Literature Review, Vision, and the Road Ahead." ACM Transactions on Software Engineering and Methodology 34.5 (2025): 1-22.

[10] Lacoste, Alexandre, et al. "Quantifying the carbon emissions of machine learning." arXiv preprint arXiv:1910.09700 (2019).

[11] De Vries, Alex. "The growing energy footprint of artificial intelligence." Joule 7.10 (2023): 2191-2194.

[12] Vartziotis, Tina, et al. "Learn to code sustainably: An empirical study on llm-based green code generation." arXiv preprint arXiv:2403.03344 (2024).

[13] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for modern deep learning research." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 09. 2020.

[14] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[15] Zhang, Peiyuan, et al. "Tinyllama: An open-source small language model." arXiv preprint arXiv:2401.02385 (2024).

[16] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

[17] Team, Gemma, et al. "Gemma 2: Improving open language models at a practical size." arXiv preprint arXiv:2408.00118 (2024).

[18] Bai, Shuai, et al. "Qwen2. 5-vl technical report." arXiv preprint arXiv:2502.13923 (2025).

[19] Chaplot, Devendra Singh. "Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed." arXiv preprint arXiv:2310.06825 3 (2023).

[20] https://github.com/mlco2/codecarbon/blob/master/codecarbon/data/privateinfra/globalenergymix.json

[21] https://huggingface.co/datasets/Salesforce/wikitext

[22] https://calculator.aws//createCalculator/ec2-enhancement