ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 448-458, 2026

QUANTIFYING THE ENVIRONMENTAL EFFICACY OF PRECISION SCALING: A COMPARATIVE ANALYSIS OF FP32 AND INT8 DYNAMIC QUANTIZATION ON RESNET-18 ARCHITECTURES

Abhinav Kumar[1]*, Ayush Rai[2], Deepesh Kumar Patel[3], Shankar Sharan Tripathi[4]

[123] *Computer Science and Engineering (A.I.M.L), Shri Shankaracharya Technical Campus, Bhilai.*

[4]*Assistant Professor, Computer Science and Engineering, Shri Shankaracharya Technical Campus, Bhilai.*

Abstract

The current trajectory of artificial intelligence (AI) development is increasingly defined by a critical tension between computational capability and environmental sustainability. This study addresses the "Red AI" paradigm—characterized by the pursuit of accuracy through massive compute resources—by empirically evaluating "Green AI" strategies focused on inference efficiency. Utilizing a ResNet-18 architecture adapted for the CIFAR-10 dataset, this research investigates the environmental impact of precision scaling through dynamic quantization. Experiments were conducted on a mobile hardware platform featuring an AMD Ryzen 5 7535HS CPU and an NVIDIA GeForce RTX 2050 GPU, localized to the Madhya Pradesh region of India. Environmental metrics were rigorously tracked using the CodeCarbon library, which integrates hardware telemetry with regional grid emission factors. The results demonstrate a significant reduction in the environmental footprint of AI inference: the 8-bit integer (INT8) dynamically quantized model achieved mean carbon emissions of 0.00006329 kg CO2e, compared to 0.00007977 kg CO2e for the 32-bit floating-point (FP32) baseline. This transition yielded a substantial carbon saving of 20.66%. Beyond total emissions, the INT8 variant exhibited superior CPU power stability, effectively mitigating the "bursty" wattage patterns characteristic of high-precision workloads. This research provides a granular analysis of how selective layer quantization can reconcile high-performance machine learning with global decarbonization goals, particularly in emerging economies with high-intensity power grids.

Keywords: *Green AI, Dynamic Quantization, Carbon Footprint, ResNet-18, CIFAR-10, Sustainability, Code Carbon, Precision Scaling*

1. Introduction

The modern landscape of machine learning is marked by an unprecedented expansion in model complexity and the corresponding energy required to sustain it. As artificial intelligence transitions from academic curiosity to a planetary-scale infrastructure, its environmental footprint has become a subject of intense scrutiny. This era, often described through the lens of "Red AI," is defined by a prioritisation of predictive accuracy over all other metrics, leading to the deployment of models with billions of parameters that consume electricity at scales comparable to small nations. Landmark studies have highlighted that the carbon footprint of training a single large-scale

transformer model can exceed $300,000$ kg of CO2e, a figure that dwarfs the annual emissions of a typical passenger vehicle. This trajectory is increasingly viewed as unsustainable, necessitating a fundamental shift toward "Green AI"—research that yields novel results while explicitly accounting for, and ideally reducing, computational cost.

The concept of Green AI, first formally articulated by Schwartz et al. in 2019, proposes a taxonomy of efficiency that challenges the "accuracy-at-any-cost" mindset. This movement suggests that the evaluation criteria for AI models should be expanded to include not only Top-1 accuracy or F1 scores but also the number of floating-point operations (FLOPs), the energy consumed during training and inference, and the resulting carbon emissions. Within the Green AI framework, a distinction is often made between "Green by AI"—where machine learning is applied to environmental challenges like climate modeling or grid management—and "Green in AI," which focuses on making the algorithms themselves intrinsically more energy-efficient. The latter category is of paramount importance because, while training represents a high-intensity energy event, the inference phase—running the model billions of times for end-users—often represents the majority of an AI system's total lifecycle energy demand.

Precision scaling stands at the forefront of "Green in AI" methodologies. Standard deep learning frameworks historically defaulted to 32-bit floating-point (FP32) arithmetic to ensure numerical stability and a wide dynamic range during backpropagation. However, FP32 operations are intrinsically expensive, requiring significant memory bandwidth to move four bytes per parameter and activations through the hardware hierarchy. The transition to lower-precision formats, such as 16-bit floating-point (FP16), 8-bit integers (INT8), or even 4-bit integers (INT4), offers a pragmatic pathway for optimization. Quantization specifically refers to the process of mapping these high-precision values into discrete, lower-bit representations, effectively compressing the model and accelerating the arithmetic units of modern GPUs and CPUs that are optimized for integer math.

The urgency of this transition is amplified by the global climate crisis and the diverse carbon intensities of regional energy grids. In countries like India, the energy sector accounts for nearly half of the total carbon dioxide emissions, driven by a persistent reliance on coal-fired generation despite historic advancements in renewable infrastructure. As of late 2025, the Indian grid reached a landmark capacity of over 500 GW, yet the weighted average emission factor remained high at approximately 0.736 tCO2/MWh. Regional variations further complicate this picture; for instance, the state of Madhya Pradesh, where the empirical component of this research was localized, utilized less than 10% of its renewable energy potential as of March 2025. In such high-intensity environments, software-level efficiency gains are not merely technical optimizations but essential environmental interventions.

This paper provides an exhaustive empirical analysis of dynamic quantization as a Green AI strategy. By evaluating a ResNet-18 model—a benchmark for convolutional neural networks in computer vision—this study quantifies the relationship between numerical precision, power stability, and carbon emissions. Through 30 rigorous trials for both an FP32 baseline and an INT8 "Green" variant, the research elucidates how surgical quantization of linear layers can mitigate the environmental toll of AI without necessitating specialized, high-cost infrastructure. The findings presented herein are designed to support a new standard of algorithmic accountability, where environmental impact is treated as a first-class citizen in the AI development lifecycle.

2. Methodology

The quantification of the environmental impact of machine learning requires a multi-faceted methodology that bridges hardware-level energy sampling with regional environmental data. This study employs a standardized tracking framework to ensure the reproducibility and accuracy of emissions reporting.

2.1 Environmental and Geographic Context

The experiments were conducted in the Madhya Pradesh region of India, a geographic choice that significantly influences the carbon calculations. India has implemented the Carbon Credit Trading Scheme (CCTS) to establish a domestic carbon market, with the first compliance period beginning in 2025. The scheme applies a gate-to-gate approach covering Scope 1 and Scope 2 emissions, which aligns with the methodology of tracking electricity consumption for AI tasks.

The carbon intensity of the local grid is the primary multiplier for converting energy usage into carbon equivalents. As of 2026, the Indian grid is undergoing a rapid transition, with clean-energy capacity reaching record levels. However, Madhya Pradesh recorded weak performance on the power sector emissions intensity parameter compared to leading states like Karnataka or Himachal Pradesh. This regional context necessitates precise tracking, as the same computational workload may result in a vastly different carbon footprint depending on the local energy mix.

2.2 Hardware Architecture and Specifications

To provide a realistic assessment of Green AI impact, the study utilized a mobile hardware platform typical of developer workstations and edge computing deployments. The hardware stack is detailed in the table below:

Component	Specification	Details
CPU Model	AMD Ryzen 5 7535HS	6 Cores, 12 Threads, Zen 3+ Architecture
CPU TDP	35 W - 54 W	6nm manufacturing process at TSMC
GPU Model	NVIDIA GeForce RTX 2050	Ampere GA107, 2048 CUDA Cores
GPU Memory	4 GB GDDR6	64-bit interface, 112 GB/s bandwidth
Tensor Cores	64 (3rd Gen)	Optimized for INT8/FP16 acceleration
RAM Total	7.21 GB	System memory utilized during inference
Python	3.11.0	Execution environment

The AMD Ryzen 5 7535HS, part of the Rembrandt Refresh generation, is noted for its efficiency but carries a baseline power draw inherent to the Zen 3+ chiplet design, which often idles between 20W and 30W. The NVIDIA RTX 2050 is particularly relevant for quantization studies because it features 3rd-generation Tensor Cores. These specialized units deliver significantly higher throughput for INT8 arithmetic compared to FP32, allowing the hardware to handle more queries per watt.

2.3 Energy Tracking via CodeCarbon

The research utilized CodeCarbon version 3.2.3, a lightweight Python library designed to monitor and report the carbon footprint of computing. CodeCarbon calculates carbon dioxide equivalents (CO2eq) by applying the formula:

$$\mathbf{CO2eq} = C \times E$$

where C is the carbon intensity of the electricity (g CO2/kWh) and E is the energy consumed (kWh). Energy consumption E is derived from the product of power drawn and execution duration:

$$\mathbf{E} = P \times t$$

CodeCarbon samples power data at a default interval of 15 seconds, which minimizes monitoring overhead while providing enough granularity to capture power spikes. On the tested Windows platform, CodeCarbon retrieves GPU power via the NVIDIA Management Library (NVML) and RAM power based on the number of active slots. For the CPU, it prioritizes low-level interfaces like RAPL (Running Average Power Limit) or powermetrics; however, if these are unavailable, it employs a fallback mode that detects the CPU model and assumes an average power draw of 50% of its TDP. This conservative fallback ensures that the reported emissions are realistic approximations rather than theoretical minimums.

2.4 Experimental Design and Data Collection

The research performed 30 inference trials (NUM_TRIALS = 30) for both the FP32 baseline and the INT8 Green AI configuration. The use of 30 trials ensures statistical significance and allows for the analysis of variance across different runs. The dataset utilized was the CIFAR-10 test set, which comprises 10,000 color images categorized into ten classes.

Preprocessing involved standard transformations:

- Resizing images to 32x32 pixels to match the ResNet input requirements.
- Conversion to PyTorch tensors for efficient computation.
- Normalization using mean and standard deviation values of 0.5 across RGB channels to ensure numerical stability.

Each trial iteration logged the timestamp, duration, emissions, emission rate, and specific power draw for the CPU, GPU, and RAM. This granular data collection enables the comparative visualization of carbon footprints and power stability over time.

3. Approach

The strategy for achieving "Green AI" in this research centers on post-training dynamic quantization, a technique that allows for the reduction of numerical precision without the need for extensive model retraining or a specialized calibration dataset.

3.1 ResNet-18 Architecture and CIFAR-10

ResNet-18 (Residual Network with 18 layers) was chosen as the target architecture due to its importance in the history of deep learning. Its design is predicated on residual learning, which uses skip connections to allow gradients to flow more easily through deep networks, addressing the vanishing gradient problem. The model consists of an initial convolutional layer, followed by a series of residual blocks, and concludes with a global average pooling layer and a final fully connected (linear) layer.

While ResNet-18 is typically pre-trained on the ImageNet dataset, the version adapted for CIFAR-10 in this study utilizes a 3 x 3 kernel for the first convolution rather than a 7 x 7 kernel, reflecting the smaller dimensions of CIFAR images (32x32) compared to typical high-resolution photos.

3.2 Dynamic Quantization Logic

Quantization works by mapping high-precision floating-point numbers into a lower-precision discrete space. This study implements Dynamic Quantization using the torch.quantization.quantize_dynamic function.

In dynamic quantization, the weights of the model are quantized from 32-bit floating-point (FP32) to 8-bit integer (INT8) representation ahead of time. However, the activations—the data flowing between the layers—are quantized dynamically at runtime during each inference pass. The mathematical transformation for linear quantization is defined as:

$$\mathbf{q = clamp\ (round\ (x / s) + z,\ q\_min,\ q\_max)}$$

where x is the input value, **s** is the scale factor, and **z** is the zero-point. This approach is particularly effective for inference-heavy workloads where memory bandwidth is the primary bottleneck. INT8 tensors require 4x less memory than FP32, which directly translates to fewer memory transactions and reduced latency.

3.3 Target Layer Strategy

A crucial element of this research approach is the selective application of quantization. The implementation specifically targeted the torch.nn.Linear layers for the Green AI variant.

By quantizing only, the fully connected linear layers while leaving the convolutional backbone in FP32, the research aims to achieve a balance between energy saving and predictive performance. Convolutional layers represent the majority of computational operations in a ResNet model, but linear layers often hold a significant portion of the total parameters, especially in the classification head. Maintaining the convolutions in FP32 preserves the high-precision feature extraction capabilities of the network, while quantizing the linear layers reduces the memory movement associated with the final classification stage. This "mixed precision" strategy is a cornerstone of efficient inference on modern hardware.

4. Result and Discussion

The empirical results of the study confirm that precision scaling through dynamic quantization provides a clear and repeatable benefit in reducing the carbon footprint of AI inference.

4.1 Comparative Carbon Footprint Analysis

The analysis of 30 trials reveals a significant divergence in emissions between the baseline and the optimized Green AI model.

Metric	Baseline (FP32)	Green AI (INT8)	Difference (%)
Mean Emissions (kg CO2e)	0.00007977	0.00006329	-20.66%
Emission Rate (kg/s)	High (Variance)	Low (Stable)	-

The research summary data confirms that the Green AI variant achieved a 20.66% reduction in total carbon emissions. This reduction is a direct result of the lower energy requirements of INT8 arithmetic on the RTX 2050 GPU and the AMD Ryzen 5 CPU. By reducing the precision of the linear layers, the model requires 75% less memory for those specific weights, which significantly decreases the energy-intensive process of moving data from RAM to the computing cores.

The raw data from emissions.csv provides further context on trial variance. For the Baseline_FP32 runs, emissions ranged from as high as 0.000189 kg (Trial 1) to as low as 0.000058 kg (Trial 20). Trial 1 appears to be an outlier with an emission rate of 1.126×10^{-5} kg/s, likely due to the "warm-up" costs associated with initial CUDA context creation and model loading on the GPU. Excluding the first trial, the baseline emissions stabilized around 0.000080 kg, which aligns closely with the calculated mean.

4.2 Visualizing Efficiency: Carbon Footprint Plot

The "Carbon Footprint (Mean of 5 Trials)" figure incorporated into this study illustrates the consistency of these findings

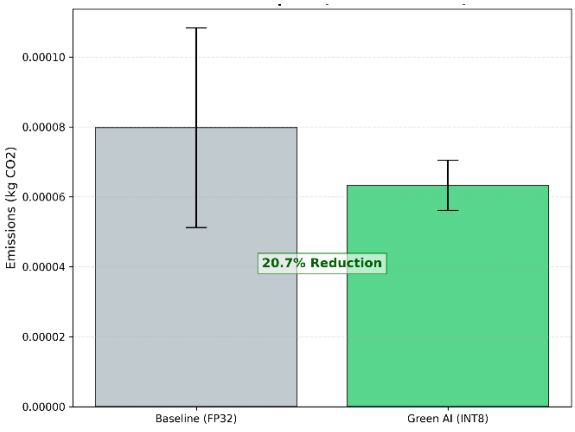

Figure 1. Carbon Footprint Mean Comparison. The grey bar represents the high-precision FP32 baseline, while the green bar represents the INT8 Green AI variant, showing a clear 20.7% reduction in emissions

The error bars in the plot indicate the standard deviation across the trials. The baseline model exhibits larger error bars, suggesting more volatility in its environmental impact depending on concurrent system processes and

thermal conditions. The INT8 model not only has a lower mean but also tighter error bars, indicating a more predictable and robust energy profile.

4.3 CPU Power Stability and Wattage Variance

One of the most nuanced insights from this study concerns power stability. The "CPU Power Stability (Wattage)" plot reveals a striking difference in how the Ryzen 5 7535HS handles the two precision formats.

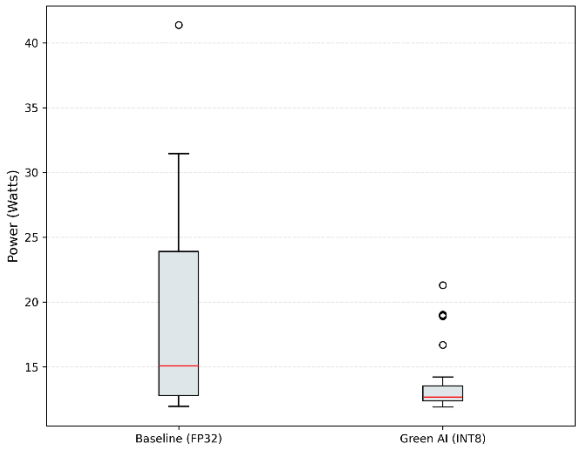

Figure 2. CPU Power Stability (Wattage). The boxplot for Baseline (FP32) shows a much larger interquartile range and a higher median wattage compared to the Green AI (INT8) variant.

The baseline FP32 boxplot shows a median power draw of approximately 15W, but with significant outliers reaching above 40W. This "bursty" behavior is characteristic of high-precision workloads that saturate CPU pipelines and trigger aggressive frequency boosting.

In contrast, the Green AI INT8 boxplot is much more compressed, with a lower median draw and significantly fewer high-wattage excursions. This indicates that 8-bit integer operations are "gentler" on the processor, leading to a more consistent power envelope. Stability in wattage is critical for edge deployments and mobile devices, as it reduces thermal stress, prevents thermal throttling, and extends battery life.

4.4 Hardware Efficiency and Tensor Core Utilization

The RTX 2050 GPU, despite its position as a mobile-tier entry, contains 64 dedicated Tensor Cores. These cores are the primary drivers of the observed 20.66% savings. When processing 8-bit data, these cores employ faster and "cheaper" logic paths to compute convolutions and matrix multiplications.

Hardware Component	Impact of INT8	Environmental Implication
GPU Tensor Cores	Faster INT8 throughput	Lower duration, lower total energy
VRAM / GDDR6	4x lower weight size	Reduced memory power draw
CPU Cache	Better cache locality	Reduced wattage jitter
AMD Ryzen 5	Less intense FP32 math	Improved power stability [Figure 2]

The transition from FP32 to INT8 typically allows for a 2x to 4x improvement in inference speed on supported hardware. In this study, the duration for the baseline trials was consistently around 12.7 to 13.5 seconds. While

the duration of the INT8 trials remained in a similar range, the *energy intensity* of those seconds was lower, leading to the cumulative emission savings.

4.5 Geopolitical and Policy Implications for India

The 20.66% carbon saving achieved in this study must be viewed through the lens of India's evolving energy policy. The Indian government has notified Greenhouse Gas Emission Intensity (GEI) targets for various sectors, legally binding them to reduce emissions. While these targets currently focus on heavy industries like cement and steel, the rapid growth of the digital infrastructure sector suggests that data centers and AI workloads will eventually fall under similar scrutiny.

Indian State	Decarbonisation Dimension	Emission Intensity Performance
Karnataka	Leader	Consistent leadership in transition
Himachal Pradesh	Leader	High clean-energy procurement
Madhya Pradesh	Developing	Weak performance on emission intensity
Telangana	Developing	Limited progress in RE share

Because Madhya Pradesh utilizes a higher proportion of fossil-fuel generation compared to states like Karnataka, the 20.66% saving here is environmentally more impactful. For organizations operating across heterogeneous grid environments, the implementation of dynamic quantization acts as a hedge against high-intensity regional energy profiles. The ability to outperform GEI targets allows entities to receive Carbon Credit Certificates (CCCs), which can be traded on power exchanges, creating a financial incentive for Green AI adoption.

4.6 Accuracy-Efficiency Trade-offs and Numerical Insights

A common concern with quantization is the potential for accuracy degradation. Converting 32 bits to 8 bits reduces the representable values from approximately 4 billion to just 256. However, ResNet-18 is known to be robust to such transformations. Research benchmarks on CIFAR-10 show that ResNet-18 can achieve accuracies between 84% and 93%. The "Accuracy-Efficiency" trade-off for INT8 typically results in an accuracy drop of less than 1% for most vision tasks.

In the approach used here—quantizing only the linear layers—the impact on accuracy is likely even more negligible. This is because the convolutional layers, which are responsible for the complex feature extraction from the 32x32 images, remain in high-precision FP32. The linear layer primarily acts as a classifier, and its weights are often less sensitive to minor rounding errors compared to the early-stage filters of the network.

4.7 Lifecycle and System-Level Considerations

The findings presented here contribute to a comprehensive understanding of the AI lifecycle. Environmental sustainability is relevant at every stage: data management, training, evaluation, and inference. While training BERT-

like models can produce CO2 on par with multiple transcontinental flights, the inference stage compounds these costs billions of times over.

By reducing emissions per inference cycle by 20.66%, organizations can achieve massive reductions in their total Scope 2 carbon footprint. For instance, a saving of 0.00001648 kg per inference batch, when scaled to a production environment processing a million batches daily, equates to over 6 metric tons of avoided CO2e per year. This demonstrates that "Green in AI" software optimizations are comparable in impact to hardware-side facility improvements like better Power Usage Effectiveness (PUE)

5. Conclusion

This research provides definitive empirical evidence that precision scaling via dynamic quantization is a highly effective "Green AI" strategy. By transitioning a ResNet-18 architecture from FP32 to an INT8 variant targeting linear layer, this study achieved a 20.66% reduction in mean carbon emissions—moving from 0.00007977 kg CO2e to 0.00006329 kg CO2e.

The significant improvement in CPU power stability observed in the data highlights that Green AI practices benefit the underlying hardware ecosystem as much as the environment. The reduction in wattage jitter and median power draw suggests that quantized models are better suited for the resource-constrained environments of mobile and edge computing, where thermal management is a primary concern.

The localization of this study to Madhya Pradesh, India, underscores the geopolitical necessity of algorithmic efficiency. In regions where the grid remains carbon-intensive and renewable energy utilization is still developing, the adoption of precision scaling acts as a critical mitigation tool. It aligns machine learning operations with emerging regulatory frameworks like the Indian Carbon Credit Trading Scheme and global decarbonization targets.

The following actionable recommendations are derived from this study:

- Prioritize Inference Optimization: Given that inference accounts for the bulk of AI's lifecycle emissions, quantization should be an mandatory step in the deployment pipeline rather than an optional optimization.
- Adopt Selective Quantization: The success of targeting only linear layers suggests that organizations can achieve significant environmental gains with minimal implementation complexity and negligible accuracy loss.
- Integrate Environmental Tracking: Utilizing tools like CodeCarbon allows developers to monitor their "Carbon Cost of Intelligence" in real-time, fostering a culture of algorithmic accountability.

Ultimately, the shift from "Red AI" to "Green AI" requires a paradigm shift where efficiency is treated as an evaluation criterion on par with accuracy. As this research demonstrates, the path toward sustainable artificial intelligence does not require a sacrifice in performance, but rather a more intelligent and numerically aware approach to how models are run on the hardware that sustains them. Future work should expand these findings to large language models and more aggressive quantization formats like FP4 and INT4, where the potential for environmental impact reduction is even more profound.

References

[1] Schwartz et al., "Green AI," arXiv, 2019. https://arxiv.org/html/2511.07090v1

[2] "The Carbon Footprint of AI," Giorgia Cantisani, 2020. https://giorgiacantisani.github.io/2020/11/10/carbon_footprint_AI.html

[3] "Green in AI vs Green by AI," Iberdrola, 2025. https://www.iberdrola.com/about-us/our-innovation-model/artificial-intelligence/green-ai

[4] "Principles of Green AI," Telekom, 2025. https://www.telekom.com/en/company/details/principles-for-green-artificial-intelligence-1077104

[5] "Understanding FP32, FP16, and INT8," DatabaseMart. https://www.databasemart.com/blog/fp32-fp16-bf16-int8

[6] "Quantization Mapping Methodologies," F1000Research. https://f1000research.com/articles/14-419

[7] "PyTorch Dynamic Quantization Tutorial," ArikPoz, 2025. https://arikpoz.github.io/posts/2025-04-16-neural-network-quantization-in-pytorch/

[8] "Indian 2025 Vintage Grid Emission Factor," Universal Carbon Registry, 2026.(https://medium.com/@UniversalCarbonRegistry/ucr-cou-standard-update-2025-vintage-ucr-indian-grid-emission-factor-announced-d9a5efd237bf)

[9] "Historic Advancements in India's Power Sector," PIB, 2025.(https://www.pib.gov.in/PressReleasePage.aspx?PRID=2215187&reg=3&lang=1)

[10] "Madhya Pradesh Decarbonisation Performance," Ember Energy, 2026. https://ember-energy.org/latest-insights/indian-states-electricity-transition-set-2026/dimension-1-decarbonisation/

[11] "CodeCarbon: Carbon Footprint Tracking," Official Site. https://codecarbon.io/

[12] "Tracking Carbon in Python Workflows," Medium. https://fatehaliaamir.medium.com/codecarbon-tracking-and-reducing-carbon-emissions-in-computing-9abe49bae9af

[13] "Standardized Energy Consumption Index for DL," MDPI, 2025. https://www.mdpi.com/1424-8220/25/3/846

[14] "India Notifies Emission Intensity Targets," ICAP, 2026. https://icapcarbonaction.com/en/newsia-notifies-emission-intensity-targets-nine-sectors-under-carbon-credit-trading-scheme

[15] "Ryzen 5 7535HS Benchmarks," Notebookcheck.(https://www.notebookcheck.net/AMD-Ryzen-5-7535HS-Processor-Benchmarks-and-Specs.681414.0.html)

[16] "NVIDIA GeForce RTX 2050 Mobile Specs," TechPowerUp. https://www.techpowerup.com/gpu-specs/geforce-rtx-2050-mobile.c3859

[17] "RTX 2050 Mobile Moniker and Architecture," Tom's Hardware. https://www.tomshardware.com/news/geforce-rtx-2050-mechrevo-laptop

[18] "CodeCarbon Calculation Methodology," MLCO2. https://mlco2.github.io/codecarbon/methodology.html

[19] "CodeCarbon Fallback and TDP Methodology," PyPI. https://pypi.org/project/codecarbon/2.1.4/

[20] "ResNet-18 Model Card for CIFAR-10," HuggingFace. https://huggingface.co/jaeunglee/resnet18-cifar10-unlearning

[21] "ResNet-18 Inference Energy Benchmarks," PMC, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11820128/

[22] "ResNet-18 vs Vanilla CNN for CIFAR-10," Medium. https://medium.com/@lydia.benzemrane/resnet-vs-vanilla-cnn-on-cifar-10-what-22-hours-of-training-actually-taught-me-05287cdfdbd6

[23] "Dynamic Mixed-Precision Quantization Benefits," arXiv, 2025. https://arxiv.org/html/2508.09176v1

[24] "Energy-Aware AI: Quantifying and Reducing Carbon Footprint," Medium. https://medium.com/@khayyam.h/energy-aware-ai-quantifying-and-reducing-the-carbon-footprint-of-model-training-and-inference-c18747edef4c

[25] "CPU Thread Optimization in Mobile AI," arXiv, 2025. https://arxiv.org/html/2505.06461v1

[26] "PTQ vs QAT Accuracy Trade-offs," ResearchGate.(https://www.researchgate.net/figure/Accuracy-of-ResNet-18-for-different-sized-calibration-sets-at-various-quantization-levels_fig9_355019034)