Fairer AI Carbon Accounting: Incorporating Market-based Attribution and Uncertainty in Embodied and Operational Carbon Footprint

Xiaoyang Zhang
Hong Kong Polytechnic University
Hong Kong, China
xiaoyang.zhang@connect.polyu.hk

Yang Deng
Hong Kong Polytechnic University
Hong Kong, China
yang2.deng@connect.polyu.hk

Fang He
Hong Kong Polytechnic University
Hong Kong, China
fangf.he@connect.polyu.hk

Dan Wang
Hong Kong University of Science and Technology
Hong Kong, China
wangdan@ust.hk

Abstract

The computational demands of large-scale AI models raise significant concerns about their carbon footprint. Most carbon accounting methods for large-scale AI models suffer from three key limitations: they overlook embodied carbon (from hardware manufacturing) or model it simplistically, rely on location-based carbon attribution that fails to reflect individual corporate efforts to decarbonize (e.g., via Power Purchase Agreements (PPAs)), and are deterministic, ignoring inherent uncertainties. This paper proposes PUMA, a **P**robabilistic **U**ncertainty **M**arket **A**ttribution carbon accounting model for large-scale AI models. PUMA integrates market-based carbon intensity to accurately account for the impact of PPAs and employs probabilistic modeling to capture uncertainties in the carbon accounting for AI models arising from spatiotemporal variations in manufacturing and operation, as well as evolving efficiency. We make an effort to develop a comprehensive carbon dataset by aggregating related data from diverse sources, and then we implement a simple yet effective Kernel Density Estimate (KDE) on the distribution of the parameters from the collected dataset. We compare PUMA with LLMCarbon, the state-of-the-art carbon accounting model for large AI models. The deviation of the accounting result is significant, reaching up to around 201%.

CCS Concepts

• **Social and professional topics → Sustainability**; • **Computing methodologies → Artificial intelligence**.

Keywords

Carbon Accounting, Large AI model, Sustainable Computing

ACM Reference Format:
Xiaoyang Zhang, Yang Deng, Fang He, and Dan Wang. 2026. Fairer AI Carbon Accounting: Incorporating Market-based Attribution and Uncertainty in Embodied and Operational Carbon Footprint. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3774904.3793002

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3793002

1 Introduction

Large-scale AI models have shown remarkable effectiveness in diverse applications (e.g., video [30], speech [43, 44], recommender system [14], etc.). However, the increasing scale of model parameters and training data sharply raises computational demand, resulting in considerable carbon footprints. In alignment with the United Nations' Sustainable Development Goals, the AI community is increasingly focusing on social good [24], particularly on decarbonizing society [12, 42].

Carbon accounting, which quantifies a product's carbon footprint, is essential for assessing environmental impacts and guiding carbon reduction strategies. This quantification process enables organizations to establish emission reduction targets, ensure regulatory compliance, and showcase their dedication to sustainability. For large-scale AI models, the total footprint includes two components: operational carbon arising from electricity use during the operation of AI models, and embodied carbon associated with the manufacture of AI hardware that runs the AI models.

Although recent studies have begun to quantify the carbon footprint of large-scale AI models [16, 42], most methods have three key limitations. First, they emphasize operational emissions while omitting embodied emissions or modeling them with oversimplified assumptions. In particular, many approaches apply coarse, class-level averages from Life Cycle Assessment (LCA) reports to represent embodied carbon. As power grids decarbonize and data centers procure carbon-free energy, operational carbon is likely to fall, implying that embodied emissions will account for an increasing share of the total footprint of large-scale AI models. Second, current methods all rely on location-based carbon attribution, which assigns a uniform, grid-average carbon intensity to all electricity consumers within the grid. They fundamentally fail to account for individual corporate efforts to decarbonize, such as investments in renewable energy via Power Purchase Agreements (PPAs). These approaches overestimate the carbon footprint of the AI model run by the entities that invest in renewables via PPAs. Conversely, they underestimate the carbon of the AI models that run by the entities without PPAs. This underestimation occurs because the location-based method fails to account for the residual grid mix, which is the energy sources mix remaining after PPA-contracted renewable energy is subtracted. As renewable energy procurement through PPAs increases, the residual grid becomes more carbon-intensive, leading to a growing disparity between the location-based average and the actual carbon intensity applicable to non-purchasers. Third, existing carbon models for AI are deterministic and thus fail to represent inherent uncertainty in the carbon footprint of large-scale AI models. Key sources of uncertainty include: (1) geotemporal manufacturing variability: a hardware product instance can be fabricated in diverse time periods and from different regions, resulting in different embodied carbon, due to the spatial-temporal dynamics in the carbon intensity of electricity consumed in the manufacturing process; (2) dynamic manufacturing evolution: annual PPAs volume changes, yield and energy efficiency improvement across time will affect the embodied carbon; (3) dynamic operating context: the operational carbon of AI models can vary significantly depending on when and where the operation occurs as the carbon intensity of electricity varies in the spatial-temporal dimension.

This paper proposes PUMA, a **P**robabilistic **U**ncertainty **M**arket **A**ttribution carbon accounting model for large-scale AI models. PUMA is designed to capture the effects of Power Purchase Agreements (PPAs) and generate probabilistic carbon accounting outcomes. Specifically, we introduce the market-based carbon intensity to integrate PPAs into the carbon footprint accounting of large-scale AI models. Then, we develop parameter models for individual hardware components, such as processors, memory, and storage, that contribute to the AI models' embodied carbon. To estimate the parameter distributions, a comprehensive hardware and electricity dataset is constructed by integrating information from multiple sources. These include Environmental, Social, and Governance (ESG) reports from hardware suppliers, statistical data from grid operators, industry analyses, and peer-reviewed literature, as summarized in Table 1. A straightforward yet efficient distribution modeling approach is employed: collected data are first converted into frequency histograms and subsequently processed using Kernel Density Estimation (KDE) to derive continuous probability density functions for the parameters.

We evaluate the performance of PUMA against LLMCarbon, the state-of-the-art carbon accounting approach for large-scale AI models. The evaluation is performed based on four representative large AI models (XLM, T5, GPT-3, and Switch) [29, 42]. We compare the performance at key distribution percentiles (minimum, 20th, 50th, 80th, and maximum) of PUMA's probabilistic outputs against LLMcarbon. The results show that the deviations are significant, reaching up to around 145% for the embodied carbon and around 201% for the operational carbon, which highlight the critical importance of uncertainty-aware modeling for the carbon footprint of large-scale AI models.

Our contributions are summarized as follows:

- We propose a new uncertainty-aware carbon accounting model with market-based attribution for large AI models. This model produces distribution-based estimates of carbon footprint rather than point estimates, enabling AI company to incorporate risk assessment into sustainability decisions.
- We formalize and integrate market-based carbon attribution into AI carbon accounting, moving beyond the conventional location-based method. This allows for a more accurate reflection of an AI company's individual renewable energy investments (via PPAs).
- We make an effort to develop a comprehensive carbon dataset comprising PPA data, and AI hardware-related parameters across multiple technology nodes, drawing on diverse sources including technology reports, ESG and LCA reports, as well as real-world electricity residual carbon intensity data from 30 regional grid operators.

2 Background

2.1 Large-scale AI Model Carbon Footprint.

The carbon footprint of large-scale AI models includes two categories: operational carbon and embodied carbon. Operational carbon is from the electricity consumed during model deployment. Deploying a large-scale AI model demands not only considerable electricity but also significant computational hardware, such as high-performance GPUs (e.g., NVIDIA A100 [28]) and storage. Embodied carbon is the emissions produced during the manufacturing of the hardware required to run these AI models. The processors supporting large-scale AI models are often fabricated using cutting-edge semiconductor processes (such as 5nm technology), which contribute substantially to the embodied carbon of AI models. As the share of renewable energy in power grids increases and more data centers transition to carbon-neutral power sources, the operational carbon emissions from running AI models are expected to decrease. Consequently, the embodied carbon will represent a growing share of AI models' total carbon footprint [18, 46].

2.2 Carbon Attribution

To reduce carbon footprint, some companies are shifting electricity use to regions or periods with low grid carbon intensity, while many companies are investing in renewable energy via PPAs. PPAs are usually long-term contracts for renewable energy between a consumer and an electricity producer, wherein the consumer can claim renewable energy credits for their investment and lower the emissions caused by the electricity they consume. Under the Scope 2 GHG Protocol guidance [1], carbon attribution to consumers can follow two distinct approaches, depending on how renewable generation and the underlying grid mix are allocated. The location-based carbon attribution assigns an identical electricity mix to all consumers within a defined geographic area. Under this approach, green energy is credited to the grid as a whole, and the carbon intensity is calculated based on the average emissions of the entire grid mix, incorporating both renewable and non-renewable sources in proportion to their actual generation. Importantly, this method does not account for individual green energy investments made by specific consumers. Instead, any renewable energy contributions are shared collectively among all consumers in the region. The market-based carbon attribution enables consumers who invest in renewable energy to claim the environmental attributes of that electricity and account for lower carbon emissions, even if the physical power they consume comes from the grid, which comprises both renewable and non-renewable sources. For any remaining electricity demand not covered by such investments, or for consumers without renewable contracts, carbon emissions are calculated using the residual grid mix. This residual mix excludes all electricity that has been claimed under contractual instruments. The market-based attribution allows electricity to be attributed according to investment sources, resulting in varying carbon intensities across different consumers.

2.3 Location-based Carbon Intensity

The location-based carbon intensity (CI_l) of electricity is defined as the carbon emission rate (in g/kWh) during power generation. This intensity is calculated as the weighted average of the carbon intensities of all contributing energy sources, based on their respective shares of electricity generation. The mathematical formulation for the carbon intensity is as follows:

$$CI_l = \frac{\sum ef^k \times E^k}{\sum E^k} \tag{1}$$

where ef^k and E^k represent the carbon emission factor and the electricity generated by energy source k, respectively.

3 Methodology

3.1 The Boundary of PUMA

We examine the carbon footprint across four standard lifecycle stages: (1) Hardware manufacturing; (2) Hardware transport; (3) Operational use, covering emissions resulting from software execution, primarily due to electricity consumption; and (4) End-of-life processing. Among these, the operational use phase corresponds to the operational carbon of AI models, while the remaining stages contribute to their embodied carbon. Manufacturing and operational use are the dominant sources of emissions, accounting for the vast majority [3]. While others make negligible contributions and are therefore omitted. Within manufacturing, we focus on the main emissions from materials, energy, and chemical gases, while excluding ancillary components such as buildings, cooling infrastructure, and human labor.

3.2 Market-based Carbon Intensity

When certain renewable energy sources are contracted out via PPAs, the electricity source mix that remains within the grid is referred to as the residual grid mix. The carbon intensity associated with this residual mix is the residual carbon intensity (CI_{res}) [25]. Consider a grid where ef^k denotes the constant carbon emission factor of source k (e.g., wind), E^k represents the electricity generated by source k, of which E^k_{ppa} represents the portion contracted under PPAs. Let E denote the total electricity generation in the grid, and E_{ppa} be the total PPA-contracted electricity; it follows that $\sum E^k_{ppa} = E_{ppa}$. Under the market-based attribution, entities holding PPAs can claim the low-carbon benefits associated with their procured electricity. As a result, all renewable energy covered by PPAs is subtracted prior to computing CI_{res}. Following that, we have Eq. 2.

$$CI_{res} = \frac{\sum_{k \in \mathcal{E}} ef^k \cdot (E^k - E^k_{ppa})}{E - E_{ppa}} \tag{2}$$

Then, the carbon intensity of a consumer using the market-based attribution (CI_m) in that grid can be calculated by Eq. 3

$$CI_m = CI_{res} \cdot (1 - f_{ppa}) \tag{3}$$

where $f_{ppa} \in [0, 1]$ denotes the proportion of a consumer's electricity supply covered by PPAs. Unlike the location-based attribution, the market-based carbon intensity varies across consumers within the same grid. Consumers without PPAs (i.e., $f_{ppa} = 0$) have a carbon intensity $CI_m = CI_{res}$, whereas those able to meet their entire electricity demand via PPAs achieve $CI_m = 0$. This approach systematically benefits investors in renewable energy over non-investors. For more about location-based carbon intensity, please refer to Appendix 2.3.

Figure 2 presents a histogram showing the relative increase in CI_{res} compared to CI_l across 265 global regions in 2024, under the scenario where all renewable energy sources are contracted out (data source: ElectricityMaps [26]). Regions with greater integration of solar and wind power exhibit more significant rises in CI_{res}. As renewable energy continues to expand in power systems worldwide, coupled with the growing procurement of electricity through PPAs, this gap is anticipated to widen further in the future. Figure **??** illustrates the difference between location-based carbon intensity and residual carbon intensity over a typical week in Germany, where the renewable energy is 100% contracted out in the power grid in 2024 [2]. The grid derives a significant proportion of its electricity from solar generation, resulting in considerably lower grid carbon intensity during daylight hours. As the share of PPAs grows, not only does CI_{res} increase, but the temporal fluctuation in carbon intensity also becomes less pronounced. Figure 1 shows histograms of hourly location-based carbon intensity and residual carbon intensity data in 2024 in four regions with their individual kernel density estimates, where we can also find that the difference between location-based and market-based carbon intensity is significant.

3.3 Embodied Carbon Modeling

We model the embodied carbon of AI models (EC_{model}) from three key components as Eq.4 shows, i.e., the carbon caused by processors (EC^p_{model}), memory (EC^m_{model}), and storage (EC^s_{model}).

$$EC_{model} = EC^p_{model} + EC^m_{model} + EC^s_{model} \tag{4}$$

3.4 Embodied Carbon of AI models Associated with Processors

The embodied carbon of AI models associated with processors used are modeled based on the following components: (a) the operational duration of AI models on the processors, represented as t_p, relative to the processor's total lifetime T_p; (b) carbon resulting from electricity consumption during hardware fabrication, determined by multiplying the electricity consumed Per unit of die Size (EPS) by the market-based carbon intensity of the manufacturing electricity supply, CI_m; (c) carbon from raw material usage, quantified as MPS (Material required Per Size); (d) emissions from specialty gases such as fluorinated compounds, expressed as GPS (Gas emitted Per Size); (e) the processor's die size, denoted as S; and (f) the fabrication

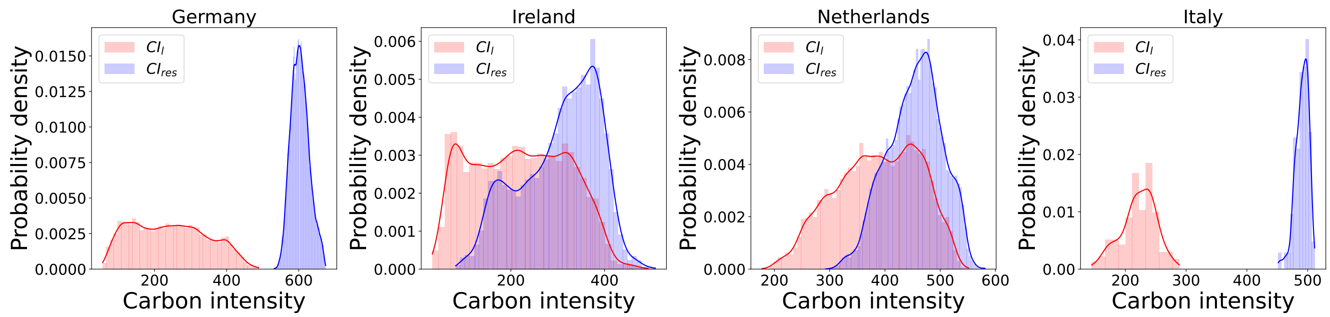

Figure 1: Histograms of hourly location-based carbon intensity and residual carbon intensity in 2024 with their individual kernel density estimates.

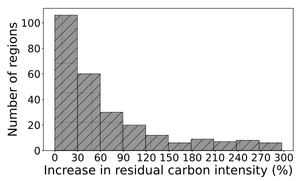

Figure 2: Average increase in residual carbon intensity when all renewable energy is contracted out.

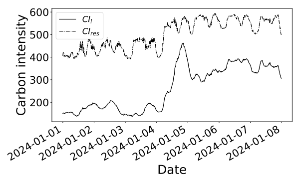

Figure 3: Weekly trace of location-based carbon intensity and residual carbon intensity in Germany.

yield Y. Then, the embodied carbon of processors attributable to AI models can be formulated as follows:

$$EC^{p}_{model} = \frac{t_p \cdot S}{T_p \cdot \tilde{Y}} \cdot (\tilde{CI}_m \cdot \tilde{EPS} + GPS + MPS) \quad (5)$$

Here, we denote $\tilde{CI}_m$ as the probabilistic model of market-based carbon intensity CI_m. Similarly, we have $\tilde{Y}$ and $\tilde{EPS}$. A detailed characterization of uncertainty for these parameters will be provided below.

Market-based carbon intensity distribution. In embodied carbon accounting, CI_m represents the carbon intensity of electricity used in semiconductor manufacturing. This key factor depends on the composition of energy sources (e.g., solar, wind, etc.) for power generation, along with related PPAs. The carbon intensity varies spatiotemporally, influenced by production timelines, geographic placement of manufacturing sites, and different PPAs. A major source of uncertainty stems from temporal fluctuations across yearly cycles, which arise from seasonal changes in renewable energy supply and variations in electrical demand and PPAs. Figure 4 shows the carbon intensity of processors manufactured by TSMC in Taiwan and Intel in the USA with different carbon attribution, which is derived from PPAs data and carbon intensity data from 2021 to 2024.

Yield Distribution. Semiconductor yield is the fraction of good, defect-free dies on a wafer relative to the total die count. This parameter is inherently uncertain, driven mainly by time-varying defect densities (D) across fabs. To study this systematically, we analyze TSMC's historical defect density records for four technology nodes [13]. We construct defect density histograms and apply Kernel Density Estimation (KDE) [41] to obtain probability density functions of the defect density distributions, as shown in Figure 5. Then we can compute yields using the Poisson yield mode: $Y = e^{(-S \cdot D)}$ [15].

EPS Distribution. Uncertainty in *EPS* primarily arises from temporal fluctuations in energy efficiency across semiconductor fabrication stages. We construct *EPS* distributions using *EPS* estimates from the STEC model [46], ACT model [18], and imec model [8], together with annual efficiency improvement data reported in TSMC's ESG disclosures [38] for multiple technology nodes (e.g., 5 nm, 10 nm). The workflow comprises three transformations: (a) normalize per-node energy efficiency over time relative to a baseline year; (b) adjust raw *EPS* values by dividing by these normalized efficiency to reflect technological progress; and (c) model the distribution by implementing a dual-stage distribution modeling approach, i.e., converting the processed data into frequency histograms, then applying KDE to obtain continuous probability density functions of $\tilde{EPS}$, shown in Figure 6.

3.5 Embodied Carbon of AI models Associated with Memory

We model the embodied carbon of AI models associated with memory from following components: (a) the fraction of the device's lifetime utilized by the model, captured by the ratio of memory runtime t_m to memory lifetime T_m; (b) memory fabrication-related emissions attributable to electricity use, computed by multiplying the electricity per unit size (*EPS*) by the market-based carbon intensity during manufacturing (CI_m) and then divided by bit density (*BD*); (c) emissions released independent of electricity, including materials, packaging, denoted α_m; and (d) the installed memory capacity C_m. Then, the embodied carbon of AI models associated with memory can be modeled as follows:

$$EC^{m}_{model} = t_m / T_m \cdot C_m \cdot (\tilde{CI}_m \cdot \tilde{EPS} / BD + \alpha_m) \quad (6)$$

Unlike processors, where manufacturing is often concentrated at a single foundry (e.g., most GPUs at TSMC in Taiwan), memory is produced by multiple vendors (e.g., SK Hynix, Samsung, Micron, etc.). When the fabrication location is uncertain, we model the region as a discrete random variable. Regional probabilities are assigned in proportion to each area's share of global IC capacity for the relevant process node [7], using capacity splits reported

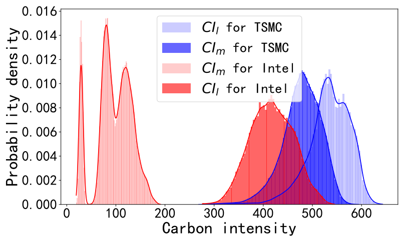

Figure 4: Location and market-based *CI* for processors.

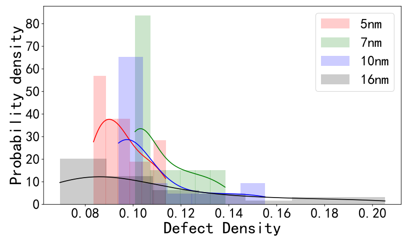

Figure 5: Defect Density (D) for processors.

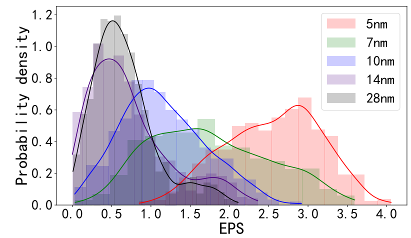

Figure 6: Energy per size (EPS) for processors.

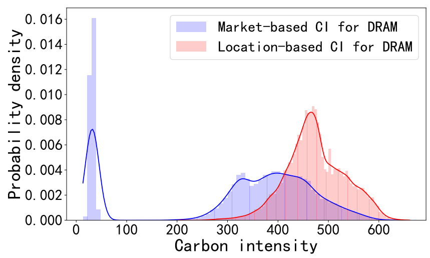

Figure 7: The location-based and market-based Carbon Intensity for memory.

by industry sources [6] as weights. We then construct a composite carbon intensity distribution for each node via a mixture approach that integrates: (a)regional distributions: the distributions developed based on historical carbon intensity data for each major manufacturing region and PPAs data of each manufacturer from 2021 to 2024; (b) capacity-weighted sampling: a Monte Carlo sampling strategy where region selection follows normalized capacity shares; (c) mixture aggregation: combining the sampled observations across regions and applying kernel smoothing to obtain the final distribution.

This framework captures uncertainty from PPAs, geographical production distributions, and temporal energy mix variations. As a result, Figure 7 compares the market-based and location-based CI for memory. We can find obvious differences between the two attributions.

3.6 Embodied Carbon of AI Models Associated with Storage

We model the embodied carbon of AI models associated with storage from following components: (a) the utilization ratio of the storage's lifetime, given by t_s/T_s, where t_s is the model's storage usage time and T_s is the storage lifetime; (b) manufacturing emissions of storage from to electricity consumed, computed as $EPG \cdot CI_m$, where EPG is the electricity consumed per GB during fabrication and CI_m is the market-based carbon intensity; (c) the emissions released independent of electricity (e.g., materials), denoted as α_S and available from industry reports [32]; and (d) installed storage capacity C_s. Then, the embodied carbon of AI models associated with storage can be calculated as Eq. 7. Similarly, we can also get the distribution of market-based and location-based CI for storage as Figure 8, where differences between the two attributions are also obvious.

$$EC^s_{model} = t_s/T_s \cdot C_s \cdot (\tilde{CI}_m \cdot EPG + \alpha_S) \tag{7}$$

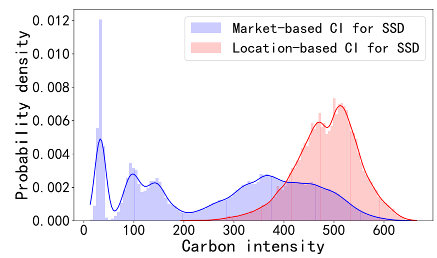

Figure 8: The location-based and market-based Carbon Intensity for storage.

3.7 Operational Carbon

The operational carbon of an AI model (OC_{model}) is the emissions attributable to the electricity it consumes during operation. It depends on the model's operational electricity consumption (E_o) and the market-based carbon intensity of the supplied electricity (CI_m), and can be expressed as:

$$OC_{model} = \tilde{CI}_m \cdot E_o = \tilde{CI}_m \cdot \sum_{i \in HardwareSet} (P_i \cdot eff_i \cdot n_i \cdot t_i) \cdot PUE \tag{8}$$

where $\tilde{CI}_o$ denotes the probability distribution of the electricity's carbon intensity, reflecting uncertainty in when and where training or inference occurs. P_i is the peak power of hardware i; eff_i is its efficiency (obtainable from hardware efficiency models [16]); n_i is the count of units of hardware i; and t_i is the runtime on hardware i, which can be estimated using FLOP-based models [16].

4 Evaluation

We make an effort to develop a hardware-electricity dataset by aggregating data from diverse sources, including ESG reports disclosures from hardware vendors, statistics released by power system operators, industry reports, and peer-reviewed literature, as summarized in Table 1. We compare PUMA against LLMCarbon, the current state-of-the-art carbon model for large-scale AI models. LLMCarbon accounts for the embodied carbon of AI models using the ACT model [18], which is designed for the embodied carbon

Table 1: The data source of the carbon dataset we developed for PUMA

Parameter	Description	Unit	Source
CI	Carbon Intensity data across 256 regions	g/kWh	ElectricityMaps[26]
ef^k	carbon emission factors	g/kWh	Research paper [46]
PPAs	Power Purchase Agreements data	%	Industrial reports [34] ESG reports [27] [21, 22, 31, 33, 40]
CI_{res}	residual carbon intensity	%	Industrial report [2]
Fabrication capacity	Global wafer fabricatioin capcity by regions	%	Industrial report [6]
Energy effciency	Annual improvement of procees energy effciency	%	ESG reports [36, 37, 39]
EPG	Electricity consumed per GB	kWh/GB	LCA reports [20, 32]
Die size	GPUs and CPUs	mm2	Industrial reports [35]
Process nodes	GPUs and CPUs	nm	Industrial reports [35]
Defect density	Defect density trend across time	%	Industrial reports[4]
EPS	Electricity consumed per die Size	kWh/cm2	Research paper [8, 18, 46]
GPS	Emission from Gas per die Size	g/cm2	Research paper [18, 46]
MPS	Emission from Material used per die Size	g/cm2	Industrial reports [9]
BD	Bit density	GB/cm2	Industrial reports [11]

Table 2: The comparison between PUMA and LLMCarbon on embodied carbon accounting.

Hardware information			
Hardware	Die size/Unit	Number	Technology
CPU	1.47 cm2	512	12nm
GPU	8.15 cm2	64	16nm
Storage	32TB	64	SSD
Memory	256GB	64	10nm ddr4

Accounting result						
Component	Model	Embodied Carbon (kg) at each Percentile				
		Min.	20th	Median	80th	Max.
Total	PUMA	138.60	179.43	238.37	295.40	498.11
	LLMCarbon			271.01		
	|Deviation|	132.41	91.58	32.64	24.39	227.10
	Deviation (%)	48.86%	33.79%	12.05%	9.00%	83.80%
CPU	PUMA	0.97	1.27	1.48	1.81	2.84
	LLMCarbon			1.16		
	|Deviation|	0.19	0.11	0.32	0.65	1.68
	Deviation (%)	16.22%	9.76%	27.84%	56.29%	145.08%
GPU	PUMA	83.67	111.15	133.50	168.88	329.2
	LLMCarbon			138.39		
	|Deviation|	54.72	27.24	4.89	30.49	190.76
	Deviation (%)	39.54%	19.68%	3.53%	22.03%	137.85%
Memory	PUMA	5.43	5.83	8.54	10.73	18.27
	LLMCarbon			10.53		
	|Deviation|	5.10	4.70	1.99	0.20	7.74
	Deviation (%)	48.40%	44.65%	18.88%	1.89%	73.52%
Storage	PUMA	48.22	61.14	94.84	113.98	147.88
	LLMCarbon			120.94		
	|Deviation|	72.72	59.80	26.10	6.96	26.94
	Deviation (%)	60.13%	49.45%	21.58%	5.75%	22.27%

accounting of computer systems. Our evaluation is based on the publicly available training data from four representative models: XLM, T5, GPT-3, and Switch [29, 42]. Evaluation is conducted by comparing PUMA's probabilistic outputs with LLMCarbon at major distribution percentiles (min., 20th, 50th, 80th, and max.) based on the collected carbon dataset.

4.1 Embodied Carbon Evaluation

We evaluate the embodied carbon accounting performance of PUMA and LLMCarbon using the publicly available XLM training data [42], which, to our knowledge, represents the only publicly available source disclosing hardware-level embodied carbon data for large-scale AI model training. The setup involves 64 servers (details in Table 2) used over 20.4 days, with hardware assumed to have a 5-year lifetime [42].

As shown in Table 2, a marked contrast exists between the single-point deterministic model (LLMCarbon) and the probabilistic PUMA model. LLMCarbon estimate consistently falls within the distribution provided by PUMA. PUMA reveals a much wider potential range, frequently showing that emissions can be over 145% higher at the maximum percentile compared to the LLMCarbon estimate. Significant deviations are also observed at the component level. For GPUs and storage, the dominant sources of uncertainty, PUMA yields a range of 83.67–329.2 t and 48.22–147.88 t, respectively, while LLMCarbon gives a simple average value of 138.39 t and 120.94 t, corresponding to maximum deviations of 137.85% and 60.13%.

We can find that LLMcarbon gives a relatively large estimate in the distribution of total embodied carbon (271.01t vs. 238.37t at the median). The reason behind this phenomenon is that many hardware manufacturers invest in renewables through PPAs, thereby reducing the carbon intensity of their electricity consumption. PUMA can capture this effect on the operational carbon accounting of large-scale AI models, whereas LLMcarbon lacks this capability. Our evaluation also reveals critical limitations in deterministic accounting and emphasizes the need for probabilistic approaches. We recommend that environmental assessments for AI systems report uncertainty intervals to better inform sustainable development practices.

4.2 Operation Carbon Evaluation

We compare PUMA with LLMCarbon on the operational carbon accounting for four large-scale AI models, XLM, T5, GPT3, and Switch, using published training details [29] as shown in Table 3. We evaluate along two dimensions: (i) spatial variation: fixing the training date and varying location across 30 countries based on real-world grid residual energy mix data [2]; and (ii) temporal variation: fixing location (Germany) while varying time (in 2024). Here, we use real-world residual carbon intensity data from the power grid to study its impact on the carbon footprint accounting of large-scale AI models, without considering the investment behavior of individual AI companies in green energy. The impact of individual companies' investment in green energy will be studied in the next section.

Table 3: The comparison between PUMA and LLMCarbon on operational carbon accounting

AI moodels	Training information	Accounting Models	Operational Carbon at each Percentile (ton)									
			Spatial Dimension					Temporal Dimension				
			Min.	20th.	Median	80th.	Max.	Min.	20th.	Median	80th.	Max.
XLM	Tarinin Day: 20.4; PUE: 1.1;Ave. Power: 342 kw; Num. of device: 512	LLMCarbon			17.63					20.69		
		PUMA	3.19	16.97	26.65	36.23	53.13	7.75	20.86	30.05	35.74	47.93
		|Deviation|	14.44	0.66	9.02	18.60	35.50	12.94	0.17	9.36	15.05	27.24
		Deviation (%)	81.91%	3.74%	51.16%	105.50%	201.36%	62.54%	0.82%	45.24%	72.74%	131.66%
T5	Tarinin Day: 20; PUE: 1.12;Ave. Power: 310 kw; Num. of device: 512	LLMCarbon			15.95					18.72		
		PUMA	2.89	15.36	24.11	32.78	48.08	7.01	18.87	27.19	32.34	43.37
		|Deviation|	13.06	0.59	8.16	16.83	32.13	11.71	0.15	8.47	13.62	24.65
		Deviation (%)	81.88%	3.70%	51.16%	105.52%	201.44%	62.55%	0.80%	45.25%	72.76%	131.68%
GPT3	Tarinin Day: 14.8; PUE:1.1;Ave. Power: 330 kw; Num. of device:10K	LLMCarbon			241.08					282.90		
		PUMA	43.71	232.12	364.43	495.37	726.53	106.01	285.26	410.95	488.71	655.38
		|Deviation|	197.37	8.96	123.35	254.29	485.45	176.89	2.36	128.05	205.81	372.48
		Deviation (%)	81.87%	3.72%	51.17%	105.48%	201.36%	62.53%	0.83%	45.26%	72.75%	131.66%
Switch	Tarinin Day: 27; PUE:1.1;Ave. Power: 245 kw; Num. of device: 1K	LLMCarbon			32.65					38.31		
		PUMA	5.92	31.43	49.35	67.09	98.40	14.35	38.63	55.65	66.19	88.76
		|Deviation|	26.73	1.22	16.70	34.44	65.75	23.96	0.32	17.34	27.88	50.45
		Deviation (%)	81.87%	3.74%	51.15%	105.48%	201.38%	62.54%	0.84%	45.26%	72.77%	131.69%

PUMA offers a comprehensive distribution of operational carbon across both spatial and temporal dimensions, whereas LLMCarbon produces only an average single-point estimate. As shown in Table 3, the estimates from LLMCarbon exhibit substantial deviations from those of PUMA across all evaluated models and percentiles. These relative deviations are especially pronounced at upper percentiles, frequently surpassing 200% in the spatial dimension and 130% in the temporal dimension. Even at lower percentiles, deviations remain notable, generally approximating 80% spatially and 0% temporally. For example, in the spatial dimension for GPT-3, PUMA yields a range of 43.71t–726.53t across percentiles, while LLMCarbon gives a single value of 241.08t, leading to a relative deviation of around 201% at maximum percentile.

We can find that LLMcarbon always gives a relatively small estimate in the distribution of operational carbon for each large-scale AI mode. The reason behind this phenomenon is that renewable energy in the grid has been contracted out via PPAs, resulting in a higher carbon intensity in the residual electricity. PUMA can capture the effect of PPAs on the operational carbon accounting of large-scale AI models, whereas LLMcarbon lacks this capability.

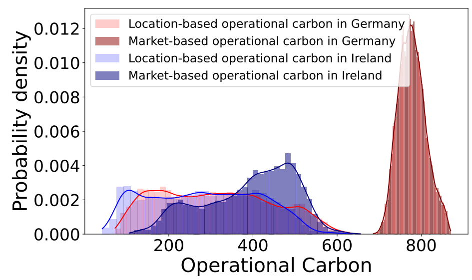

Figure 9: The distribution of operation carbon over PPAs excluded.

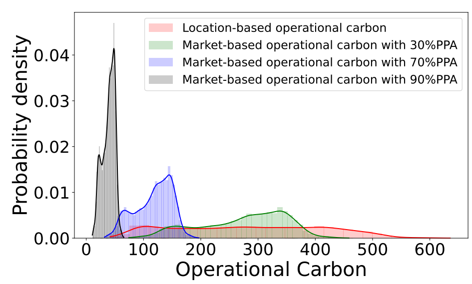

Figure 10: The distribution of operation carbon over PPAs.

4.3 The Effect of PPAs

In this section, we examine the impact of the investment in green energy via PPAs from an individual organization that operates a large-scale AI model on the carbon accounting for the AI model. Firstly, figure 9 presents the operational carbon distribution of GPT-3 training by a company without investing in green energy via PPAs in Germany or the Netherlands in 2024. It can be observed that the difference between location-based and market-based emissions remains significant. For instance, the median differs by 152.89% (307.04 kg vs. 776.49 kg) in Germany and 45.45% (282.51 kg vs. 410.91 kg) in the Netherlands. We can observe that the residual carbon intensity has a more significant effect on the operational carbon of GPT-3 when it is trained in Germany. The reason behind this is that the green energy is 100% contracted out in the German power grid in 2024 (vs. 55% in the Netherlands).

Figure 10 presents the operational carbon emissions from GPT-3 training within the Netherlands grid under the scenario where the AI companies with different sustainable actions (i.e, investments in green energy via PPAs with varying percentages). A substantial discrepancy exists between location-based and market-based carbon accounting methods. For example, at 90% PPA coverage, the location-based median can be over six times the market-based value (282.51 kg vs. 41.09 kg). The results show that operational carbon emissions decrease as the share of PPAs increases, suggesting that investing in renewable energy via PPAs can effectively reduce the carbon footprint of AI models, and PUMA can capture this effect on the carbon accounting for AI models.

5 Insights and Potential Use Case for PUMA

The development and evaluation of PUMA reveal several critical insights into carbon accounting for large-scale AI models and open up a range of practical applications for stakeholders across the AI ecosystem.

5.1 Key Insights

The Impact of Market-Based Attribution: Our results highlight the significant influence of corporate renewable energy investments, particularly through PPAs, on carbon accounting. The location-based method, which ignores such investments, systematically misrepresents the carbon footprint of both green and non-green energy consumers. PUMA's market-based approach not only rewards decarbonization efforts but also exposes the growing carbon intensity of residual grids, offering a more equitable and accurate accounting framework.

Uncertainty-aware Accounting: PUMA demonstrates that uncertainty is not merely a statistical nuance but a central feature of carbon accounting in the large-scale AI models. The significant deviations (up to 251.58% in embodied carbon and 138.76% in operational carbon) between probabilistic and deterministic models underscore the risk of relying on point estimates. By providing distributional outputs, PUMA enables AI developers and operators to quantify and communicate the variability and reliability of their carbon footprints, facilitating more robust sustainability planning and risk management.

Embodied Carbon Cannot Be Ignored: As operational carbon decreases due to grid decarbonization and efficiency gains, embodied carbon becomes an increasingly dominant component of the total footprint. PUMA's detailed modeling of hardware components, processors, memory, and storage shows that embodied emissions are subject to significant spatial, temporal, and technological uncertainties. Ignoring these factors, as current methods often do, leads to incomplete and potentially misleading environmental assessments.

5.2 Potential Use Cases

AI Company Sustainability Strategy: PUMA can help AI companies make informed decisions regarding hardware procurement, energy sourcing, and operational scheduling. By simulating different PPA scenarios and hardware choices under uncertainty, companies can optimize their carbon budgets, set risk-aware targets, and report sustainability metrics with confidence intervals, enhancing transparency and credibility.

Policy and Regulation Support: Regulators and standard-setting bodies can use PUMA as a benchmark for developing more nuanced carbon accounting guidelines for the AI industry. The model's ability to differentiate between market-based and location-based emissions can inform policies that incentivize renewable energy investments and penalize carbon-intensive operations.

Green AI Model Development: Researchers and engineers can integrate PUMA into the AI development lifecycle to evaluate the environmental impact of training scheduling and deployment options. By incorporating carbon awareness early in the design process, the community can foster the creation of lower-carbon AI systems.

Third-Party Auditing and Certification: PUMA's transparent sources and probabilistic framework provide a foundation for independent verification and certification of AI carbon footprints. Auditors can use the tool to validate corporate sustainability claims and issue certifications that reflect both average emissions and associated uncertainties.

Investment and ESG Reporting for AI company: Investors and stakeholders increasingly demand accurate environmental performance data. PUMA enables AI companies to report carbon footprints in a manner that reflects their actual renewable energy investment and operational contexts, supporting better ESG ratings and more aligned investment decisions.

6 Related Work

Current carbon accounting methods for AI models primarily focus on operational emissions, calculated as the product of electricity consumption and grid carbon intensity. Most of the existing literature is dedicated to tracking or estimating electricity consumed. Several studies have introduced software-based tools to monitor CPU/GPU power consumption in real-time during model training or inference [5, 10, 17, 19], while others estimate energy use based on hardware specifications such as thermal design power [23]. Yet these frameworks consistently overlook the embodied carbon emissions originating from AI hardware infrastructure, which represent a significant portion of the total footprint, especially as grids decarbonize and data centers adopt more renewable energy. For embodied carbon, current methods like SustainableAI [42] rely on coarse-grained manufacturer-reported average emission factors for hardware components. LLMCarbon [16] employs a deterministic parametric model for processors while using averaged data from ESG reports for memory and storage devices. Yet these approaches are deterministic and location-based, neglecting the role of PPAs and the inherent uncertainty in the carbon footprint of large AI models. PCAM [45] proposes a probabilistic carbon accounting model to quantify the uncertainties of the carbon footprints of large AI models. Yet PCAM is also location-based and neglects the uncertainty associated with PPAs in power grids. The effect of PPAs on carbon accounting for AI models is further demonstrated in Section 4.3 by comparing the accounting results from the market-based approach (PUMA) and location-based approach (PCAM).

7 Conclusion

In this paper, we propose PUMA, an uncertainty-aware carbon accounting model with market-based attribution for large-scale AI models. PUMA integrates PPAs into the accounting process and can capture the uncertainty in both operational and embodied carbon. Our evaluation revealed significant deviations up to 251.58% for embodied carbon and 138.76% for operational carbon, between PUMA's probabilistic outputs and the deterministic estimates of LLMCarbon. This highlights the critical limitations of existing location-based methods and underscores the necessity of uncertainty modeling. PUMA provides a more realistic and equitable accounting framework by: formally integrating market-based carbon intensity to reward decarbonization efforts and employing probabilistic modeling to capture spatiotemporal and technological uncertainties. By enabling risk-aware decisions and transparent reporting, PUMA supports the development of sustainable AI practices.

Acknowledgments

Dan Wang's work is supported in part by RGC GRF 15201322, 15230624, 15239925, CRF C5020-25GF, ITC ITS/052/23MX, and PolyU 1-CDKK, G-SAC8, K-ZYAP, HKUST Start up fund R9117.

References

[1] United Satets Environmental Protection Agency. 2021. Greenhouse Gas Protocol: Scope 2 Guidance. https://ghgprotocol.org/scope-2-guidance.

[2] AIB. 2024. European Residual Mixes. https://www.aib-net.org/facts/european-residual-mix/2024.

[3] Husam Alissa, Teresa Nick, Ashish Raniwala, Alberto Arribas Herranz, Kali Frost, Ioannis Manousakis, Kari Lio, Brijesh Warrier, Vaidehi Oruganti, TJ DiCaprio, et al. 2025. Using life cycle assessment to drive innovation for sustainable cool clouds. *Nature* (2025), 1–8.

[4] Anandtech. 2020. better yield on 5nm than 7nm tsmc update on defect rates for n5. https://www.anandtech.com/show/16028.

[5] Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051* (2020).

[6] SIA BCG. 2024. *EMERGING RESILIENCE IN THE SEMICONDUCTOR SUPPLY CHAIN*. https://www.semiconductors.org/wp-content/uploads/2024/05/Report_Emerging-Resilience-in-the-Semiconductor-Supply-Chain.pdf.

[7] Anvita Bhagavathula, Leo Han, and Udit Gupta. 2024. Understanding the Implications of Uncertainty in Embodied Carbon Models for Sustainable Computing. In *HotCarbon*.

[8] L Boakes, M Garcia Bardon, V Schellekens, IY Liu, B Vanhouche, G Mirabelli, F Sebaai, L Van Winckel, E Gallagher, C Rolin, et al. 2023. Cradle-to-gate Life Cycle Assessment of CMOS Logic Technologies. In *2023 International Electron Devices Meeting (IEDM)*. IEEE, 1–4.

[9] Sarah B Boyd. 2011. *Life-cycle assessment of semiconductors*. Springer Science & Business Media.

[10] Semen Andreevich Budennyy, Vladimir Dmitrievich Lazarev, Nikita Nikolaevich Zakharenko, Aleksei N Korovin, OA Plosskaya, Denis Valer'evich Dimitrov, VS Akhripkin, IV Pavlov, Ivan Valer'evich Oseledets, Ivan Segundovich Barsola, et al. 2022. Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. In *Doklady Mathematics*, Vol. 106. Springer, S118–S128.

[11] Jeongdong Choe. 2017. XPoint Memory Comparison Process Architecture. https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FR12_Choe.pdf.

[12] Paolo Cornale, Michele Tizzani, Fabio Ciulla, Kyriaki Kalimeri, Elisa Omodei, Daniela Paolotti, and Yelena Mejova. 2025. The role of science in the climate change discussions on Reddit. *PLoS Climate* 4, 5 (2025), e0000541.

[13] Ian Cutress. 2020. *Better yield on 5nm than 7nm: TSMC update on defect rates for N5*. https://www.anandtech.com/show/16028/.

[14] Yicheng Di, Xiaoming Wang, Hongjian Shi, Chongsheng Fan, Rong Zhou, Ruhui Ma, and Yuan Liu. 2025. Personalized consumer federated recommender system using fine-grained transformation and hybrid information sharing. *IEEE Transactions on Consumer Electronics* (2025).

[15] EESemi. 2005. *Test yield model*. https://www.eesemi.com/test-yield-models.htm.

[16] Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2024. LLMCARBON: MODELING THE END-TO-END CARBON FOOTPRINT OF LARGE LANGUAGE MODELS. In *The Twelfth International Conference on Learning Representations*. ICLR.

[17] Zhenxiao Fu, Fan Chen, Shan Zhou, Haitong Li, and Lei Jiang. 2025. Llmco2: Advancing accurate carbon footprint prediction for llm inferences. *ACM SIGENERGY Energy Informatics Review* 5, 2 (2025), 63–68.

[18] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*. 784–799.

[19] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21, 248 (2020), 1–43.

[20] SK Hynix. 2021. Sustainability Report. https://www.skhynix.com/sustainability/UI-FR-SA1601/.

[21] SK hynix. 2025. SK hynix ESG reports. https://www.skhynix.com/sustainability/UI-FR-SA1601/.

[22] Kioxia. 2024. Kioxia ESG reports. https://www.kioxia-holdings.com/content/dam/kioxia-hd/en-jp/sustainability/asset/2024-Sustainability-KIOXIA-EN.pdf.

[23] Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science* 8, 12 (2021), 2100707.

[24] Zhixiang Lu, Yulong Li, Feilong Tang, Zhengyong Jiang, Chong Li, Mian Zhou, Tenglong Li, and Jionglong Su. 2025. DeepGB-TB: A Risk-Balanced Cross-Attention Gradient-Boosted Convolutional Network for Rapid, Interpretable Tuberculosis Screening. *arXiv preprint arXiv:2508.02741* (2025).

[25] Diptyaroop Maji, Noman Bashir, David Irwin, Prashant Shenoy, and Ramesh K Sitaraman. 2024. Untangling carbon-free energy attribution and carbon intensity estimation for carbon-aware computing. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*. 580–588.

[26] Electricity Maps. 2025. Carbon intensity of power grids. https://app.electricitymaps.com/.

[27] Micron. 2025. Micron ESG reports. https://www.micron.com/about/sustainability.

[28] Nvidia. 2024. *NVIDIA A100 Tensor Core GPU*. https://www.nvidia.com/en-us/data-center/a100/.

[29] D Patterson, J Gonzalez, Q Le, C Liang, LM Munguia, D Rothchild, D So, M Texier, and J Dean. 2021. Carbon emissions and large neural network training. arXiv 2021. *arXiv preprint arXiv:2104.10350* (2021).

[30] Yuriel Ryan, Rui Yang Tan, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2025. Humor in Pixels: Benchmarking Large Multimodal Models Understanding of Online Comics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 14024–14050.

[31] Samsung. 2025. Samsung ESG reports. https://samsungbiologics.com/sustainability/esg-report.

[32] SEAGATE. 2024. LCA Product Reports by SEAGATE. https://www.seagate.com/esg/planet/product-sustainability/.

[33] Seagate. 2024. Seagate ESG reports. https://www.seagate.com/content/dam/seagate/assets/esg/resources/files/seagate_fy2024_esg_performance_report_pages_521.pdf.

[34] statista. 2025. Intel renewable energy purchase. https://www.statista.com/statistics/1200881/intel-renewable-energy-purchase-share-of-electricity-use-worldwide/.

[35] Techpowerup. 2024. GPU and CPU specs database. https://www.techpowerup.com/gpu-specs/.

[36] TSMC. 2021. ESG Reports by TSMC. https://esg.tsmc.com/en-US/file/public/e-all_110.pdf.

[37] TSMC. 2022. ESG Reports by TSMC. https://esg.tsmc.com/en-US/file/public/e-all_2022.pdf.

[38] TSMC. 2023. ESG Reports by TSMC. https://esg.tsmc.com/en-US/resources/ESG-data-hub?tab=reportbuilder.

[39] TSMC. 2023. ESG Reports by TSMC. https://esg.tsmc.com/en-US/resources/ESG-data-hub?tab=reportbuilder.

[40] TSMC. 2024. TSMC ESG reports. https://esg.tsmc.com/en-US/file/public/2024-TSMC-Sustainability-Report-e.pdf.

[41] Jake Vanderplas. 2024. *Density Estimation-scikit-learn.org*. https://scikit-learn.org/stable/modules/density.html.

[42] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* 4 (2022), 795–813.

[43] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023).

[44] Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. 2024. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7945–7949.

[45] Xiaoyang Zhang, He Fang, Yang Deng, and Dan Wang. 2025. Unveiling the Uncertainty in Embodied and Operational Carbon of Large AI Models through a Probabilistic Carbon Accounting Model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[46] Xiaoyang Zhang, Yijie Yang, and Dan Wang. 2024. Spatial-Temporal Embodied Carbon Models for the Embodied Carbon Accounting of Computer Systems. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*. 464–471.