Carbon Literacy for Generative AI: Visualizing Training Emissions Through Human-Scale Equivalents

Mahveen Raza
Independent Student Researcher
Toronto, Canada
mahveen.raza10@gmail.com

Maximus Powers Clarkson University
Arizona, United States
maximuspowersdev@gmail.com

Abstract

Training large language models (LLMs) consumes vast energy and produces substantial carbon emissions, yet this impact remains largely invisible due to limited transparency. We compile reported and estimated training emissions (2018–2024) for 13 state-of-the-art models and reframe them through human-friendly comparisons, such as trees required for absorption, and per-capita footprints, via our interactive demo. Our findings highlight both the alarming scale of emissions and the lack of standardized reporting. We position this work as a contribution to **Creative Practices**, advancing public awareness and encouraging model reporting transparency of generative AI (GenAI). Our demo is available: `https://neurips-c02-viz.vercel.app/`.

1 Introduction

As GenAI becomes widely adopted, its environmental impact, particularly during training, remains overlooked. Studies show that large models generate substantial carbon emissions; for example, GPT-4 alone is estimated to have produced 5,184 t CO_2 [1]. Yet discussions around AI largely emphasize benchmarking [2] and safety [21], with limited attention to sustainability. Recent work has begun addressing this gap through analyses of Google's Gemini model [14] and broader ecological surveys [6, 19]. Building on these efforts, we evaluate the environmental impact of 13 state of the art GenAI models released between 2018 and 2024 to highlight the true scale of their carbon footprint.

Background and Related Work Prior work on model emissions [1, 15] reveals two major gaps in sustainability studies. First, developers rarely disclose raw training emissions, especially for closed-source models. Second, available data is fragmented: model cards may report FLOPs, GPU hours, or hardware, but omit energy or carbon values. For example, GPT-3 reported parameters (175B) [3] but no emissions, leaving researchers to rely on indirect estimates [15]. Broader surveys such as the Stanford AI Index [1] aggregate such estimates, yet only LLaMA-2 [11] and LLaMA-3 [12] provide official disclosures. This lack of transparency highlights a critical gap for advancing sustainable AI.

Contribution. We compile reported and estimated training emissions for 13 GenAI model families and reframe them through human-friendly comparisons, such as tree absorption [8] and per-capita footprints [19]. This translation offers an accessible view of the scale of emissions and their broader environmental consequences.

2 Methodology

Model Selection (2018–2024). We analyzed 13 prominent GenAI models, from early architectures like BERT [5] and GPT-2 [16] to large-scale releases such as GPT-4 [13], the LLaMA family [18, 12], and DeepSeek v3 [4], using reported or estimated training emissions.

Preprint.

Emission Data Collection For each model, we gathered available training emissions data from published reports, technical documentation, or prior sustainability analyses. In cases where official carbon emissions were disclosed (e.g., LLaMA-2 and LLaMA-3), we report the values as *reported (R)*. For all other models, emissions were estimated (*E*) based on published details such as FLOPs, GPU hours, or hardware configurations [15, 1]. Where raw data was incomplete, we relied on standardized estimation approaches outlined in sustainability studies. To make emissions more interpretable, we translate raw CO_2 values into two equivalents:

Tree Absorption. We estimate the number of trees required to absorb a model's emissions using the assumption that one tree absorbs $\approx$ 25 kg of CO_2/year [8], i.e., Trees Required $= \frac{\text{Emissions (kg)}}{25}$.
Human Equivalence. We compare emissions to an average human's yearly CO_2 footprint of $\approx$ 4.8 t (4800 kg) [19], i.e., Human Years $= \frac{\text{Emissions (kg)}}{4800}$.

3 Results

Table 1: Training emissions of state-of-the-art GenAI models (2018–2024), with equivalent environmental impacts. Tree absorption assumes 25kg CO_2/year [8], and average per-capita footprint is 4,800kg/year [19]. R = reported, E = estimated. *Note: Each row cites (Model) the original model paper/card for dataset and training details, and (CO_2 (kg)) the source of the reported or estimated emissions.*

Model	Year	CO_2 (tonnes)	CO_2 (kg)	Type (R/E)	Tree Eq. (25kg/yr)	Human Eq. (4,800kg/yr)
BERT (base) [5]	2018	0.652	652 [7]	E	26.1 trees	0.14 yrs ($\approx$1.7 mo.)
BERT-Large [5]	2018	2.6	2,600 [1]	E	104 trees	0.54 yrs ($\approx$6.5 mo.)
GPT-2 (OpenAI) [16]	2019	0.735	735 [7]	E	29.4 trees	0.15 yrs ($\approx$1.8 mo.)
RoBERTa [10]	2019	5.5	5,500[1]	E	220 trees	1.15 yrs
GPT-3 (175B) [3]	2020	502	502,000 [15]	E	20,080 trees	104.6 yrs
BLOOM (176B) [9]	2022	25	25,000 [15]	E	1,000 trees	5.21 yrs
OPT (175B) [20]	2022	70	70,000 [15]	E	2,800 trees	14.6 yrs
Gopher (280B) [17]	2022	352	352,000 [15]	E	14,080 trees	73.3 yrs
GPT-4 (OpenAI) [13]	2023	5,184	5,184,000 [1]	E	207,360 trees	1080 yrs
LLaMA-2 (70B) [18]	2023	539	539,000 [11]	R	21,560 trees	112.3 yrs
LLaMA-3.1 (405B) [12]	2024	8,930	8,930,000 [1]	E	357,200 trees	1860 yrs
LLaMA-3 (70B) [12]	2024	2,290	2,290,000 [12]	R	91,600 trees	477 yrs
DeepSeek v3 [4]	2024	597	597,000 [1]	E	23,880 trees	124 yrs

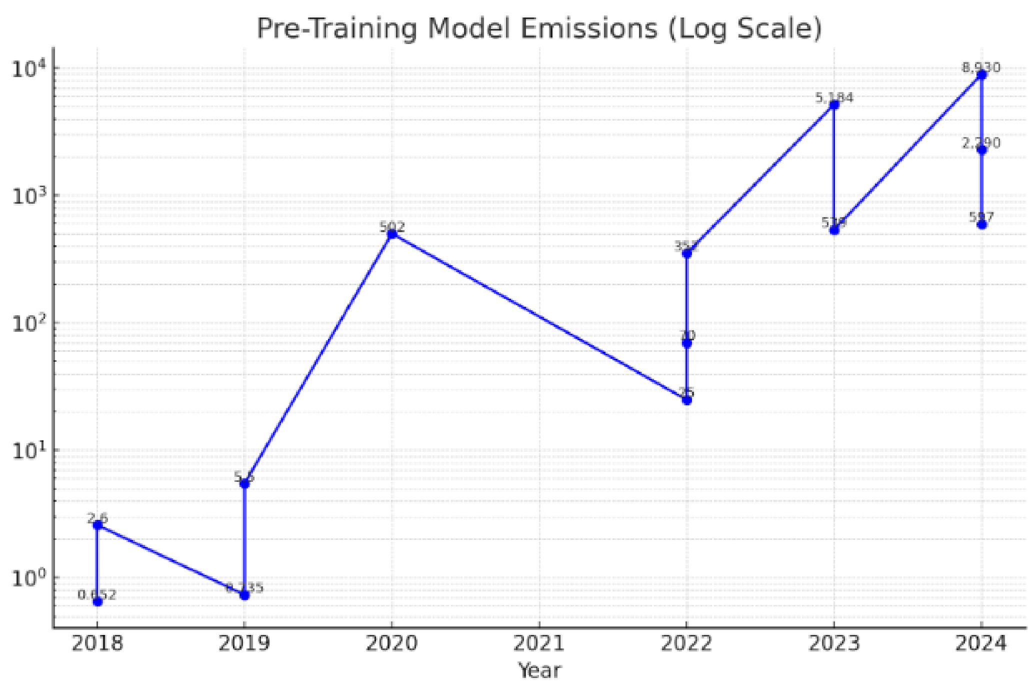

Figure 1: Exponential growth of training emissions (t CO_2) across major LLMs (2018–2024).

The results in Table 1 and Figure1 visualize an exponential rise in training emissions as model size increases, emphasizing the environmental impact. Early models like BERT (2018) and GPT-2 (2019) produced under 1 tonne of CO_2, while frontier models such as GPT-4 and LLaMA-3.1 reached thousands of tonnes, requiring hundreds of thousands of trees for absorption. Only LLaMA-2 and LLaMA-3 reported official emissions; all others rely on indirect estimates [15, 1], underscoring the need for standardized reporting. Human-scale equivalents make this tangible: GPT-4's footprint equals $\approx 1{,}080$ years of an average human's emissions, LLaMA-3.1 $\approx 1{,}850$ years, while GPT-2 corresponds to less than two months. These sharp jumps reflect how compute demands outpace linear scaling laws.

4 Discussion

Environmental and Social Implications. Scaling GenAI carries significant environmental costs, with training emissions of frontier models rivaling those of entire communities. LLaMA-3 alone produced $\approx 2{,}290$ t CO_2, equivalent to ≈ 477 average human-years of emissions, raising equity concerns as the environmental burden is global while the benefits are concentrated among a few corporations. The visual trend underscores how model scaling has rapidly accelerated emissions, emphasizing the need for efficiency-focused research and transparent reporting. This connects directly to the broader issue of disclosure and accountability.

Transparency and Accountability. Of the 13 models reviewed, only LLaMA-2 and LLaMA-3 disclosed official emissions. The rest required indirect estimation from FLOPs, GPU hours, or hardware specs, revealing a pressing need for standardized emissions reporting in model documentation.

Assumptions and Limitations. Our estimates are approximations: tree absorption (25 kg CO_2/year) varies by species and region, and human per-capita footprints differ globally. We use global averages for comparability, providing proxies that make model emissions more accessible despite inherent uncertainty.

Future Directions. Sustainable AI requires transparent reporting, efficiency-focused research, and integration of sustainability into governance frameworks. Future work should account for lifecycle emissions, including inference and deployment.

References

[1] AI Index Steering Committee. Ai index report 2025. Technical report, Stanford Institute for Human-Centered Artificial Intelligence (HAI), 2025. URL `https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf`. Accessed: 2025-08-26.

[2] Debarag Banerjee, Pooja Singh, Arjun Avadhanam, and Saksham Srivastava. Benchmarking llm powered chatbots: methods and metrics. *arXiv preprint arXiv:2308.04624*, 2023.

[3] Tom B. Brown et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL `https://arxiv.org/abs/2005.14165`.

[4] DeepSeek-AI. Deepseek llm: Scaling open-source language models with 10t tokens. *arXiv preprint arXiv:2405.04434*, 2024. URL `https://arxiv.org/abs/2405.04434`.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. URL `https://arxiv.org/abs/1810.04805`.

[6] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. Measuring the carbon intensity of AI in cloud instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894, 2022.

[7] Register Dynamics. Artificial footprints series: The environmental impact of ai, 2024. URL `https://www.register-dynamics.co.uk/blog/artificial-footprints-series-the-environmental-impact-of-ai`. Accessed: 2025-08-27.

[8] EcoTree. How much co_2 does a tree absorb? https://ecotree.green/en/how-much-co2-does-a-tree-absorb, 2025. Accessed August 2025; estimates annual CO_2 absorption of 10–40 kg per tree, approximately 25 kg/year on average.

[9] Teven Le Scao, Angela Fan, Christopher B. Akiki, Ellie Pavlick, Suzana Ilić, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2023. URL https://arxiv.org/abs/2211.05100.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

[11] Meta AI. Llama 2: Open foundation and fine-tuned chat models. https://huggingface.co/meta-llama/Llama-2-70b, 2023. Accessed: 2025-08-27.

[12] Meta AI. Llama 3: Advancing open foundation models. https://huggingface.co/meta-llama/Meta-Llama-3-70B, 2024. Accessed: 2025-08-27.

[13] OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

[14] Organic Consumers. How much energy does google's AI use? we did the math. *Organic Consumers Association*, 2025. Accessed via web; based on Google's technical report.

[15] David Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. URL https://arxiv.org/abs/2104.10350.

[16] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. OpenAI Technical Report.

[17] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. URL https://arxiv.org/abs/2112.11446.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

[19] Worldometers. Co2 emissions per capita. https://www.worldometers.info/co2-emissions/co2-emissions-per-capita, 2025. Accessed August 2025; reports global per-capita CO_2 emissions of 4.8 tons for 2022 (and includes country-level data).

[20] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

[21] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. MultiTrust: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models, December 2024. URL http://arxiv.org/abs/2406.07057. arXiv:2406.07057 [cs].

NeurIPS Paper Checklist

1. **Claims**
2. Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The abstract and introduction clearly state the contribution: compiling reported and estimated emissions for 13 models (2018–2024), translating them into human-friendly comparisons, and highlighting transparency gaps. The claims match the presented results.

3. **Limitations**
4. Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses assumptions (average tree absorption, global per-capita averages) and limitations (incomplete reporting, reliance on estimates), framing them as approximations.
5. **Theory assumptions and proofs**
6. Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results or formal proofs; it is empirical and descriptive.
7. **Experimental result reproducibility**
8. Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All emission values are either cited from official sources or estimated using documented methods (FLOPs, GPU hours, standardized formulas). Tables and formulas are provided to allow reproducibility of calculations.
9. **Open access to data and code**
10. Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

    Answer: [Yes]

    Justification: The interactive demo and dataset of compiled emissions are made available online through an anonymized demo URL that is provided in the paper, enabling public access to the data and results.
11. **Experimental setting/details**
12. Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

    Answer: [Yes]

    Justification: The methodology section details model selection, emission data collection, and equivalence calculations. No hyperparameters or training runs are involved since the work is observational.
13. **Experiment statistical significance**
14. Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

    Answer: [NA]

    Justification: The paper does not present stochastic experiments requiring error bars or confidence intervals. It is based on reported or estimated fixed values.
15. **Experiments compute resources**
16. Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

    Answer: [NA]

Justification: The paper does not involve new model training. It compiles previously reported/estimated training emissions of existing models.

17. **Code of ethics**
18. Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics

    Answer: [Yes]

    Justification: The work conforms to NeurIPS ethics guidelines, as it uses publicly available reports and sustainability studies, cites all sources, and raises awareness of environmental costs.
19. **Broader impacts**
20. Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The discussion covers both positive impacts (raising awareness, promoting sustainability, informing creative practices) and negative aspects (large emissions, inequitable distribution of burdens).
21. **Safeguards**
22. Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release high-risk models or datasets, only emissions summaries and visualizations.
23. **Licenses for existing assets**
24. Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing models, datasets, and reports are properly cited with their original sources, licenses, or official technical reports (e.g., OpenAI, Meta, Stanford AI Index).
25. **New assets**
26. Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:

    Justification: The paper introduces a compiled dataset and interactive demo of emissions, with documentation describing data sources, calculation methods, and assumptions, along with references for transparency and reproducibility.
27. **Crowdsourcing and research with human subjects**
28. For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subject research.
29. **Institutional review board (IRB) approvals**
30. Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subject experiments are conducted.

31. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not part of the core methodology. Any LLM use was limited to writing/editing support, not central to the scientific contribution.