Measuring Energy Consumption of Deepfake Data Generation

Sean Cheng
Tufts University
sean.cheng@tufts.edu

Darien Parris
Tufts University
darien.parris@tufts.edu

Subhanga Upadhyay
Tufts University
subhanga.updahyay@tufts.edu

Abstract

Modern ML models used for deepfake generation require a ton of compute power and energy. We present a study that quantifies the energy use of two deepfake image generation models - StyleGAN2 ADA and BigGAN - to estimate the rate of carbon emissions during model inference. We demonstrate that there is a considerable impact on the environment when using these large models and convey the ethical implications of continued usage. Furthermore, we found non-trivial variability between GPUs and architectures when it comes to carbon emissions: for example, when generating 50 images over 100 batches with the BigGAN model, there's a difference of **0.347901** grams of CO_2 between the 256 and 512 versions of the model. We discuss the need to research more energy efficient innovations in the space of AI development to protect the environment for future generations.

1 Introduction

The rise of deepfake technology has sparked serious ethical concerns, from its potential to be misused to spread misinformation, create non-consensual explicit content, and manipulate public opinion, leading to ethical dilemmas around consent, privacy, and trust. The technology's ability to produce convincing fake content poses threats to security, as it can facilitate identity theft, cyber fraud, and political destabilization (Chapagain et al., 2024). While these social implications are widely discussed, an often-overlooked dimension is the environmental cost of generating deepfake content. These issues are deeply concerning because their circulation has been found to have doubled every year since 2019 and, some projections have estimated that the number of deepfakes created could reach millions(Jacobson, 2024).

The data does not lie, and, given the forecasts, it is a good idea to get a grasp on the baseline environmental impact of these technologies, especially in the context of assigning accountability. As generative models become more complex and widely deployed, they demand increasing amounts of computational power-raising questions about their carbon footprint and long-term sustainability. Focusing on the environmental costs of inference seems most logical to us because:

1. Evidence from literature corroborates the logic that inference is the primary consumer of energy
2. Inference creates most CO_2 emissions, electricity, and water usage
3. The actual generation of deepfakes can be more easily measured as compared to the training process, as the latter would need more comprehensive data from the model creators, how they tuned hyper parameters, where they deployed their models, among other factors [1]

In this project, we will seek to answer the question of how the choice of generative architecture and GPU model affect the energy consumption and environmental impact of deepfake content generation. We expect different deepfake generation models to exhibit varying levels of energy consumption due to differences in architectural complexity and inference speed. We hypothesize that the StyleGAN2 ADA model will produce lower levels of carbon emissions because it is slightly smaller and newer than the BigGAN model. We also believe that when comparing emission levels on different GPUs, the newest A100 GPU will produce the least amount of carbon emissions because we hope to find that technology has been increasing in efficiency and power over time.

[1]The specific data centers on which the models are hosted, the location, and other factors all contribute to carbon emissions and WUE/PUE (Water/Power Usage Effectiveness) values of the trained model

Once trained, inference can be done on a model any number of times to create deepfakes. Therefore, by quantifying the impact of inference for deepfake generation models, we aim to highlight how architectural choices impact environmental costs and how such insights could promote more sustainable AI development practices.

2 Literature Review

2.1 Holistically Evaluating The Environmental Impact Of Creating Language Models

In their paper, (Morrison et al., 2025)[2] examine the environmental footprint associated with developing large language models (LLMs). They assess factors such as carbon emissions and water consumption across various stages, including model development, training, and hardware manufacturing. Their findings reveal that creating a series of models resulted in 493 metric tons of CO_2 emissions and consumed approximately 2.769 million liters of water, with model development accounting for about 50% of the total impact. The study also highlights significant fluctuations in power usage during training, ranging from 15% to 85% of hardware capacity, which poses challenges for energy infrastructure planning. However, the research has limitations, including potential underestimation of total power consumption due to measuring only GPU usage and not accounting for other components like CPUs and cooling systems. Additionally, the study assumes that the environmental impact of hardware manufacturing can be accurately amortized over the hardware's lifetime, which may not capture the full complexity of production emissions. Future work could involve more comprehensive measurements encompassing all data center operations and explore strategies to mitigate the environmental impact of LLM development.

2.2 Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model

(Luccioni et al., 2023)[3] analyze the environmental impact of developing, training, and deploying the BLOOM language model. To account for all sources of greenhouse gas emissions, they use a standardized measurement called carbon dioxide equivalents (CO_2eq). They estimate that the dynamic power consumption of the final training phase emitted approximately 24.7 tonnes of CO_2eq, and an additional 25.8 tonnes to a total of 50.5 tonnes when accounting for embodied emissions and idle consumption. The study also examines the energy requirements and carbon emissions during BLOOM's real-time inference deployment via an API, which they calculated to be approximately 19 kgs of CO_2eq per day. The authors discuss the challenges of accurately estimating the carbon footprint of machine learning models and suggest directions for future research to improve carbon emissions reporting. The authors acknowledge the difficulty in precisely quantifying the carbon footprint due to factors such as varying energy sources and the complexity of tracking all contributing processes. They highlight the need for standardized methodologies in reporting emissions and suggest that future work should focus on developing more accurate tools and frameworks for assessing the environmental impact of large-scale AI models.

2.3 Green AI

(Schwartz et al., 2020)[4] aim to emphasize the importance of energy efficiency and environmental considerations in artificial intelligence research which they refer to as "Green AI". In this paper, they highlight the rapid increase in computational resources required for state-of-the-art AI models or "Red AI", leading to significant carbon footprints and accessibility issues. They point out that a linear gain in model performance requires an exponential increase in computational power, which is not sustainable. The authors propose incorporating efficiency as a primary evaluation criterion alongside accuracy, suggesting metrics like Floating Point Operations (FPO) to quantify computational cost. They argue that prioritizing efficiency can make AI research more environmentally friendly and inclusive, enabling broader participation from researchers with limited resources. They discuss that future work could focus on developing and validating specific methodologies for reducing computational costs without compromising model performance, as well as creating standardized benchmarks for evaluating the energy efficiency of AI models.

[2] https://arxiv.org/pdf/2503.05804

[3] https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf

[4] https://dl.acm.org/doi/pdf/10.1145/3381831

2.4 Energy and Policy Considerations for Deep Learning

(Strubell et al., 2020) analyze the financial and environmental costs associated with training large neural network models in natural language processing (NLP). They estimate that training a single large model can emit approximately 626,000 pounds of CO_2, comparable to the lifetime emissions of five average American cars. The authors highlight that the substantial computational resources required for such models not only have significant environmental impacts but also create barriers to entry for researchers with limited access to these resources. They advocate for increased transparency in reporting energy consumption and carbon emissions in AI research and suggest prioritizing more computationally efficient models to promote sustainability and equity in the field.

2.5 Deep learning for deepfakes creation and detection: A survey

(Nguyen et al., 2025) delves into the deep learning techniques employed in generating and detecting deepfakes. The authors categorize deepfake generation into three primary types: face-swapping, lip-syncing, and puppet-mastering. Face-swapping involves replacing a person's face with another person's face in a video, lip-syncing modifies mouth movements to match a different audio track, and puppet-mastering animates a target's facial expressions and movements based on another individual's actions. These techniques predominantly utilize Generative Adversarial Networks (GANs) and autoencoders, to produce highly realistic synthetic media. The survey highlights how advancements in these models, combined with the accessibility of large datasets and computational resources, have made deepfake creation more attainable for a broader audience. They also emphasize the need for ongoing research to develop more robust detection methods and to understand the societal implications of deepfake proliferation. The rapidly evolving nature of deepfake technology means that assessing the real-world impact of deepfakes necessitates continuous research and adaptation in this field.

2.6 LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models

(Faiz et al., 2024) created a new carbon footprint modeling library for LLMs that extends the carbon footprint analysis that previous libraries (such as mlco2) focused on. That is, previous work had not factored in model architectures or what the authors call *embodied carbon footprint*, a measure that looks at the carbon costs generated by the creation of semiconductor chips for the models, transport to the data center, and other costs unrelated to operation.

On case studies done by Faiz et al, it was observed that simply running parallelization techniques so often done at data centers running these big models leads to a decrease in operational carbon usage. Furthermore, picking a model type also is a big factor, so are the architectural pieces, etc: **MoE LLMs can obtain a lower test loss with the same training carbon footprint**, a big conclusion because usually a lower test loss requires more parameters, training, and tuning, which generally increases carbon footprint.

Although this paper does experiments and analyses **on LLMs**, these techniques are model-agnostic and could be equally applicable to the GANs or diffusion models that Deepfakes are created by. As a final push, the recommendation is also to take care to calculate embodied costs of AI more so than operational costs because companies have the mechanisma (and already are) investing in green data centers (Meta data centers have made 97% of their operational energy usage green): focusing on embodied costs is better because these will dominate net energy costs for LLMs in the future.

2.7 From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference

In this paper, LLMs are once more the subject of discussion; however, the focus shifts towards inference, making (Samsi et al., 2023) a noteworthy paper to analyze. Through in-depth analysis, the authors prove that LLM inference is more costly than training and yet receives much less attention from organizations and experts creating benchmarks for sustainable AI.

The authors conduct experiments with the 65 billion parameter version of LLaMA. Primarily, the authors stuck to using nvidia-smi for monitoring the GPU usage as well as NVIDIA DCCM for energy and power draws. It was discovered that the sharding strategy for these models make quite a big difference in terms of energy usage. It's recommended that GPU capping should be more wide-spread as a strategy to constrain model inferences.

2.8 Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

In this fascinating paper, (Li et al., 2025) talk about the water footprint of AI, something that is completely novel in terms of researchers discussing it yet hugely important. The demand and use of AI as it has exploded and will continue to do so for the foreseeable future, is projected to account for 4.2-6.6 billion cubic meters of water withdrawal in 2027, which is the amount 406 Denmark's and half of the UK would need in a year.

The authors continue to explain that such water usage arise from similar models that track AI carbon emissions: scope-1, scope-2, and scope-3 water consumption. When distilled, one can think of these water costs as the cost of cooling data center servers that heat up during huge amounts of inference. These are also the lesser thought about factors like water required for AI chip and server manufacturing water usage. These need to be factored in too because if these LLMs didn't exist, we wouldn't have produced all these GPUs and additional servers they need.

Furthermore, many different company and high profile reports are utilized to weigh numbers in terms of power draws, WUE and PUE metrics (as defined in footnote #2), in combination to the actual model parameters and types themselves, so the analysis was very complete and in-depth. One interesting conclusion made was that there could be a tradeoff between water and CO_2 footprint, as one depends on the other. However, it should be possible to find a better balance than what's being done now. In the context of our paper, we will try to include these alternative costs wherever possible, as these are wholly dependent on third-party organizations who collect these metrics.

3 Experiment & Results

To test our hypothesis, we use the CodeCarbon library (Courty et al., 2024), an open-source python package that allows us to track the equivalent

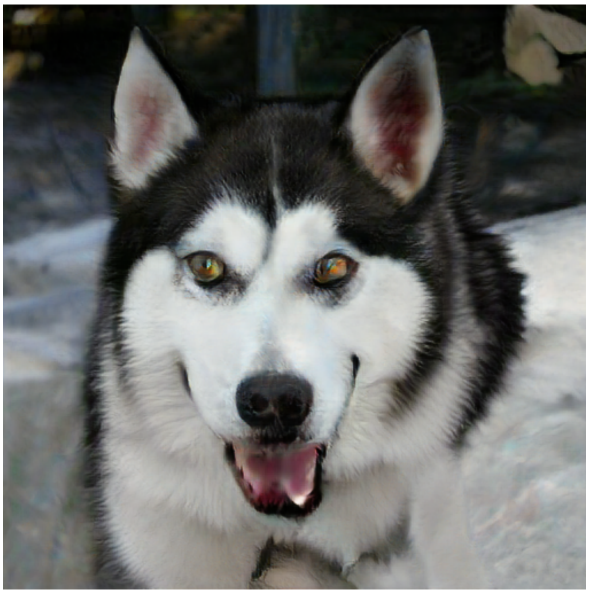

Figure 1: A sample image generated by BigGAN

amount of CO_2 emissions from running a model. The CodeCarbon library generates useful metrics such as emissions, emissions_rate, gpu_energy by sampling the local machine's power draws combined with publicly available power grid data to estimate data center energy usage. In order to account for all sources of greenhouse gas emissions (e.g. methane, nitrous oxide, etc.), the measurement is standardized to a single unit based on the level of global warming potential compared to that of carbon dioxide (Luccioni et al., 2023). The carbon output depends on the source of electricity, for instance, renewable energy sources emit less carbon into the atmosphere than natural gas or coal power plants.

Our experiments were performed on the Tufts University high-performance computing (HPC) cluster with a Linux operating system, located at the Massachusetts Green High Performance Compute Center (MGHPCC). The MGHPCC itself is dedicated to using green energy to power its systems, and has been awarded with the LEED Platinum Certification, the highest level award by the Green Building Council's Leadership in Energy and Environmental Design Program(Center, 2025). Although we do not know the exact mix of electricity sources for this facility, this location has been designed to yield 30% less CO_2 equivalent emissions than standard and is a state-of-the-art facility in power usage effectiveness. As a result, all our emissions measurements are likely an underestimation compared to running this experiment on other

Figure 2: A sample image generated by StyleGAN2 ADA

GPU model	Memory
Tesla V100	32GB
Tesla P100	16GB
NVIDIA A100	80GB

Table 1: These GPU models were readily available within the Tufts HPC and configurable with our version of CUDA.

high-powered clusters.

The HPC allows us access to multiple GPU models to run labor-intensive deep learning image generation models on various different GPU models and to specify the number of nodes. Table 1 shows the GPU models we tested on, and the respective memory allocations. We also try running model inference with 1-4 V100 GPU nodes to check if adding nodes changes the amount of energy consumed.

We chose recently developed image synthesis models that are popular in the literature and provide publicly available pre-trained models for testing. The BigGAN PyTorch model developed by Google was released in Feb. 2019 and contains 56M parameters (Brock et al., 2019). On the other hand, the StyleGAN2 ADA PyTorch model developed by NVIDIA was released in Oct. 2020 and contains 54M parameters (Karras et al., 2020). We estimated that, through the BigGAN experiments, we'd find that deeper model architectures and synthesis of images with higher dimensions would draw more carbon emissions.

BigGAN and StyleGAN2-ADA are both advanced Generative Adversarial Networks (GANs). These are a class of machine learning frameworks designed to generate realistic data samples. They consist of two neural networks: the generator and the discriminator, which are trained simultaneously through a process of competition. The generator creates fake data samples, while the discriminator evaluates them against real data, in order to distinguish between the two. The generator's goal is to produce data that is so convincing that the discriminator cannot tell that it is fake. Over time, as both networks improve, the generator becomes adept at creating highly realistic data. This adversarial training process enables GANs to generate images, music, and other types of data that closely mimic real-world examples, making them powerful tools in various applications, including deepfake content generation(Patel et al., 2023).

BigGAN[5] is an extension of the original GAN architecture, designed to generate high-resolution images with improved quality and diversity. It is trained on ImageNet and uses larger batch sizes, more layers, and a higher number of parameters, incorporating class-conditional generation to produce images based on specific categories. This results in highly detailed and diverse images, but requires substantial computational resources and training data. An example image generated by BigGAN is shown in Figure 1. Using the pytorch-pretrained-BigGAN library, we chose the *biggan-deep-512* - which generates images at 512×512 pixels.

StyleGAN2-ADA[6], on the other hand, builds on StyleGAN2 by introducing Adaptive Discriminator Augmentation (ADA), which dynamically adjusts augmentation probabilities based on the discriminator's feedback. This makes StyleGAN2-ADA more robust to overfitting and enhances image quality, especially when training data is limited. While BigGAN excels in generating diverse images across various categories, StyleGAN2-ADA is renowned for producing photorealistic images with fewer artifacts and maintaining high quality even with constrained datasets. A sample image produced by StyleGAN2 ADA is shown in figure 2.

(Luccioni et al., 2023) argues that the carbon footprint life cycle of a model begins during the extraction of raw materials for hardware manufacturing. Although other segments of the model life

[5]https://github.com/ajbrock/BigGAN-PyTorch

[6]https://github.com/NVlabs/stylegan2-ada-pytorch

cycle, such as training, may require a large amount of energy, these stages tend to occur in limited intervals. In this paper, we focus on the energy use of model inference, which has been shown to be the bulk of the energy consumption within a model's life cycle. A deepfake content generation model may take a year to develop and train but can be used to synthesize deepfakes for a decade or more. Therefore, our study aims to quantify the power consumption of computationally demanding models such as StyleGAN2 ADA and BigGAN to reveal the true impact of the generation of deepfake content on our environment. All the code for our experimentation is provided in our Github repository[7].

We used both pre-trained models to generate 50 images of size 512 x 512, and calculate the average emissions across 100 epochs.[8] We tested the StyleGAN ADA model on all 3 GPU models, but the BigGAN model was only tested on the Tesla V100 and P100 GPUs because it is incompatible with the newer A100 GPU. In order to test the StyleGAN ADA model on the Tesla V100 and P100 GPUs, we used Python 3.6.8 with torch 1.10.2, but the NVIDIA A100 GPU required Python 3.10 with torch 1.13.1+cu117. Both environments were equipped with CUDA 11.7. For the BigGAN model, we tested using Python 3.6.8, torch 1.0.1 and CUDA 11.7.

Table 2 shows the average emissions between the different ML models and GPU models to generate 50 images. For comparison, humans expel about 1kg of CO_2 per day (Withers, n.d.), which is equivalent to 6.95×10^{-4} kg CO_2 per minute. It was unexpected to find that the StyleGAN2 ADA model and the A100 GPU actually produced much more carbon emissions for the same amount of work. Additionally, there was much more variance in the carbon emissions generated by BigGAN-512 on the P100 GPU (0.016856 kg), as compared to V100 (0.000921 kg). The P100's per-batch carbon variance is over eighteen times greater than the V100's. That's nearly a 20× difference in volatility - more than a full order of magnitude in how wildly each run can swing on the older hardware.

This could possibly be due to the P100 GPU being older than V100, which once again showcases the big differences model architectures and even GPUs can make.

GPU model	StyleGAN2 ADA	BigGAN
V100	6.42×10^{-4}	3.96×10^{-4}
P100	6.95×10^{-4}	5.59×10^{-4}
A100	1.17×10^{-3}	N/A

Table 2: The average kg CO_2eq emissions for generating 50 images using different GPU and ML models. The NVIDIA A100 GPU with StyleGAN2 ADA had the highest overall emissions, and V100 GPU with BigGAN had the lowest overall emissions. For both the V100 and P100 GPUs, the BigGAN model produced less emissions.

# of nodes	StyleGAN2 ADA	BigGAN
1	6.42×10^{-4}	3.96×10^{-4}
4	1.07×10^{-3}	6.48×10^{-4}

Table 3: The average kg CO_2eq emissions for generating 50 images using different number of V100 GPU nodes and ML models. The use of 4 GPU nodes increases the overall emissions for both models by about 165%.

Another interesting finding was that the RAM power metric remained exactly constant throughout our many experiments. This wasn't the case for other similar metrics, such as CPU or GPU power. This likely is because of how CodeCarbon models RAM draws; in practical life, RAM variability does exist, but literature states that its effects are small compared to elative to GPU/CPU swings.

Furthermore, we analyzed data collected from running the same experiment with different V100 configurations shown in Table 3. An interesting finding is that the configuration with 4 V100 GPUs produced about a gram of CO_2 and the newer A100 GPU produced a bit more CO_2 . The A100 seems to be more efficient in its energy consumption using half of the wattage required to power 4 V100s, but they both took about 30 seconds to generate 50 images, succinctly adding more nodes for the same amount of work emitting more emissions.

4 Discussion & Ethical Implications

Deepfake technology presents both positive and negative aspects. On the positive side, deepfakes offer innovative applications in entertainment, education, and accessibility. They enable the creation of realistic visual and audio content, enhancing virtual experiences and providing new tools for creative expression(Chapagain et al., 2024).

However, the negative aspects are significant and concerning. These issues are compounded when

[7] https://github.com/dp0920/cs239_deepfake

[8] We generated 254x254 images generated as well, in the case of BigGAN. This allowed us to test how significantly image size affects carbon emissions.

one takes into consideration that deepfakes (typically generated via GANs) have doubled every 6 months since 2019 and have been forecasted to continue to do so. Besides disproportionately harming members of vulnerable communities, they have noteworthy impacts on the environment, not just limited to carbon emissions and electricity, but also water consumption (Li et al., 2025). Balancing the benefits and risks of deepfake technology requires robust countermeasures and ethical guidelines to mitigate its potential harm while leveraging its positive applications.

The generation of deepfake content, particularly through resource-intensive models like GANs which are pivotal in creating realistic synthetic media, raises significant ethical concerns regarding its environmental impact. These models requires substantial computational power for inference, predominantly sourced from fossil fuels. This leads to considerable greenhouse gas emissions, contributing to climate change. Ethically, this poses a dilemma as the benefits of deepfake technology in entertainment, research, and creative industries must be weighed against its environmental footprint.

The responsibility to mitigate these emissions falls on both developers and users, urging the adoption of more energy-efficient algorithms, renewable energy sources, and carbon offset initiatives. It is equally important to also be mindful about other meta choices, which one wouldn't usually attribute importance to, such as the cloud providers we're utilizing, model architectural patterns and their second order effects, or having a dedicated carbon emissions comparative analysis done to check how one's model fares against other baselines.

Additionally, the broader implications of climate change, such as its disproportionate effects on vulnerable communities and future generations, highlight the need for sustainable practices in AI development. Balancing innovation with environmental stewardship is crucial to ensure that technological advancements do not come at the expense of our planet's health.

5 Conclusion & Recommendations

In this study, we aimed to quantify the environmental impact of deepfake content creation. As the use of AI deepfake technology grows, we must remain cognizant of the effects on climate change. We use an emissions tracking library called CodeCarbon to measure the amount of CO_2 equivalent emissions when running our deepfake image generation models. We test two GAN models, StyleGAN2 ADA and BigGAN to show how much energy gets used when creating 50 images of size 512 x 512. We test on various types of GPU models with various amount of GPU nodes to find the optimal combination with the lowest emission rate.

Due to time and resource constraints, we limited the scope of our experiments to deepfake image generation. We leave it to future work to experiment with deepfake video and audio generation. There was also a significant amount of models built for deepfake detection, so this would be an interesting avenue for future work to discover the carbon emission of deepfake detection models, as those become more prevalent.

We advocate for a greater awareness and transparency of the environmental impacts of ML. While these models we use are currently at the forefront of deepfake image synthesis, newer more powerful models will eventually take over that will require more energy to perform inference. We recognize the complexity of conducting empirical studies to test energy consumption and carbon emissions. Our results are highly specific to the computing center that houses the university cluster. We hope that new model developers will take into account these ethical concerns and work towards building more energy-efficient AI algorithms.

References

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Massachusetts Green High Performance Computing Center. 2025. Green design.

Devendra Chapagain, Naresh Kshetri, and Bindu Aryal. 2024. Deepfake disasters: A comprehensive review of technology, ethical concerns, countermeasures, and societal implications. In *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*.

Benoit Courty, Victor Schmidt, Sasha Luccioni, GoyalKamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, and 11 others. 2024. mlco2/codecarbon: v2.4.1.

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. *Preprint*, arXiv:2309.14393.

Neill Jacobson. 2024. Deepfakes and their impact on society.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. In *Proc. NeurIPS*.

Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2025. Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models. *Preprint*, arXiv:2304.03271.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Jacob Morrison, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell, and Jesse Dodge. 2025. Holistically evaluating the environmental impact of creating language models. In *Proceedings of the 2025 Conference of the Association for Computational Linguistics (ACL)*.

Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. 2025. Deep learning for deepfakes creation and detection: A survey. *Journal of Artificial Intelligence Research*.

Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvala, and Vrince Vimal. 2023. Deepfake generation and detection: Case study and challenges. *IEEE Access*, 11:143296–143312.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696.

Nikki Withers. n.d. How much does human breathing contribute to climate change?