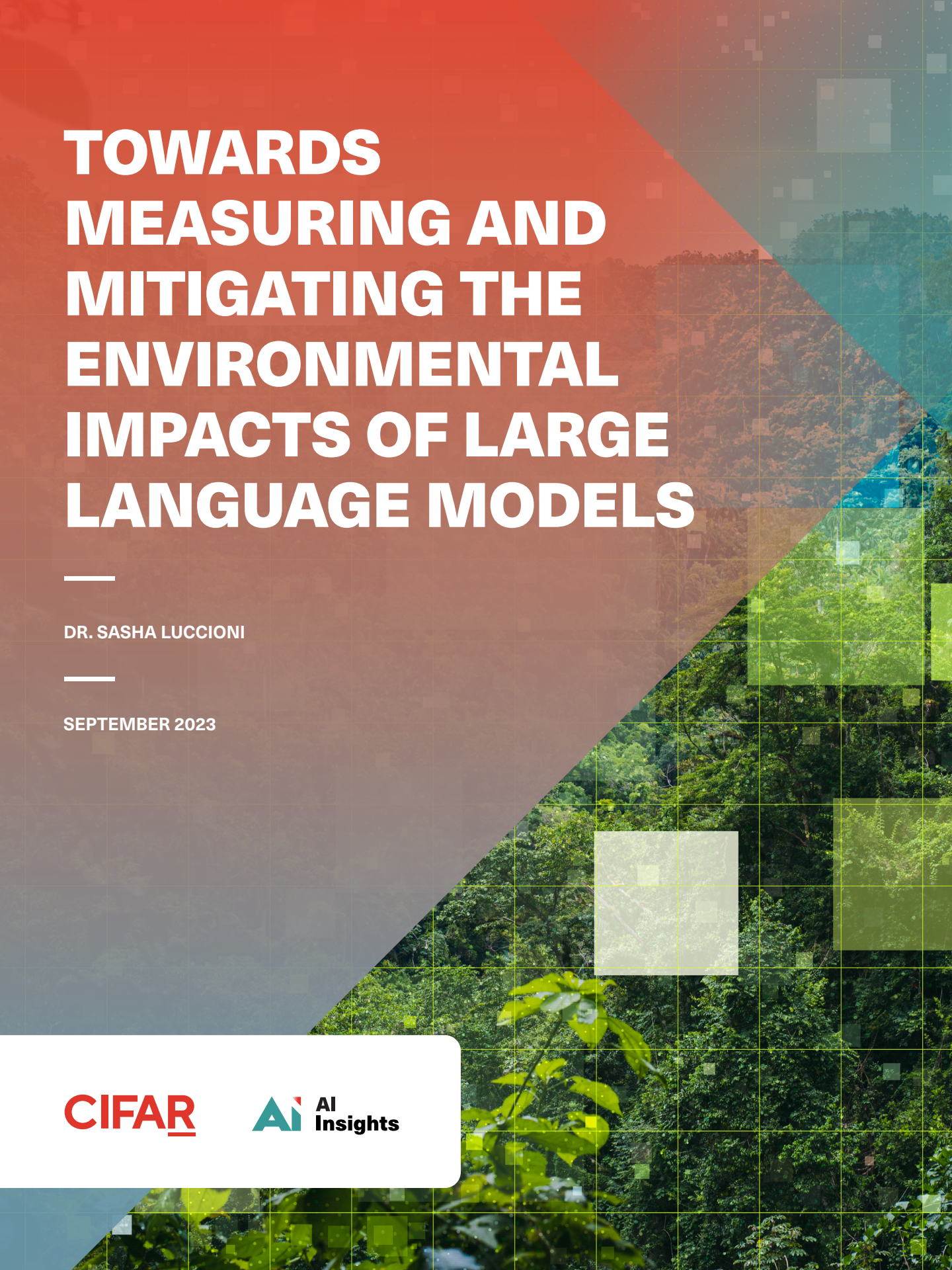
TOWARDS MEASURING AND MITIGATING THE ENVIRONMENTAL IMPACTS OF LARGE LANGUAGE MODELS

DR. SASHA LUCCIONI

SEPTEMBER 2023



CIFAR

AI Insights

THE AUTHOR

DR. SASHA LUCCIONI

Photo Credit: Joseph Viviano

Climate Lead, Hugging Face
sasha.luccioni@huggingface.co

LAND ACKNOWLEDGMENT

We wish to acknowledge this land on which CIFAR operates. For thousands of years it has been the traditional territory of many nations including the Mississaugas of the Credit, the Anishnabeg, the Chippewa, the Haudenosaunee and the Wendat peoples and is now home to many diverse First Nations, Inuit and Métis peoples. We are grateful to have the opportunity to work on this land. We also acknowledge we are all responsible for reconciliation. CIFAR's AI & Society program seeks to advance our understanding of the societal implications of AI to design a future of responsible AI. A future of responsible AI includes one that centres the concerns of Indigenous communities. CIFAR is committed to prioritizing Indigenous perspectives in the development and design of responsible AI.

DISCLAIMER

This brief expresses the opinions of the author only; it is informed by their previous research as well as by research from other sources, which are cited in the references.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In recent years, large language models (LLMs) have been exponentially growing in size, with recent large models like GPT-3[1] and OPT[2] spanning hundreds of billions of parameters. Training models of this size typically requires millions of hours of computation, which consumes large amounts of energy and emits many tonnes of carbon emissions in the process. Despite this, few LLM releases publicly share the environmental cost of model training or the logs necessary to replicate these environmental cost evaluations. Additionally, other carbon-intensive processes that are part of the broader model life cycle are underexplored, such as manufacturing the specialized hardware necessary for training models, as well as the computational requirements (and ensuing emissions) of their large-scale deployment and maintenance. Going forward, establishing standards around the environmental impacts of LLMs and requiring more transparency from model creators are key steps towards ensuring model users can make more informed decisions.

1.0 INTRODUCTION

Artificial intelligence (AI) has been one of the biggest technological advances in recent years, with fundamental breakthroughs made in industry and academia being used in a multitude of consumer-facing applications and tools ranging from Web search to smart devices. The most recent AI models, such as GPT-4, have gained considerable popularity due to their generative capabilities, being used tens of millions of times within the first months of their launch.

In order to better comprehend the sustainability of these technologies, it is useful to understand how these models are created and deployed. In the current brief, we will start with a short description of the evolution of AI models across the last decades and the paradigm shift that has taken place with the advent of generative models in recent years. We will then present the different steps in the life cycle of generative models, from the manufacturing of GPUs used for their training to their deployment in products and services. This will be followed by a presentation of recent findings regarding the carbon footprint of different kinds of AI models as well as the pre-training and fine-tuning processes. We will conclude with recommendations on steps to be taken to ensure the sustainability of AI models in research and development.

[1] Brown, T. et al., (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[2] Zhang, S. et al., (2022) OPT: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068

1.1

THE EVOLUTION OF LARGE LANGUAGE MODELS

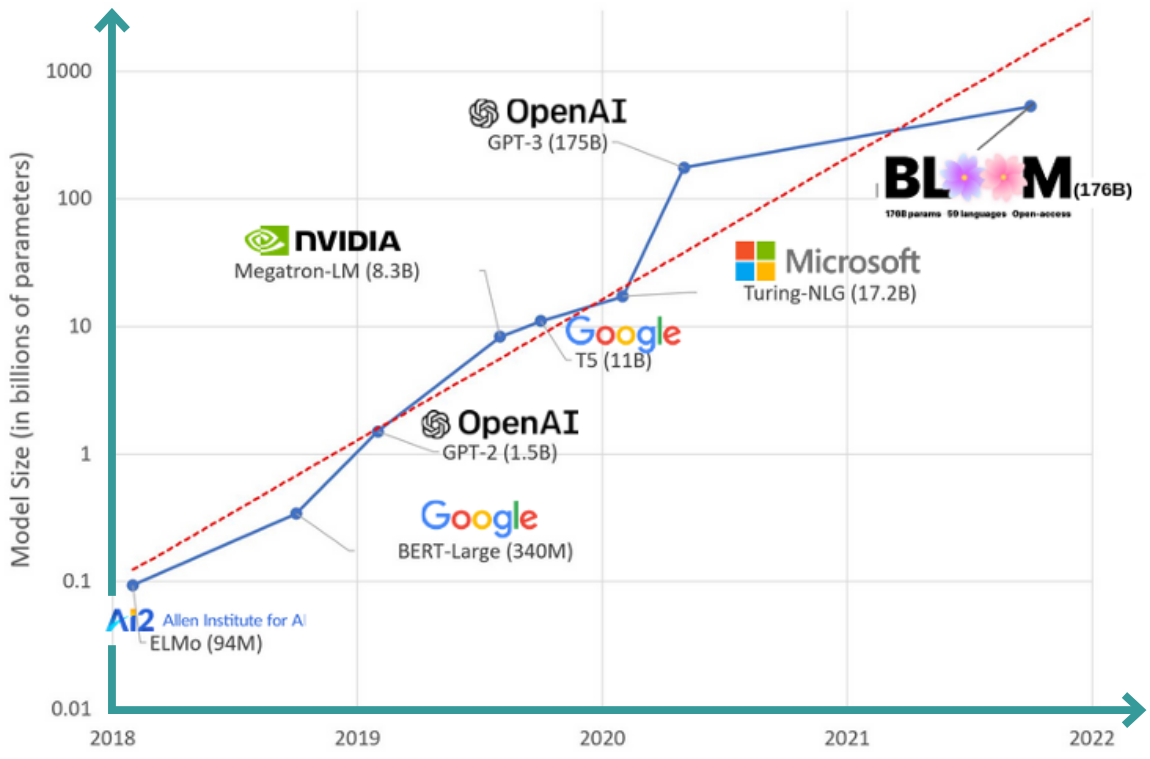
As a starting point, it is important to define what Large Language Models (LLMs) are, how they work and how they are trained. LLMs (also called foundation models[3]) are Transformer[4]-type neural networks that are trained on large amounts of text, allowing them to build statistical models of language. In practice, what this means is that given a sufficiently large amount of general text data (for instance, all of Wikipedia), LLMs will learn to predict the most probable word based on an input text, e.g., 'the cat sat on the _____' will result in the LLM returning 'chair' or 'mat'. This is called the pretraining step, which can be followed by a fine-tuning step (e.g., to adapt an LLM to a specific domain by training it on data from that domain), or an instruction step (e.g., to train an LLM to better respond to human feedback and instruction). The resulting LLM can then be deployed in an AI tool or system, for instance one such as ChatGPT, or be queried directly via a user-interface or program.

Recent years have seen a steady increase in the size of LLMs. This is often measured in the number of parameters (i.e., connections) that they contain, which are distributed across thousands of layers that constitute the model. Whereas a few years ago, these were measured in the millions – for instance, BERT-Large, a LLM released by Google Research in 2018, had 340 million parameters, whereas BLOOM, the first open-access LLM trained collaboratively by the Big Science project and released in 2022, had 176 billion parameters – over five thousand times bigger in 4 years.

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[4] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

FIGURE 1



The evolution of Large Language Model size over time, from 2018 to 2022

This increase in size did not happen in isolation but was accompanied by improvements in hardware as well. Each new generation of Graphical Processing Units, or GPUs, which are specialized computer chips used for training LLMs, is more powerful than the last, not only in terms of speed but also increased connections between the different layers of the GPU. This last point is particularly important for LLMs since they involve billions of calculations at each pre-training step, and parallelization makes it possible to update the model's internal representation concurrently across multiple model layers at once. The most recent generations of LLMs are trained for thousands of hours on thousands of GPUs, resulting in many scholars to worry about the sustainability of such computational demands in the long term.[5]

Another important contribution towards the growth and development of LLMs has been the increased access to collections of data large enough in size to enable their training. For instance, whereas the BERT-Large model trained in 2018 used approximately 16 GB of training data consisting of books and Wikipedia, newer generations of models have been trained largely on Web-scraped data such as that from the Common Crawl, which is 400 TB in total. This data is complemented with data gathered from human feedback, and all of it contributes towards the quality of model generations. And alongside model size and hardware improvement, data has been an important catalyst in the recent progress in large language models. But this progress also comes with a price in terms of environmental impacts, which we describe in more detail below.

[5] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. arXiv preprint arXiv:2007.05558.

2.0

FINDINGS

CALCULATING THE CARBON EMISSIONS OF LLM TRAINING

Most of the carbon footprint carried out in the field of AI to date has focused on the emissions produced by generating the energy necessary for training models. This is the most straightforward estimation, since model training has a well-defined start and end time, even if it can take weeks and even months for a model to train. There is no single agreed-upon methodology in the field, and estimates vary depending on the factors taken into account by model creators. Generally speaking, the relevant factors are:

1. The model training time - measured in hours.
2. The power used by the hardware used to train the model.
3. The carbon intensity of the energy grid used for training.

Multiplying these three factors provides an initial estimate of what is called 'dynamic power consumption', since it only accounts for the energy consumed by the hardware running model training, but not any of the overhead energy costs (e.g., datacenter cooling, storage, networking, etc.).

This number is sometimes reflected by the PUE (Power Usage Effectiveness) of datacenters, which is an average factor that will vary depending on the efficiency of the server instance where training took place. PUE is taken into account in some estimations of LLM carbon footprints (e.g., T5[6]), but not others (e.g., OPT).

The first article to carry out this kind of estimation, by Strubell et al[7]., estimated that the training of BERT, the first large language model, emitted approximately 284 tonnes of carbon dioxide equivalents (CO_{2eq}). When tallying up different sources of carbon emissions they are converted to a single unit of measure - carbon dioxide equivalents (CO_{2eq}). These are calculated by comparing the global-warming potential (GWP) of different greenhouse gases to that of carbon dioxide (CO_2) – e.g., methane has a GWP 25 times larger than that of CO_2, which means that 1 gram of methane is equal to 25 grams of CO_{2eq}. In recent years, there have been a number of further studies regarding both the energy consumption and carbon emissions of popular LLMs – we present the available results in the Table below, which compares the emissions of 4 recent LLMs: GPT-3, Gopher, OPT and BLOOM, in terms of their size, power consumption and CO_{2eq} emissions.

TABLE 1

MODEL NAME	MODEL YEAR	MODEL SIZE (# PARAMETERS)	POWER CONSUMPTION TRAINING	CO_{2EQ} EMISSIONS TRAINING
GPT-3	2020	175B	1,287 MWh	502 tonnes
Gopher	2021	280B	1,006 MWh	352 tonnes
OPT	2022	175B	324 MWh	70 tonnes
BLOOM	2022	176B	433 MWh	25 tonnes

Comparison between Large Language Models from recent years, in terms of model Size, power consumption and carbon emissions emitted during training.

We can see that while the size of these models has largely stayed the same over the last 2 years (with Gopher slightly larger than its contemporaries), their power consumption is actually decreasing, with OPT training consuming roughly 75% less energy compared to GPT-3. This is largely due to new generations of hardware, which have gotten more efficient and less energy-intensive (as noted by Patterson et al. in 2022[8]). The same can be said for the carbon emissions of models, which have gone down over 20 times between GPT-3 and BLOOM – this can be attributed to the energy mix used by either model, with GPT-3 largely being powered by coal and natural gas, whereas BLOOM used nuclear energy. This illustrates the potential of renewable energy in improving the sustainability of LLMs, which can have significant impacts on the overall carbon emissions of model training, even as model training time remains measured in the millions of GPU hours.

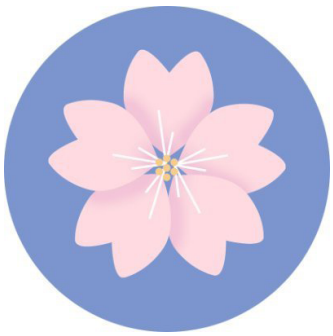
[6]Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

[7]Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650).

[8]Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), 18-28

CASE STUDY:

THE BLOOM MODEL

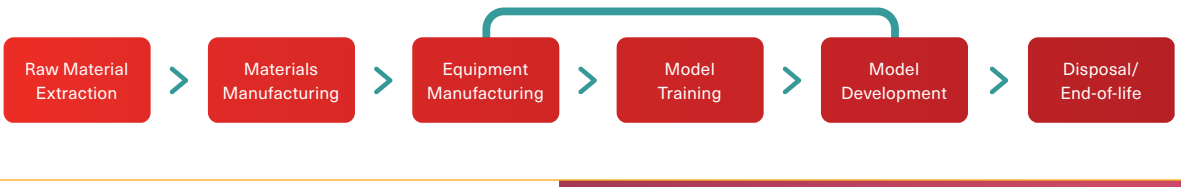


The BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) is a 176 billion parameter language model. It was trained on 1.6 terabytes of data in 46 natural languages and 13 programming languages as part of the BigScience Workshop, a year-long initiative that brought together over a thousand researchers from around the world. Thanks to a generous computing grant from the French government, BLOOM was trained over 3 months on over 1,000 GPUs, using over a million GPU hours of compute.

METHODOLOGY

Taking a step back from the LLM training described in the previous section, it is also worth carrying out a life cycle assessment (LCA)[9] of language models, which look at the resources they require in all stages of their life cycle. These resources include power usage during model training and deployment, but also the planetary impacts of mining the rare metals and plastic required for creating the hardware, and the water needed to cool data centers that run the models.

FIGURE 2



The phases considered under life cycle assessment of LLMs.

When it comes to estimating the carbon footprint of AI models, there is a lack of information when it comes to the various steps in models' lifecycles. For instance, in order to quantify the emissions engendered by manufacturing the equipment used for training LLMs (e.g., GPUs, servers, etc.), it is necessary to gather Scope 3 emissions data from the designers of this hardware, which means that they, in turn, need to gather this information from their suppliers. Because of the lack of data regarding the sources of emissions at different steps in an AI models' lifecycle, the BLOOM carbon estimation is the first of its kind to go above and beyond model training to also consider other parts of the LCA process.

[9]Klöpffer, W. (1997). Life cycle assessment: From the beginning to the current state. Environmental Science and Pollution Research, 4, 223-228.

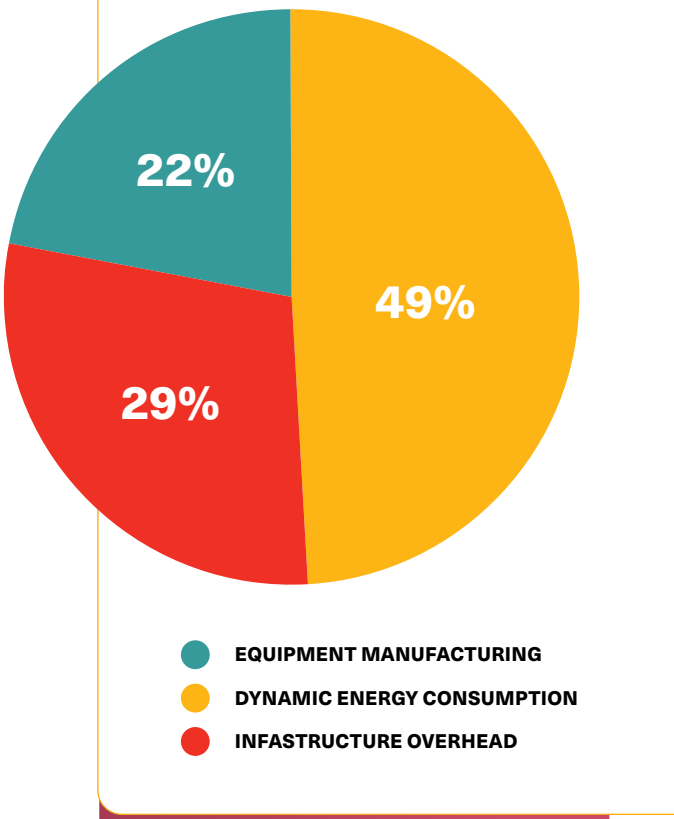
ESTIMATING MANUFACTURING EMISSIONS AND OVERHEAD

Based on our estimates, equipment manufacturing was responsible for 11 tonnes of CO_{2eq}, dynamic consumption from model training emitted 25 tonnes of CO_{2eq}, and the infrastructure overhead a further 14 tonnes, for a total of **50 tonnes of CO_{2eq}**.

In order to estimate the emissions produced by manufacturing the hardware used for training the BLOOM model, we based ourselves on estimates provided by hardware manufacturers such as HPE, since the exact figures for the Nvidia GPUs that were used for training were not provided by the manufacturer. Given that BLOOM training lasted a total of 1.08 million hours using, on average, 384 GPUs across 48 computing nodes, we can estimate that this produced approximately **11.2 tonnes of CO_{2eq}** to its carbon footprint. This assumes a replacement rate of 6 years and an average usage of 85%, which vary depending on how often hardware is updated and how intensely it is utilized. However, given the importance of hardware towards the success of LLM training, many organizations are buying tens of thousands of recent GPUs and discarding older, less efficient hardware versions, which adds a significant amount of emissions to their overall carbon footprint.

Furthermore, in order to calculate the emissions of different components of datacenter hardware, we ran a series of experiments, comparing the total energy consumption of devices on the computing cluster (e.g., network, GPUs, storage, cooling/heating and computation nodes) when they were idle to the total consumption of the same devices while they were dynamic, i.e., running the model training code. Based on this, we found that around 54% of the power consumption can be attributed to running model training, with the remaining 46% is used for keeping the computing nodes on. This means that the estimates presented above, which only focus on dynamic power consumption, fail to account for almost half of the overhead necessary to power the rest of the computing infrastructure. This figure will vary depending on the generation of hardware used and the efficiency of computing infrastructure, but given the size of LLMs, these are mostly trained in a distributed way across hundreds and even thousands of computing nodes, which brings with it added overhead costs. However, work pursued by Google Deepmind endeavors to reduce this overhead by employing AI, and has proven to reduce overhead costs by up to 40%.

FIGURE 3

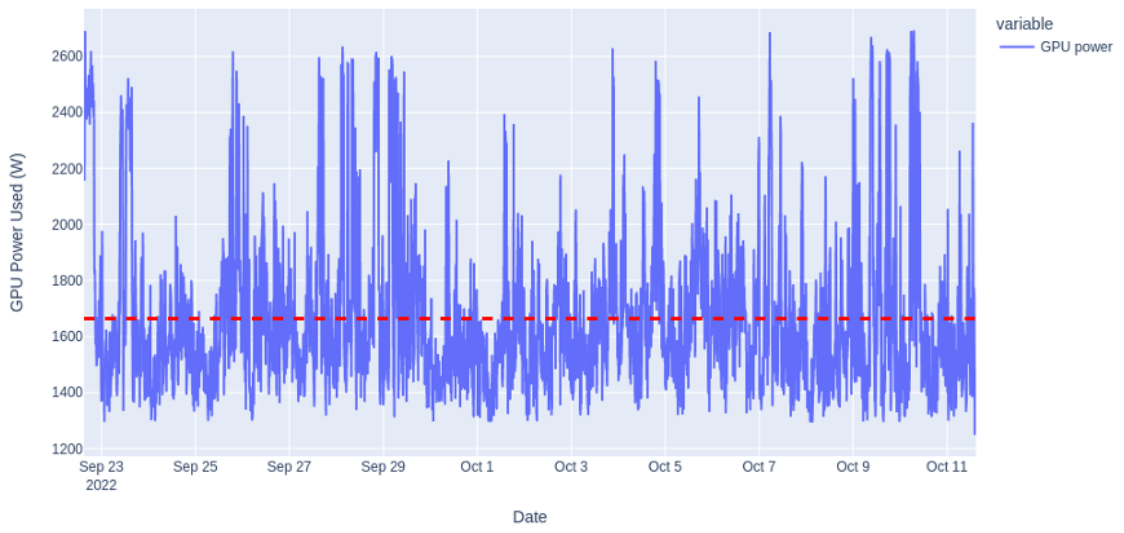


A comparison of the proportion of the carbon emissions from BLOOM model.

ESTIMATING THE EMISSIONS OF MODEL DEPLOYMENT

A 2019 article estimated that 80–90% of Nvidia's ML workload is model deployment, but the true numbers of energy usage and carbon emissions of model inference are largely unknown because of the distributed and dynamic nature of model deployment – at any given time, thousands of models can be deployed across multiple cloud compute instances, and scaled up and down depending on user demand. This means that to date, the major missing piece of LLM carbon footprint is that of model deployment, which was not covered by a single carbon footprint study. In order to estimate the carbon emissions incurred by deploying BLOOM, we deployed the model on a cloud compute instance and tracked the energy usage of the instance over a period of approximately 18 days. During this period, the model received an average of 558 requests per hour, which were handled in real time, for 230,768 requests in total. This represents a realistic use case of real-time LLM deployment in applications such as chatbots, where they are expected to respond to a constant, varying flux of user queries.

FIGURE 4



The fluctuation of mean power used to power the GPUs running the BLOOM model. With the mean power consumption in red (1664W) in dotted red.

As can be seen in the plot above during the 18-day period for which we carried out our analysis, the power consumed by the BLOOM model fluctuated between 1252 W to 2735 W, with the mean power consumption in red (1664W) in dotted red. In total, the instance used for the BLOOM model API consumed 914 kWh of electricity, which fluctuated depending on usage of the model. However, 75% of this amount was dedicated to maintaining BLOOM in memory, given that models have to always be ready to respond to user queries. Furthermore, given that the cloud instance that we used for deploying the BLOOM model has a carbon intensity of 394 gCO_{2eq}/kWh this resulted in approximately 19 kgs of CO_{2eq} emitted per day of model deployment, or 340 kg over the total period. While this may not seem like a huge amount compared to the 50 tonnes emitted during model training, it adds up when LLMs like BLOOM are deployed in user-facing applications like Web search and navigation, which can get queried millions of times a day and therefore require many instances of LLMs deployed in parallel to respond to user demand.

3.0

POLICY DISCUSSION

In order to continue responsible LLM innovation with sustainability in mind, the following elements must be considered:

1 Creating standards and frameworks for evaluating and reporting the carbon footprint of large language models: as illustrated in the findings, there is no single framework or methodology used for calculating the carbon emissions of LLMs, making it hard to meaningfully compare emissions from different models. While there are initiatives like the Green Software Foundation, which aim to develop standards for computing in general, there are no specific ones for AI in general and LLMs in particular, which need to take into account the specificities of these technologies.

2 Developing tools for accurate energy estimation: while tools such as Code Carbon and Carbon Tracker exist, the extent to which they reflect actual energy consumption is debated – recent work[10] has shown that the measurements they make vary depending on the type of hardware used, and do not reflect the true energy consumption of devices (when measured using physical devices). Also, none of the existing tools can adequately measure the energy consumption of model deployment since the time interval used for measuring consumption is not granular enough. Improving and calibrating existing tools and harmonizing them to correspond to standards such as those described above, is important to reflect energy consumption and carbon emissions more accurately across models and different steps of the model lifecycle.

3 Mandates for environmental impact with the release of LLM-based systems: the BLOOM model is the latest LLM for which we have sufficient information in order to meaningfully estimate its carbon footprint. Subsequent generations of LLMs, such as LLaMa[11], GPT-4[12] and PaLM 2[13], do not provide sufficient details about model training or architecture to attempt to estimate their carbon emissions. This is symptomatic of a general trend for recent generations of LLMs to be less permissive in terms of access and detail compared to previous ones (see Solaiman, 2022[14]). Mandating that model creators provide sufficient information regarding the environmental impacts of their models, for instance via tools such as model cards[15], is an important step towards transparency and accountability.

4 The creation of certification and ratings of AI models: there are currently no ways for model users to pick models based on factors such as efficiency and sustainability, given such information is not presented by model providers (see point above). However, many companies and organizations are taking ESG (Environmental, social, and corporate governance) into account when making business decisions and declaring their own emissions via ESG reports. Since the emissions of LLMs would constitute Scope 3 emissions – since they are the result of activities from assets not owned or controlled by the reporting organization – it is crucial that organizations be provided with the necessary information to estimate their contribution which contribute to their overall footprint, and to take this information into account when choosing between different LLM-based products and services.

5 Carbon-aware model training: Recent research[16] has shown that there are ways to reduce the carbon footprint of LLM training by factoring in the marginal carbon intensity of the energy used for training the model. This information is not readily available on most commercial cloud computing platforms, meaning that users cannot take it into account when choosing a compute instance. Tools such as the Microsoft Emissions Impact Dashboard can help provide more transparency into the emissions across different regions and resources, which can help guide both training and deployment.

6 Expanding renewable energy resources: As shown in the LLM comparison table above, the energy mix used for training LLMs can contribute to drastically reducing the training footprint (e.g., from over 70 tonnes of CO_{2eq} from training OPT to 25 from training BLOOM, ceteris paribus). Despite this, the biggest cloud computing servers globally are currently located in places whose energy grids are mostly powered by oil and natural gas[17], contributing to the high carbon footprint of training many LLMs. Training them in regions such as Quebec, which is powered by hydroelectric power, would bring down their emissions; however, for this to be feasible, more computing clusters need to be built in these regions.

[10]Bannour, N., Ghannay, S., Névéol, A., & Ligozat, A. L. (2021, November). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing (pp. 11-21).

[11]Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[12]OpenAI (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[13]Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). PaLM 2 Technical Report arXiv preprint arXiv:2305.10403.

[14]Solaiman, I. (2023, June). The Gradient of Generative AI Release: Methods and Considerations In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 111-122).

[15]Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[16]Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., ... & Buchanan, W. (2022, June). Measuring the carbon intensity of AI in cloud instances. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1877-1894).

[17]See Greenpeace Report: Oil in the Cloud (2020)

4.0
CONCLUSION

The environmental impacts of new generations of AI technologies such as LLMs are under-explored and often overlooked in research and practice. Recent research on the carbon footprint of BLOOM, a 176-billion parameter language model, showed that both the manufacturing of equipment, model training and deployment are responsible for non-negligible amounts of carbon emissions, with other models of similar size having even larger footprints. More research is needed to better understand the emissions of LLMs at different stages of their life cycle, including the manufacturing of equipment and deployment. For this to be possible, more transparency is needed in terms of the true emissions of different models and architectures, as well as the development of standardized frameworks and tools for measuring and certifying model's carbon emissions. As LLMs are deployed in more user-facing tools and applications, making informed decisions between factors such as sustainability and efficiency will become increasingly important.

CIFAR

MaRS Centre, West Tower
661 University Ave., Suite 505
Toronto, ON M5G 1M1 Canada

www.cifar.ca/ai