

The Environmental Impacts of Language Model Training Keep Rising Now is the Time to Catch Impacts on the Rebound

Clément Morand and Aurélie Névél
Université Paris-Saclay, CNRS, LISN
name.surname@lisn.fr

Anne-Laure Ligozat
ensIIE
Université Paris-Saclay, CNRS, LISN
name.surname@lisn.fr

Abstract

Recent Machine Learning (ML) approaches have shown increased performance on benchmarks at the cost of escalating compute demands. Hardware, algorithmic and carbon optimizations have been proposed to curb energy use and environmental impacts. We estimate the environmental impacts associated with training models documented in the Epoch AI database over the last decade, with a particular focus on impacts associated with Large Language Models and the hardware used to train them. We find that energy use and environmental impacts associated with training ML models have increased exponentially, even when considering impact reduction strategies such as using less carbon intensive electricity mixes or more efficient hardware. Optimization strategies do not mitigate the impacts induced by model training, suggesting rebound effect. We show that the impacts of hardware must be considered over the entire life cycle rather than the sole use phase in order to avoid impact shifting. Our study demonstrates that increasing efficiency alone does not ensure sustainability. There is an urgent need to systematically integrate environmental impacts in NLP evaluation practices to better inform the community and support the use of impact as a feature in research planning and decision making.

1 Introduction

The increasing ubiquity of [Large Language Models \(LLMs\)](#) is fueling debates in the [Natural Language Processing \(NLP\)](#) community who is questioning the risks and potential associated with global LLM use ([Bender et al., 2021](#); [Scialom et al., 2022](#); [Groeneveld et al., 2024](#); [Varoquaux et al., 2025](#)).

NLP evaluation uses community-developed tools including benchmarks and standardized performance metrics that aim for easy and reproducible comparison between models - as shown in leaderboards. Standard NLP evaluation methods over-emphasize task related performance at the

expense of user oriented metrics ([Ethayarajh and Jurafsky, 2020](#)) and real world impact evaluation of NLP applications ([Reiter, 2025](#)). The community has identified a need to rethink progress and evaluation practices. We believe that the systematic assessment of compute requirements and associated environmental impacts is a fundamental way forward. The impacts of [Machine Learning \(ML\)](#), including LLMs, come from many sources including model development, training and use ([Ligozat et al., 2022](#); [Wu et al., 2022](#)). Although model inference has received much attention with the wide use of LLMs, model training impacts are substantial compared to inferences ([Wu et al., 2022](#); [Berthelot et al., 2024](#)) and need careful assessment.

Research is underway to understand and develop assessment methods for the impacts of carbon emissions ([Strubell et al., 2019](#)), water usage ([Li et al., 2025](#)) and metallic resource depletion ([Morand et al., 2024b](#); [Berthelot et al., 2024](#)). However these studies are limited to individual models cases and, to our knowledge, no comprehensive analysis of LLM training impact trends is available.

Strategies have been developed to optimize the energy use of models training. *Hardware optimizations* leverage more recent energy-efficient hardware. *Algorithmic optimizations* adjust software to offer the same task performance using less compute. *Carbon optimizations* displace computation towards less carbon-intensive electricity (e.g., renewable energy). Major [Artificial Intelligence \(AI\)](#) companies have claimed these strategies would mitigate ([Wu et al., 2024](#)) or reduce the carbon footprint of ML ([Patterson et al., 2022](#)).

These claims overlook two important points. First, frequent hardware upgrades reduce the carbon footprint from the use phase at the expense of the production and end-of-life phases. They also increase metallic resource depletion, limiting the overall benefit. This phenomenon is called *impact shifting*. Second, optimization may result in

smaller-than-expected reductions, or even increase the overall environmental impacts because of *rebound effect*, a prevalent phenomenon in [Information and Communication Technology \(ICT\)](#) ([Gossart, 2015](#); [Bol et al., 2021](#)).

In this study, we address the following research question: "Do optimization strategies lead to lower impact for LLMs model training over time?".

In section 2, we review related work. In section 3 we detail the impact assessment method we combine with data from GPU manufacturers and the Epoch AI database to examine (in section 4) the environmental trajectory of ML model training over time. Finally, in section 5, we investigate the efficacy of carbon, hardware, and software optimizations in mitigating environmental impacts. Our analysis reveals that the increase in compute demand has outweighed optimization gains. Overall, optimizations are correlated with larger models, consistent with rebound effect. As carbon optimization strategies are insufficient to curb the growth in the environmental impacts of model training, we advocate integrating environmental impact into NLP evaluations to provide actionable data to guide community decision making.

2 Related Work: Measuring the Environmental Impact of AI training

After the high level of carbon emissions associated with training NLP models was reported ([Strubell et al., 2019](#)), researchers stated the need for "Green AI" ([Schwartz et al., 2020](#)), soon structured as an entire research field focused on lowering impacts ([Treviso et al., 2023](#); [Verdecchia et al., 2023](#)).

Tools to Measure the Impact of an Algorithm

Methods and tools to assess the energy use of AI activities have been developed, based on energy use measures (e.g., *CarbonTracker* ([Anthony et al., 2020](#))) or estimations (e.g., *Green Algorithms* ([Lanvelongue et al., 2021](#))). Energy use assessments are complemented with data center energy efficiency indicators such as *Power Usage Effectiveness* (PUE) ([Avelar et al., 2012](#)), and information on the carbon intensity of local electricity mixes to convert energy use into carbon footprint assessments. Reviews have compared tools to assess the energy use and associated carbon footprint of training language models and other AI activities ([Bouza et al., 2023](#); [Jay et al., 2023](#); [Bannour et al., 2021](#)). Additional efforts address the water usage associated with hardware energy use by using the *Water*

Usage Efficiency (WUE) of data centers ([Li et al., 2025](#)).

ML related impacts are also incurred by model development (including dataset curation or architecture search) and use, with environmental impacts coming from each life cycle phase of hardware used ([Ligozat et al., 2022](#)). MLCA ([Morand et al., 2024b](#)) leveraged an existing method ([Luccioni et al., 2023](#)) to assess environmental impacts associated with ML training or inferences, including carbon footprint, energy demand and metallic resource depletion over the production and use phase of the hardware. [Morrison et al. \(2025\)](#) build on [Luccioni et al. \(2023\)](#) and [Li et al. \(2025\)](#) to assess carbon footprint and water use of model training, including development and simulated inference. However, environmental assessments that scale beyond individual models are still needed.

Drivers of AI Impact Growth Remain Unclear

Technical reports from industry and academia suggest that the environmental impacts associated with AI have been rising at the global level ([Google, 2024](#); [Meta, 2024](#); [Shehabi et al., 2024](#)). [de Vries \(2023\)](#) discussed the potential growth of ML with the recent surge in demand for freely accessible LLMs. The [IEA \(2024\)](#) stated that energy use by AI and cryptocurrencies could double by 2026. [Masanet et al. \(2020\)](#) have shown that despite an exponential growth in computation demand between 2012 and 2018, the total energy use of data centers had only increased by around 5% thanks to a shift toward more energy efficient large scale data centers. While this relative decoupling is no longer valid ([IEA, 2025](#)), we lack consolidated information on the global trend for environmental impacts of AI. Information is needed on the impact trend of each step, including training.

Diverging Trends Need to be Analyzed for LLM Training

Optimization techniques have been used to produce models reaching a given level of performance at a reduced training carbon footprint ([Wu et al., 2022](#); [Patterson et al., 2022](#)). The energy efficiency of graphics cards has increased exponentially, leading to the claim that frequent hardware updates will shrink the carbon footprint of training models ([Patterson et al., 2022](#)). In spite of algorithmic optimization, [Sevilla et al. \(2022\)](#) and [Thompson et al. \(2023\)](#) report an exponential increase in compute requirements to train models. Compute requirements are increasing at the same time as energy efficiency, leading to unclear trends

on model training impacts. Optimizations are often absorbed by the growth of the ICT sector (Bol et al., 2021; Gossart, 2015). Is this also true for LLMs training? Luccioni et al. (2025) warn against risks of rebound effect in AI. A *producer rebound* effect (Coroama and Mattern, 2019) could be occurring, where efficiency increases fuel the creation of even larger models thereby lowering or even canceling the potential impact reduction.

Overall, previous work has led to the development of methods and tools for assessing the impact of specific ML algorithms. Reports suggests that the overall environmental impact of AI is rising. However, mechanisms behind such an increase remain undocumented. Deployment of optimizations concurring with growing compute demand could correspond with rebound effect in the case of model training. Herein, we present a comprehensive study of the impact of ML hardware and ML model training over a decade, including hardware production and investigating rebound effect.

3 Method: Environmental Impact Assessment

In this article, we question whether optimization strategies lead to lower training impact. This Section presents the environmental assessment method used in Section 4 to plot impact trends over time. The code and information used in the experiments are available on zenodo : <https://doi.org/10.5281/zenodo.15310868>.

3.1 Data sources

Studies on ML models training are conducted using the Epoch AI Notable systems database (Epoch AI, 2024). Epoch AI is the most comprehensive database on ML systems to our knowledge. It gathers extensive information on a large variety of *notable*¹ ML systems. The database mostly comprises data gathered from papers/technical reports presenting the models and used to yield estimates with varying levels of confidence, but some parameters cannot be reliably estimated. Information on graphics cards characteristics is based on a dataset gathering manufacturer information (Morand, 2026). Both data sources are available under Creative Commons Attribution license. Additional details to support replication are available in Appendix A.

¹Notable systems are “models that have advanced the state of the art, had a large influence in the field’s history, or had a large impact within the world”, per <https://epochai.org/data/notable-ai-models-documentation>.

3.2 General Methodology for Environmental Impact Assessment

To address our research question, we used a methodology to assess the impacts of ML system training including hardware life cycle. This process requires detailed information on hardware characteristics and specific training parameters, such as training duration.

Assessing Multiple Impacts Over Hardware Life-Cycle The MLCA tool (Morand et al., 2024b) is chosen to perform environmental assessments. It uses attributional LCA, a methodology allowing to account for multiple impacts over the life cycle of the used hardware. It provides carbon footprint and metallic resource depletion assessments, based on open-source *Life Cycle Assessment (LCA)* information from Simon et al. (2024)².

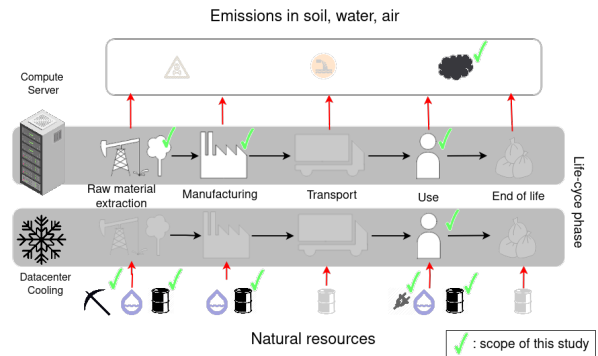


Figure 1: Scope of environmental assessments of ML training in this study. Pictograms represent fossil fuels (barrel), water (droplet), metals (pickaxe) and electricity (plug). Top row icons represent, from left to right, emissions in soil, water and air.

Figure 1 presents the scope of our study within the life cycle of hardware used for training ML models. We consider energy use, carbon footprint and metallic resource depletion over server production and usage, and data center cooling use.

Carbon footprint is assessed according to *Global Warming Potential (GWP)*,³ measured in kgCO₂ eq, for the emissions of greenhouse gases such as carbon dioxide, methane and nitrous oxide (Forster et al., 2023). Metallic resource depletion is assessed through *Abiotic Depletion Potential for el-*

²Water use and toxicity to human and non-human life are also important sources of impact for ICT hardware and services. Recent open quality information on toxicity and water use of ICT equipment and data center facilities is missing (Li et al., 2025; Lei and Masanet, 2022). End-of-life information is also lacking (Ficher et al., 2025).

³We use the IPCC standard GWP100, corresponding to impact at a time horizon of 100 years.

elements (ADPe), expressed in *kilograms antimony equivalent* (kgSb eq) (van Oers et al., 2020), which is based on the quantity of metal (e.g., copper, gold and rare earths) used to produce hardware. The amount of each metal and metal rarity are compared to Antimony, chosen as standard reference, to create and aggregated indicator.

MLCA is based on a bottom-up modeling of hardware, meaning that hardware is modeled as the sum of its parts as represented in Figure 2. Graphics cards are modeled based on GPU size (in hatched purple in the Figure), on-board memory size (in cross-hatched blue) and constant impact for other card components (in green). Their energy use during ML training is estimated using their **Thermal Design Power (TDP)**⁴ and training duration.

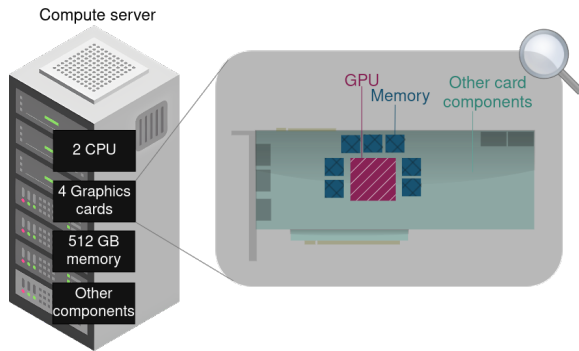


Figure 2: Hardware modeling in MLCA.

Training Duration We estimate GPU-hours required for training models using two methods, based on information from the Epoch AI database presented in Table 1. The estimates are performed on model entries majorly assessed as reporting high-quality information (74% of models).

For models specifying duration and graphics cards quantity, they can be multiplied to obtain an estimate referred to as GPU- h_1 :

$$\text{GPU-}h_1 = \text{training duration} \times \#\text{cards}$$

This estimation is the most reliable as it uses information retrieved from the papers presenting the models. However, it only covers 15% of Epoch AI models (see Table 1).

Another method to estimate the training duration amounts to dividing the training compute requirements (number of FLOP) necessary for model

⁴The TDP corresponds to the heat the chip should be able to dissipate to function at maximum load; it is used to approximate the maximum power use of the chip.

training by the peak compute power of the card used (in FLOPs, i.e., the number of FLOP the card can perform per second). This value is referred to as GPU- h_2 base:

$$\text{GPU-}h_2 \text{ base} = \frac{\text{training FLOP}}{\text{peak card compute power}}$$

GPU- h_2 base should lead to an under-estimation of the number of GPU-hours compared to GPU- h_1 , as the hardware does not always operate at peak performance, for instance due to synchronisation phases between the different nodes when compute cannot be realised.

After checking for consistency between the two estimation methods and excluding benign anomalies, we develop a linear model to correct the under-estimation from GPU- h_2 base compared to GPU- h_1 . Appendix A.2 details procedures for ensuring consistency and details the linear model. Appendix C explores the sensitivity of the estimated training duration to variations in the used performance ratio, and the influence these variations have on the assessed training carbon footprint and metallic resource depletion. The obtained model corresponds to using a quasi-constant performance ratio of $\simeq 27\%$. This leads to our second estimation method referred to as GPU- h_2 :

$$\text{GPU-}h_2 = \frac{\text{training FLOP}}{\text{peak card compute power} \times 0.27}$$

The final estimation for GPU hours is as follows: we use GPU- h_1 for the 131 models (15%) where it is available, and use GPU- h_2 for another 103 models (11%), covering 26% of Epoch AI models. Among these 234 models, 132 (56.4%) are NLP models (indicating at least language or speech modality). 123 (95.3%) of these models are language models, and the 6 other (4.7%) are speech models. 26 (20.1%) NLP models are multimodal. We include models in domains other than NLP to refine potential trends even if our primary focus is language models. In total, 30 (12.8%) of the 234 models indicate a base model, and are thus considered to be fine-tuned or distilled. Training duration is estimated using GPU- h_1 for these models and only accounts for the fine-tuning phase and not the base model training.

Illustration on GPT-4 We illustrate our methodology for the GPT-4 model, released in March 2023. Training duration is estimated at 57, 000, 000 GPU-hours on NVIDIA A100 SXM4 40 GB cards based on information in the EpochAI dataset and using

		GPU-h ₁				GPU-h ₂			
		models	training duration	card quantity	both	+ card model	training FLOP	card model	both
Number		897	199	168	131	122	451	266	230
Coverage (%)		100	22	19	15	14	50	30	26
Confidence	Confident	234	98	96	78	75	166	138	122
	Likely	103	41	34	21	17	77	51	47
	Speculative	69	21	11	9	7	51	20	16
	Unknown	491	39	27	23	23	157	57	45

Table 1: Number of entries for each information type in the Epoch AI database. Confidence scores are reported by database authors based on the quality of the Training compute, Parameters, and Training data set size fields.

GPU-h₁. Training servers are modeled as containing 4 cards, 2 CPUs and 512GB memory (consistently with NVIDIA workstation cards), and facility PUE is estimated at 1.2 (because the model was trained after 2018). Energy use impact is computed with average USA energy mix (GPT4 authors are affiliated in the US). Table 2 presents impacts associated with GPT-4 training estimated using MLCA. For this model, embodied impacts represent 20% of the total carbon footprint and close to 100% of metallic resource depletion. Carbon footprint estimation is consistent with previous assessments⁵.

	Energy (GWh)	GWP (ktCO ₂ eq)	ADPe (kgSb eq)
Embodied	-	3.3	300
Usage	27.4	10.2	2.7
Infra	5.4	2.0	0.5
Total	32.8	15	300

Table 2: Estimated production impacts for training GPT-4: impacts related to the energy use of servers (row *Usage*), infrastructure (row *Infra*) and embodied impacts (row *Embodied*) are presented.

Simulating Carbon Optimization Strategies

The study considers different scenarios for reducing **Carbon Intensity (CI)** in energy mixes over time, including shifting compute locations and decarbonizing electricity for data centers. Each scenario involves a continuous reduction in **CI**, starting from 2019, with reductions of up to 25% per year onward. The **CI** of the mix used for training a model is thus multiplied by $(1 - \text{ratio})^n$ where n stand for the number of years since 2019 at release date of the system.

⁵<https://medium.com/data-science/the-carbon-footprint-of-gpt-4-d6c676eb21ae> [Last Accessed 16/04/2026]

4 Results: Trends in ML Model Training

Now that we have data and a method for environmental assessments, we aim at answering the following questions: How have graphics cards production impacts evolved? How have model training impacts evolved?

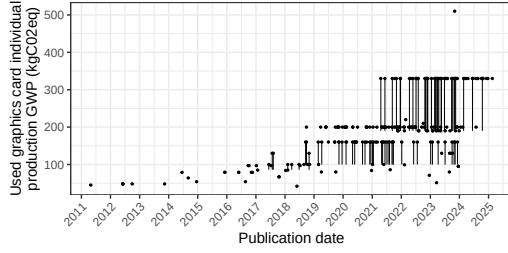
Production impacts of the used cards More recent models are trained on newer cards. The new cards are more technologically advanced, allowing for increased compute capacities and energy efficiency. As shown in Figure 3, this however leads to increases in the production impacts of graphics cards used to train **ML** models.

Figure 4 shows that hardware quantity has furthermore increased exponentially, even if some models are still trained using only a few cards. More graphic cards with higher production impacts are used, suggesting growing environmental impacts. Both production impact and energy use during usage need to be addressed.

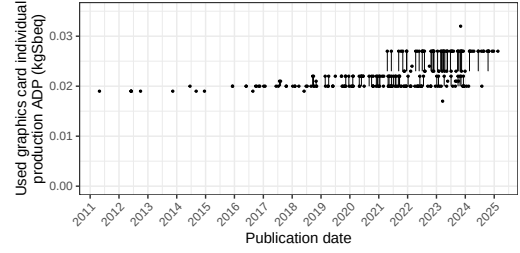
Energy use of Model Training Figure 5 presents the energy use associated with model training estimated using MLCA (see Section 3.2). Energy use has increased exponentially over time. Appendix B details regression analysis for the assessed trends in energy use and environmental impacts. While Low energy use models are still produced in recent years, the number of models released over time has increased overall.

Environmental Impacts of Model Training.

Figure 6 presents the estimated training impacts in terms of **GWP** (Figure 6a) and **ADPe** (Figure 6b). Both environmental indicators have increased exponentially between 2012 and 2025. **NLP** appear representative of other domains. Additional disaggregated experiments indicate that language, vision,



(a) GWP



(b) ADPe

Figure 3: Evolution of the production impacts of the graphics cards (assessed using the MLCA tool) used for training **ML** systems in the Epoch AI database. Value intervals correspond to cases of ambiguous card names.

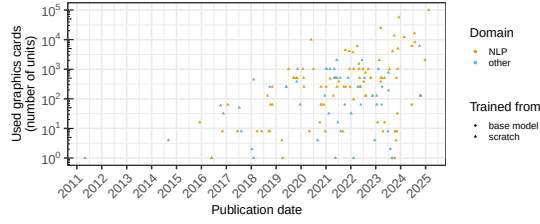


Figure 4: Number of graphics cards used for training models in Epoch AI over time. Y scale is logarithmic so a linear trend corresponds to an exponential growth.

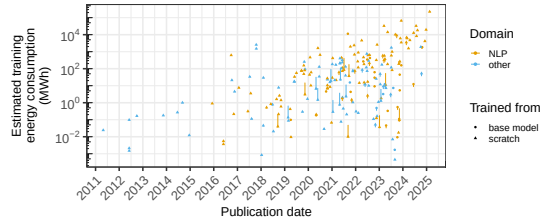


Figure 5: Energy use of training over time. Value intervals account for ambiguous card names. Y scale is logarithmic so a linear trend means exponential growth.

and multimodal models (to a lesser extent) all exhibit an exponential growth trend.

Table 3 summarizes the distribution of the shares of embodied impacts. **ADPe** comes close to exclusively from hardware production. The share of embodied **GWP** are mostly estimated around a third the total, which is consistent with previous assessments in the literature (Gupta et al., 2022; Wu et al., 2024; Luccioni et al., 2023). Variation mostly comes from differences in the **CI** of the electricity mixes of countries in which models are trained. Models trained in countries with a low **CI** for their electricity use such as France will present larger shares of embodied impacts than if trained in another country such as the USA. Embodied impacts represent a significant share of total impacts for both indicators. Thus, solely reducing the

energy impacts will not be sufficient to solve the environmental impacts of **ML** models training.

	Min	Q1	Q2	Mean	Q3	Max
ADPe	89	100	100	99	100	100
GWP	12	19	23	23	25	58

Table 3: Share (in percent) of embodied impacts on the total impacts associated with training models.

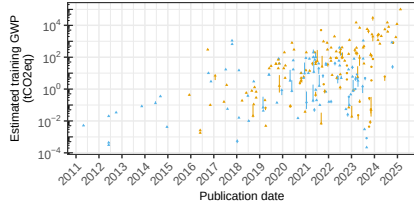
5 Result Analysis: Why do Training Impacts Rise in spite of Optimization?

Rising impact trends may appear surprising as optimisation strategies are being deployed. In this section, we investigate what role hardware and software optimization played in these trends and whether carbon optimization can mitigate impact.

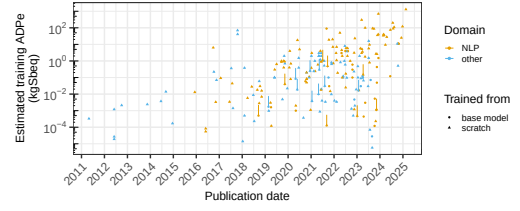
5.1 Hardware optimization

Compute efficiency is computed by dividing the peak performance of cards (maximum compute power in Single, Double, Half and Tensor precision) by **TDP**. It corresponds to the number of operations cards can perform per second per Watt. It is expressed in FLOPs/W, and larger values are better.

Figure 7 presents the evolution of the compute efficiency of the cards used to train **ML** models over time. Compute efficiency has increased exponentially, meaning that hardware optimization has been leveraged extensively, and that jobs with a given compute demand require an exponentially decreasing energy use. However, these efficiency gains appear to have mostly been used to foster larger training jobs and not to mitigate or reduce energy use. Indeed, as shown in Figure 8, the use of more compute efficient hardware is correlated with an (exponential) increase in hardware energy use during training.



(a) GWP



(b) ADPe

Figure 6: Environmental impacts in GWP (a) and ADPe (b) of training ML models over time. Value intervals account for ambiguous card names and model producers from multiple countries. Y scales are logarithmic so a linear trend corresponds to an exponential growth.

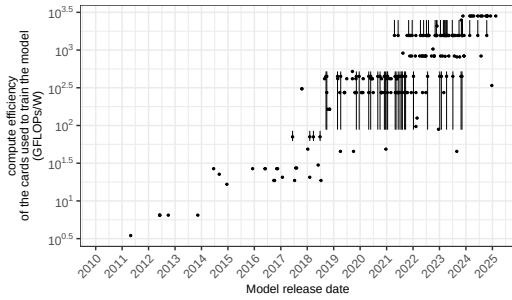


Figure 7: Evolution of the compute efficiency of the graphics cards used to train models in the Epoch AI database (2013-2025). Higher is better. Value intervals account for ambiguous card name

increasing model performance.

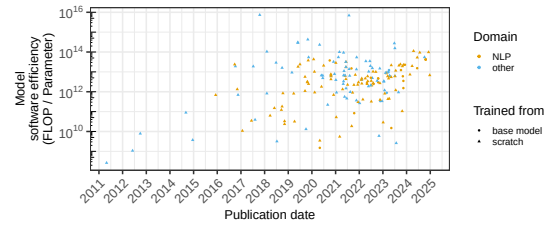


Figure 9: Software efficiency of models in Epoch AI over time. Lower is better. Y scale is logarithmic so a linear trend corresponds to an exponential growth.

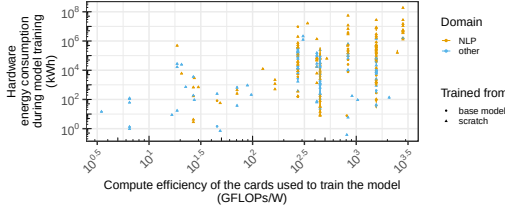


Figure 8: Hardware energy use during training over compute efficiency of the used card. Value intervals account for ambiguous card name

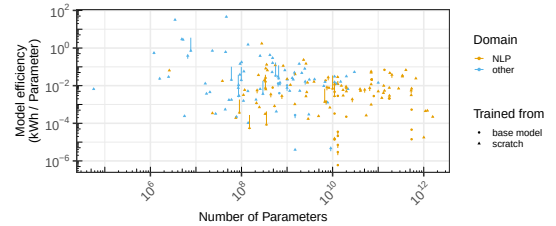


Figure 10: Combined efficiency of notable models in Epoch AI over model parameters. Lower is better. Value intervals account for ambiguous card name

5.2 Software Optimization

More efficient model architectures are continuously developed and allow to attain a set model performance with less resources (Patterson et al., 2022). Be they sparser or more efficient in other manner, these new architectures and models should require less compute per parameter to attain a given performance. We approximate model *software efficiency* as the average compute per parameter in the model training. Software efficiency is thus expressed in FLOP / parameter. We expect that more recent models have benefited from algorithmic optimization, and would therefore be more software efficient for a given performance target. It is however unclear how software optimisations fare in a context of

Figure 9 details the evolution in software efficiency of notable models in the Epoch AI database over time. Software efficiency gains appear outweighed by increased compute demand to increase model performance, with a stable or even increasing software efficiency metric when we could have expected a decrease corresponding with the roll-out of the more efficient newer architectures. As for hardware optimization, our results suggest that software optimization have historically served model growth more than impact mitigation.

Hardware energy use during model training depends both on hardware and software characteristics. Figure 10 details the combined efficiency of models training, expressed in kWh / parameter, over model size. The combined efficiency appears

to remain mostly constant with only a slight improvement with model size.

Hardware and software optimizations yield economies of scale that are outpaced by the increase in compute demand, resulting in a net increase in energy use for ML training.

5.3 Carbon optimization can only yield short-lived benefits

We compare the impact of models trained with actual electricity mixes with a simulated annual reduction of 25% of the CI of electricity mixes (see Section 3.2). Figure 11 shows that the carbon footprint of models released from 2019 is increasing regardless of carbon optimizations. Trends are very similar with or without CI reduction and regression coefficients are significantly positive. This suggests that carbon optimization is not a sufficient reduction strategy.

CI for electricity is limited by the lowest global mixes (15-20gCO₂ eq/kWh), meaning that even fast CI reductions cannot keep pace with rising energy use. While using low-impact electricity is a key Green AI practice, it cannot alone curb the growth of training impacts. Furthermore, this strategy fails to address the increasing environmental costs of hardware production. Worse, shifting data centers geographically to reduce CI may incentivize shorter lifespan for these facilities, further increasing their total environmental impact.

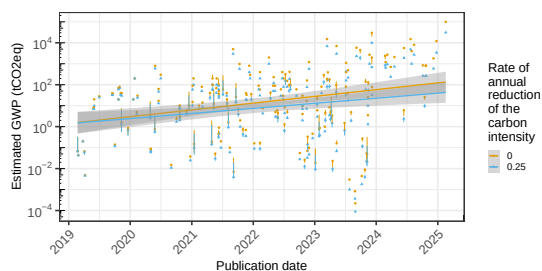


Figure 11: Estimated carbon footprint of training models released after 2019, with or without reduction of the CI of the used electricity.

6 Discussion

To our knowledge, this is the first study proving the growth in ML training impacts and explaining drivers of this growth. We show that carbon optimization strategies fail to curb training impacts.

6.1 Impact Shifting

Computing facilities are increasingly energy efficient and can perform much more computation in a given duration using less energy. This efficiency is partly obtained by regular hardware updates in data centers. Replacement can occur at the end-of-life of hardware or be more premature, shortening the lifespan of the devices. Frequently changing hardware could be understood as a form of *impact shifting*. Energy use can be reduced (at constant use) during the potentially shorter equipment lifespan, so usage impact is reduced. However, production impact in terms of carbon footprint and metallic resource depletion is increased by frequent hardware renewal. In addition, new hardware also has higher production impact due to technological advances required to reduce energy use in the usage phase.

6.2 Rebound Effect is Prevalent in AI

Our results suggest rebound effect, consistently with previous observations: Wu et al. (2022) identify a rebound effect at Facebook. Patterson et al. (2022) show that a constant share of Google’s energy use is attributable to AI when the total use increased. Efforts from de Vries (2023) and Desislavov et al. (2023) suggest that our findings should also apply to inferences.

Hardware and algorithmic optimization appear to have fueled increases in compute demand to chase better model performances. Their role thus appears not so much as a mitigation but as an enabler of growing impacts.

6.3 Greener Energy Cannot Void Carbon Impact

Figure 11 shows that reducing CI of electricity used seems insufficient to curb the exponential growth of the carbon footprint of training ML models. Furthermore, the electricity use of data centers destabilizes local electricity grids (Ortar et al., 2023), potentially causing the prolongation of fossil fuel power plants (Chu, 2024; Ambrose, 2024). On-site renewable energy production to match data centers use can also be a source of electricity grid instability (Gnibga et al., 2024). Matching the carbon footprint through carbon offsetting also has limited potential (Greenfield, 2023; Lohmann, 2009).

6.4 Community effort to document environmental impact is required

The NLP community has identified the need to rethink progress and evaluation practices. Envi-

Information	Details
Training duration	time
Card model	name
Number of cards	number
Server model and configuration	name and (number CPU, memory volume. . .)
Type of datacenter	e.g., laboratory facilities, hyperscaler,...
Hardware usage during training	ratio
Datacenter efficiency	typically PUE
Datacenter location	country or state

Table 4: Proposed standard reporting format to ensure sufficient information for environmental assessment of final model training

ronmental issues surrounding LLMs (at all the stages of their life cycles) are significant, making it pressing to include environmental impacts in NLP evaluation. Sole reporting of carbon footprint figures are insufficient as environmental assessments are complex and currently often difficult to reproduce (Morand et al., 2024a). More transparency is needed, ensuring that papers report all the necessary information to produce environmental assessments, notably including hardware used, training duration, compute localisation. A standardised reporting format could be proposed for instance in the form of a Table, or cards such as proposed by Hershovich et al. (2022), Chiu (2026) or Hender-son et al. (2020). Table 4 details the necessary information for model training in the form of a proposed standard table.

7 Conclusion

We show that the environmental impacts of training language models, and ML models in general, grew exponentially from 2013 to 2025. Despite hardware, algorithmic and carbon optimizations, impacts keep growing, consistent with rebound effect. This growth of impacts is only part of the environmental dynamics of AI. It does not account for impact sources other than individual model training, such as inference. Obtaining information to plot the trends of training impact is a grinding task that could be alleviated by a community wide effort to document environmental impacts as part of NLP evaluation. Integrating environmental impacts in NLP evaluation would give better visibility of impacts to the community and support the use of impact as a feature in research planning and decision making. Impact reduction must be combined with a broader consideration of the role of AI in a

sustainable society.

8 Limitations

8.1 Study Scope Supports Reliable Trends

Production impact assessment in MLCA has limitations, including not accounting for technological node and assuming constant memory density, but estimates are consistent with recent research (Schneider et al., 2025). The use of the PUE may lead to underestimation of infrastructure use. Adjusting for dynamic ratio as proposed by Morand et al. (2024b) would not change the overall findings since it would only mean using a higher multiplicative factor for all models. Epoch AI, although extensive, does not include all models and variants, but omissions are unlikely to alter the observed exponential increase in training impacts.

8.2 Sparse Information Still Leads to Consistent Estimations

This study estimates missing parameters for models using inferences and hypotheses, which may lead to under or over-estimation of certain parameters (e.g., training time based on number of FLOP). The validity of these estimates is checked by comparing our estimations to published information. These check yield to some significant differences but general consistency. For example, the carbon footprint for the BLOOM-176B model presents one of the greatest difference: our model estimates a carbon footprint of 220tCO₂ eq instead of 50-80tCO₂ eq, due to assumptions about its training location and electricity mix. Yet, most other models have close or very close estimations, as for the case of the GPT-4 model used as an example for our methodology where our model estimation of 15ktCO₂ eq is very close to the reference estimation of 12 to 15

15ktCO₂ eq⁶.

8.3 This Study Addresses Model Training

Morrison et al. (2025) and Luccioni et al. (2023) have shown that model development includes training multiple smaller models to converge on final architecture. These multiple training runs, that we do not account for, can account for at least half of the total footprint (Morrison et al., 2025). Future work can also address the impact of inferences in the sector, as they weight more heavily over the life cycle of models. It will however require access to user information.

8.4 Impacts Go Beyond Carbon Footprint and Resource Depletion

This study examined the carbon footprint and metallic resource depletion of training AI models, finding that both metrics have increased over time due to impacts in different phases in the life cycle of hardware. Importantly, AI environmental impacts extend beyond these metrics, including water usage (Mytton, 2021), ecosystem destruction (Comber and Eriksson, 2023), and pollution from hardware mining and disposal. Additionally, AI poses significant social consequences and ethical challenges (Bender et al., 2021; Jiang et al., 2023), highlighting the need for a more comprehensive assessment that incorporates qualitative analyses and a broader range of impacts.

Ethical Considerations

This study provides data for the footprint of training ML models, including LLMs, over a decade using estimations and information from the EpochAI database and GPU manufacturers. No re-training of the models was needed to compute impact metrics.

Acknowledgments

This work has received funding from the French "Agence Nationale pour la Recherche" under grant agreement InExtenso - ANR-23-IAS1-0004. Clément Morand was supported by a doctoral grant from ENS Rennes. The authors thank TechPowerUp for allowing us to use their data in this study. Clément Morand would like to thank Aina Rasoldier for his help in gathering the data from TechPowerUp, Loïc Lannelongue for discussions

during the experiments framing and Adrien Berthelot for insightful discussion and bibliographical advice. The authors are grateful to Aurélie Bugeau, Karën Fort, Caio Corro and Pierre Zweigenbaum, as well as the anonymous reviewers, for their comments on earlier versions of this manuscript.

References

- Jillian Ambrose. 2024. [Ireland's datacentres overtake electricity use of all urban homes combined](https://www.theguardian.com/world/article/2024/jul/23/ireland-datacentres-overtake-electricity-use-of-all-homes-combined-figures-show). The Guardian, <https://www.theguardian.com/world/article/2024/jul/23/ireland-datacentres-overtake-electricity-use-of-all-homes-combined-figures-show>.
- Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. [Carbontracker: Tracking and predicting the carbon footprint of training deep learning models](#). *Preprint*, arXiv:2007.03051.
- Victor Avelar, Dan Azevedo, Alan French, and Emerson Network Power. 2012. PUE: a comprehensive examination of the metric. *White paper*, 49.
- Nesrine Bannour, Sahar Ghannay, Aurélie Névél, and Anne-Laure Ligozat. 2021. [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*, pages 11–21, Virtual. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Adrien Berthelot, Eddy Caron, Mathilde Jay, and Laurent Lefèvre. 2024. [Estimating the environmental impact of Generative-AI services using an LCA-based methodology](#). In *Procedia CIRP*, pages 1–10, Turin, Italy.
- David Bol, Thibault Pirson, and Rémi Dekimpe. 2021. [Moore's law and ICT innovation in the anthropocene](#). In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 19–24.
- Lucía Bouza, Aurélie Bugeau, and Loïc Lannelongue. 2023. [How to estimate carbon footprint when training deep learning models? a guide and review](#). *Environmental Research Communications*, 5(11):115014.
- Guan-Ming Chiu. 2026. [Position: Carbon footprint reporting should be routine in machine learning research](#). In *Forty-third International Conference on Machine Learning Position Paper Track*.

⁶The reference estimation only accounts for energy use and not hardware production.

- Amanda Chu. 2024. [US slows plans to retire coal-fired plants as power demand from AI surges](#). Financial Times.
- Rob Comber and Elina Eriksson. 2023. Computing as Ecocide. In *Ninth Computing within Limits 2023*. LIMITS. <https://limits.pubpub.org/pub/a8h46wqy>.
- Vlad C. Coroama and Friedemann Mattern. 2019. [Digital rebound - why digitalization will not redeem us our environmental sins](#). In *Proceedings of the 6th International Conference on ICT for Sustainability, ICT4S 2019, Lappeenranta, Finland, June 10-14, 2019*, volume 2382 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Alex de Vries. 2023. [The growing energy footprint of artificial intelligence](#). *Joule*, 7(10):2191–2194.
- Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. 2023. [Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning](#). *Sustainable Computing: Informatics and Systems*, 38:100857.
- Epoch AI. 2024. [Data on notable AI models](#). Accessed: 2025-02-28.
- Kawin Ethayarajh and Dan Jurafsky. 2020. [Utility is in the eye of the user: A critique of NLP leaderboards](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4846–4853. Association for Computational Linguistics.
- Marion Ficher, Tom Bauer, and Anne-Laure Ligozat. 2025. [A comprehensive review of the end-of-life modeling in LCAs of digital equipment](#). *The International Journal of Life Cycle Assessment*, 30:20 – 42.
- P. Forster, T. Storelvmo, K. Armour, W. Collins, J.-L. Dufresne, D. Frame, D.J. Lunt, T. Mauritsen, M.D. Palmer, M. Watanabe, M. Wild, and H. Zhang. 2023. [The earth’s energy budget, climate feedbacks and climate sensitivity](#). In *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, page 923–1054. Cambridge University Press.
- Wedan Emmanuel Gnibga, Anne Blavette, and Anne-Cécile Orgerie. 2024. [Renewable Energy in Data Centers: the Dilemma of Electrical Grid Dependency and Autonomy Costs](#). *IEEE Transactions on Sustainable Computing*, 9(3):315–328.
- Google. 2024. [Environmental Report 2024](#). Accessed March 18 2025.
- Cédric Gossart. 2015. [Rebound effects and ICT: A review of the literature](#). In *ICT Innovations for Sustainability*, pages 435–448, Cham. Springer International Publishing.
- Patrick Greenfield. 2023. [Revealed: more than 90% of rainforest carbon offsets by biggest certifier are worthless, analysis shows](#). The Guardian, <https://www.theguardian.com/environment/2023/jan/18/revealed-forest-carbon-offsets-biggest-provider-worthless-verra-aoe>.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafford, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. [OLMo: Accelerating the science of language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.
- Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. 2022. [Chasing carbon: The elusive environmental footprint of computing](#). *IEEE Micro*, 42(4):37–47.
- 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](#). *J. Mach. Learn. Res.*, 21(1).
- Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. [Towards climate awareness in NLP research](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- IEA. 2024. [Electricity 2024](#). Licence: CC BY 4.0.
- IEA. 2025. [Energy and AI](#). Licence: CC BY 4.0.
- Mathilde Jay, Vladimir Ostapenko, Laurent Lefevre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel. 2023. [An experimental comparison of software-based power meters: focus on cpu and gpu](#). In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 106–118.
- Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. [AI art and its impact on artists](#). In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’23*, page 363–374, New York, NY, USA. Association for Computing Machinery.
- Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. [Green algorithms: Quantifying the carbon footprint of computation](#). *Advanced Science*, 8(12):2100707.
- Nuoa Lei and Eric Masanet. 2022. [Climate and technology-specific PUE and WUE estimations for](#)

- u.s. data centers using a hybrid statistical and thermodynamics-based approach. *Resources, Conservation and Recycling*, 182:106323.
- 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.
- Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. 2022. [Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions](#). *Sustainability*, 14(9).
- Larry Lohmann. 2009. [Toward a different debate in environmental accounting: The cases of carbon and cost-benefit](#). *Accounting, Organizations and Society*, 34(3):499–534.
- Alexandra Sasha Luccioni, Emma Strubell, and Kate Crawford. 2025. [From efficiency gains to rebound effects: The problem of jevons' paradox in ai's polarized environmental debate](#). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 76–88, New York, NY, USA. Association for Computing Machinery.
- 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.
- Eric Masanet, Arman Shehabi, Nuo Lei, Sarah Smith, and Jonathan Koomey. 2020. [Recalibrating global data center energy-use estimates](#). *Science*, 367(6481):984–986.
- Meta. 2024. [2024 sustainability report](#). Accessed March 18 2025.
- Clément Morand, Anne-Laure Ligozat, and Aurélie Névoul. 2024a. [Empreinte carbone des expériences en TAL : les défis de la reproductibilité](#). In *Actes de la journée d'étude Journée Éthique et TAL 2024*, Actes de la journée d'étude Journée Éthique et TAL 2024, Nancy, France. Karën Fort and Aurélie Névoul.
- Clément Morand, Anne-Laure Ligozat, and Aurélie Névoul. 2024b. [MLCA: a tool for Machine Learning Life Cycle Assessment](#). In *2024 10th International Conference on ICT for Sustainability (ICT4S)*, pages 227–238, Stockholm, Sweden. IEEE.
- Clément Morand. 2026. [Dataset on nvidia workstation graphics cards from 2013 to 2024 and their estimated environmental impact](#).
- Jacob Morrison, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell, and Jesse Dodge. 2025. [Holistically evaluating the environmental impact of creating language models](#). In *The Thirteenth International Conference on Learning Representations*.
- David Mytton. 2021. [Data center water consumption](#). *npj Clean Water*, 4.
- Nathalie Ortat, A. R. E. Taylor, Julia Velkova, Patrick Brodie, Alix Johnson, Clément Marquet, Andrea Pollio, and Liza Cirolia. 2023. [4 Powering 'smart' futures: data centres and the energy politics of digitalisation](#), pages 125–168. De Gruyter, Berlin, Boston.
- George Ostrouchov, Don Maxwell, Rizwan A. Ashraf, Christian Engelmann, Mallikarjun Shankar, and James H. Rogers. 2020. [Gpu lifetimes on titan supercomputer: Survival analysis and reliability](#). In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14.
- David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2022. [The carbon footprint of machine learning training will plateau, then shrink](#). *Computer*, 55(7):18–28.
- Ehud Reiter. 2025. [We should evaluate real-world impact](#). *Computational Linguistics*, pages 1–13.
- Ian Schneider, Hui Xu, Stephan Benecke, David Patterson, Keguo Huang, Parthasarathy Ranganathan, and Cooper Elsworth. 2025. [Life-cycle emissions of ai hardware: A cradle-to-grave approach and generational trends](#). *Preprint*, arXiv:2502.01671.
- Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. [Green AI](#). *Commun. ACM*, 63(12):54–63.
- Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. [Fine-tuned language models are continual learners](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. [Compute trends across three eras of machine learning](#). In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- A. Shehabi, S.J. Smith, A. Hubbard, A. Newkirk, Siddik Lei, N., B. M.A.B., Holecek, J. Koomey, E. Masanet, and D. Sartor. 2024. [2024 united states data center energy usage report](#). Technical report, Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001637.
- Thibault Simon, David Ekchajzer, Adrien Berthelot, Eric Fourboul, Samuel Rince, and Romain Rouvoy. 2024. [BoaviztAPI: a bottom-up model to assess the environmental impacts of cloud services](#). In *Hot-Carbon'24 - 3rd Workshop on Sustainable Computer Systems*, Santa Cruz, United States.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. [Energy and policy considerations for deep learning in NLP](#). In *Proceedings of the 57th*

Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 3645–3650. Association for Computational Linguistics.

Neil Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2023. The Computational Limits of Deep Learning. In *Ninth Computing within Limits 2023*. LIMITS. <https://limits.pubpub.org/pub/wm1lwjce>.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, and 3 others. 2023. *Efficient methods for natural language processing: A survey*. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Lauran van Oers, Jeroen B Guinée, and Reinout Heijungs. 2020. *Abiotic resource depletion potentials (ADPs) for elements revisited—updating ultimate reserve estimates and introducing time series for production data*. *The International Journal of Life Cycle Assessment*, 25:294–308.

Gael Varoquaux, Sasha Luccioni, and Meredith Whittaker. 2025. *Hype, sustainability, and the price of the bigger-is-better paradigm in ai*. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, page 61–75, New York, NY, USA. Association for Computing Machinery.

Roberto Verdecchia, June Sallou, and Luís Cruz. 2023. *A systematic review of green AI*. *WIREs Data Mining and Knowledge Discovery*, 13(4):e1507.

Carole-Jean Wu, Bilge Acun, Ramya Raghavendra, and Kim Hazelwood. 2024. *Beyond efficiency: Scaling AI sustainably*. *IEEE Micro*, 44(5):37–46.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, and 6 others. 2022. *Sustainable AI: environmental implications, challenges and opportunities*. In *Proceedings of Machine Learning and Systems 2022, MLSys 2022, Santa Clara, CA, USA, August 29 - September 1, 2022*. mlsys.org.

A Details for Exact Replication of Environmental Assessments

A.1 Processing Ambiguous Card Names

8 ML models are excluded from analysis because various hardware is used throughout the training process and adequate attribution is not possible.

Other models document hardware ambiguously (e.g., "A100" could refer to several cards with different features). Then, the one we expect to be most frequently used, is chosen as the reference value, and other options are used to compute a value interval.

Server Characteristics We made hypotheses on memory provisioning and number of cards per server, and hardware lifespan and utilization using values consistent with computing facilities at our university, documentation from graphic card and server manufacturers and the literature (Ostrouchov et al., 2020; Wu et al., 2022). Exact details for replication on server characteristics is available in A.3. This information serves to allocate server production impact to a specific model training. Allocated production impacts are called *embodied impacts*.

We also modeled the increase in data center efficiency estimating average PUE from 2010 to 2018 using linear interpolation and a PUE of 1.2 from 2018 onward based on Masanet et al. (2020).

Hardware Energy use. We use a value of 100% usage both for GPUs and CPUs during training as it was shown to yield accurate estimations (Jay et al., 2023).

Environmental Impact of Energy Usage To evaluate the impacts of energy use associated with training each model, we consider that models have been trained using the energy mix of the countries of the ML model producers documented by Boavizta.⁷ If multiple countries participate in model creation, the energy mixes of each implicated country are used to create a value interval, and the mix of the country indicated in first position is used as the reference value.

A.2 Training Duration Estimates

Ensuring Consistency and Excluding Anomalies: In order to validate estimates from GPU-h₂ base and correct the underestimation, we compare GPU-h₁ and GPU-h₂ base for each of the 119 models where both values are available, taking GPU-h₁ as a ground truth. This allows us to obtain an estimate for the ratio of hardware performance: $\text{ratio} = \frac{\text{GPU-h}_2 \text{ base}}{\text{GPU-h}_1}$. This ratio should in

⁷Details on the sources used for each countries can be found at: https://github.com/blubrom/MLCA/blob/main/boaviztapi/data/electricity/electricity_impact_factors.csv.

theory be lower than one as a ratio greater than one would in theory mean that the model has used the GPUs at more than their peak performance. We define *anomalies* as models where the obtained ratio is greater than one or smaller than 10%. Analyzing the obtained ratios revealed anomalies for 19 models (in 16% of cases). Anomalies especially occur with cases of fine-tuned models where GPU-h₂ base includes training the base models while GPU-h₁ only accounts for the fine-tuning process. Anomalies also occur when the number of FLOP in the EpochAI database was estimated based on compute power values different from the ones in our graphics cards database due to inconsistencies between the Epoch AI database and manufacturer data. Both types of anomalies should not pose problems as we use GPU-h₁ as our final estimation for these models.

Linear Model for Correcting Underestimations:

After checking for consistency between the two estimation methods and excluding benign anomalies, we develop a linear model to correct the underestimation from GPU-h₂ base compared to GPU-h₁. We build a linear model to predict GPU-h₁ using GPU-h₂ base. This model is computed on 100 observations excluding the anomalies. To ensure a linear relation between both variables, we estimate $\log(\text{GPU-h}_1) \sim \log(\text{GPU-h}_2 \text{ base})$ as GPU-h₁ and GPU-h₂ base both are exponentially distributed. The linear regression analysis revealed a statistically significant model ($F(1,98) = 6525$, $p < 2 \times 10^{-16}$), with an adjusted R² of 0.98, meaning that 98 percent of the variance in the observations is explained by our model. The model equation is $\log(\text{GPU-h}_1) = 1.3 + 1.0 \log(\text{GPU-h}_2 \text{ base})$ with a standard error of 0.12 for the intercept and 0.01 for the regression coefficient. This indicates that an increase of 1 for the $\log(\text{GPU-h}_2 \text{ base})$ value leads to an average increase of 1.00 units in $\log(\text{GPU-h}_1)$. This positive relationship between $\log(\text{GPU-h}_2 \text{ base})$ and $\log(\text{GPU-h}_1)$ was found to be statistically significant ($t(98) = 80.78$, $p < 2 \times 10^{-16}$), affirming the predictive power of $\log(\text{GPU-h}_2 \text{ base})$ on $\log(\text{GPU-h}_1)$. In addition to the regression analysis, a scatterplot with the fitted regression line was examined to ensure model assumptions were met. Homoscedasticity was confirmed (studentized Breusch-Pagan test = .08, $p = 0.77$) and the residuals appeared to be independent (Durbin-Watson D = 1.81, $p = .17$).

A.3 Details on Modeled Server Characteristics

We made the hypothesis that servers equipped with NVIDIA workstations cards contain 4 graphics cards, 2 CPUs and 512 GB memory. We also made the hypothesis that servers equipped with NVIDIA non-workstations cards contain 2 graphics cards, 2 CPUs and 192 GB memory. These configurations were chosen based on a review of configurations in the different generations of NVIDIA DGX servers as well as on configurations in different computing facilities, including computing facilities available in the Université Paris-Saclay.

For non-NVIDIA hardware, we searched documentation (e.g., Google Cloud Platform documentation, publications by Google) to obtain information on the number of chips and processors per server. For instance, for TPUv3, a server with two CPUs manages every four TPU chips. Missing information on the memory quantity in each server, the value of 448 GB memory is chosen based on the value for the TPU v5p for the TPU v2, v3 and v4. This value seems consistent with the value chosen for the NVIDIA servers.

We use values consistent with hyper-scale data centers for the hardware lifespan, average utilization over its life cycle and infrastructure energy use. We use a lifespan of 3 years for the hardware (Ostrouchov et al., 2020), and, using information from Meta, a hardware utilization of 50% (Wu et al., 2022).

Based on the data from Masanet et al. (2020), we model an increase in the efficiency of data centers between 2010 and 2018, accompanied with the shift from traditional data centers to hyperscale data centers. Assuming a uniform distribution between Traditional, Cloud and Hyperscale data centers, we use an average PUE of 1.75 for data centers in 2010. (This average value of 1.75 is close to the average PUE value for Cloud (non-hyperscale) data centers in 2010) Assuming that all AI workloads take place in hyperscale data centers starting in 2018, we use an average PUE of 1.2 from 2018 onwards. For dates between 2010 and 2018, we use a linear interpolation to estimate the used average PUE.

A.4 Software Used in the Analysis

Data manipulation and statistical analyses have been performed using emacs Org mode 9.1.9, Python 3.8.10 using the pandas version 2.0.3

library and R version 3.6.3 (2020-02-29) with the `ggplot2_3.4.3`, `dplyr_1.1.3`, `lmtest_0.9-40`, `stringr_1.5.0` and `zoo_1.8-12` libraries. Platform: x86_64-pc-linux-gnu (64-bit) Running under: Ubuntu 20.04.6 LTS

B Regression analyses of the observed trends

B.1 Training Energy use over time

We build a linear model to predict model training energy use (in Wh) using the model release date. To ensure a linear relation between both variables, we estimate $\log(\text{model training energy use}) \sim \text{model release date}$ as model training energy use is exponentially distributed. We use weighted least square regression to account for heteroscedasticity due to increasing variance in model training energy use in more recent years. The linear regression analysis revealed a statistically significant model ($F(1,221) = 105$, $p < 2.2 \times 10^{-16}$), with an adjusted R^2 of 0.32, meaning that 32 percent of the variance in the observations is explained by our model. The model equation is $\log(\text{model training energy use}) = -1.49e^{+01} + 8.58e^{-04} \times \text{release date}$ with a standard error of 1.54 for the intercept and $8.38e^{-05}$ for the regression coefficient. This indicates that an increase of 1 for the release date value leads to an average increase of $8.582e^{-04}$ units in $\log(\text{model training energy use})$. This positive relationship between release date and model training energy use was found to be statistically significant ($t(221) = 10.25$, $p < 2 \times 10^{-16}$), affirming the predictive power of release date on model training energy use.

B.2 Training carbon footprint

We build a linear model to predict model training carbon footprint (in kgCO_2 eq) using the model release date. To ensure a linear relation between both variables, we estimate $\log(\text{model training carbon footprint}) \sim \text{model release date}$ as model training carbon footprint is exponentially distributed. We use weighted least square regression to account for heteroscedasticity due to increasing variance in model training carbon footprint in more recent years. The linear regression analysis revealed a statistically significant model ($F(1,221) = 117.7$, $p < 2.2 \times 10^{-16}$), with an adjusted R^2 of 0.35, meaning that 35 percent of the variance in the observations is explained by our model. The model equation is

$\log(\text{model training carbon footprint}) = -1.62e^{+01} + 9.11e^{-04} \times \text{release date}$ with a standard error of 1.55 for the intercept and $8.39e^{-05}$ for the regression coefficient. This indicates that an increase of 1 for the release date value leads to an average increase of $89.11e^{-04}$ units in $\log(\text{model training carbon footprint})$. This positive relationship between release date and model training carbon footprint was found to be statistically significant ($t(221) = 10.85$, $p < 2 \times 10^{-16}$), affirming the predictive power of release date on model training carbon footprint.

B.3 Training metallic resource depletion

We build a linear model to predict model training metallic resource depletion (in kgSb eq) using the model release date. To ensure a linear relation between both variables, we estimate $\log(\text{model training metallic resource depletion}) \sim \text{model release date}$ as model training metallic resource depletion is exponentially distributed. We use weighted least square regression to account for heteroscedasticity due to increasing variance in model training metallic resource depletion in more recent years. The linear regression analysis revealed a statistically significant model ($F(1,221) = 94.31$, $p < 2.2 \times 10^{-16}$), with an adjusted R^2 of 0.30, meaning that 30 percent of the variance in the observations is explained by our model. The model equation is $\log(\text{model training metallic resource depletion}) = -1.59e^{+01} + 8.10e^{-04} \times \text{release date}$ with a standard error of 1.54 for the intercept and $8.34e^{-05}$ for the regression coefficient. This indicates that an increase of 1 for the release date value leads to an average increase of $8.10e^{-04}$ units in $\log(\text{model training metallic resource depletion})$. This positive relationship between release date and model training metallic resource depletion was found to be statistically significant ($t(221) = 9.71$, $p < 2 \times 10^{-16}$), affirming the predictive power of release date on model training metallic resource depletion.

C Sensitivity analysis on the estimated GPU average performance during training

As detailed in Section 3.1 and Appendix A.2, training duration estimates using GPU- h_2 depend on an empirically derived assumption about average GPU performance during the training process. Our

model corresponds to a quasi-constant performance ratio of around 27 %. This section details a sensitivity analysis examining the impact of variations in the used performance ratio on the obtained results and trends.

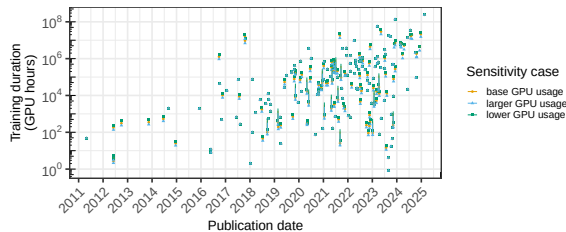


Figure 12: Estimated training duration for models in Epoch AI over time in the three considered graphics card performance scenarios.

We tested two variations corresponding to a lower usage of the cards of 20 % and a larger usage of the cards of 35 %. Larger usage means that the cards are more efficient at using their compute power, corresponding to shorter training duration. Figure 12 details the estimated training duration in the different performance scenarios. As expected, slightly shorter training duration are estimated when considering a larger usage of the cards (−16.5 % in average, compared to the base performance ratio), and slightly longer training duration are estimated when considering a lower usage of the cards (+10.2 % in average, compared to the base performance ratio). The change in hypothesis does not affect the (56 % of) models where GPU- h_1 is used. This variation in estimated training duration leads to slight variations in the final estimated training footprints (respectively −16.7 % and +10.1 % in average **GWP**, compared to the base estimated **GWP** and respectively −18.1 % and +9.5 % in average **ADPe**, compared to the base estimated **ADPe**). These variations do not change conclusions on the observed trends.

Acronyms

ADPe Abiotic Depletion Potential for elements. 3, 5, 6, 7, 16

AI Artificial Intelligence. 1, 2, 3

CI Carbon Intensity. 5, 6, 8

GWP Global Warming Potential. 3, 5, 6, 7, 16

ICT Information and Communication Technology. 2, 3

LCA Life Cycle Assessment. 3

LLMs Large Language Models. 1, 2, 3, 9, 10

ML Machine Learning. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13

NLP Natural Language Processing. 1, 2, 4, 5, 8, 9

PUE Power Usage Effectiveness. 2, 9

TDP Thermal Design Power. 4, 6