Environmental Impact of Training LLMs

Andrew Shi

February 4, 2025

1 Costs Associated with Training and Deployment

The environmental costs associated with LLMs can broadly be divided into three categories: (1) training, which entails using powerful hardware and large datasets to produce foundation models (2) resource allocation, which is the cost associated with data storage and model maintenance and (3) deployment/inference costs, which relates to integrating LLMs into existing systems through techniques like fine-tuning and sampling from the model for production use cases [5]. Large-scale GPU or TPU clusters are used in the training stage, which is computationally-expensive and an energy-intensive process. This results in significant carbon dioxide equivalent (CO_2e) contributions. The inference stage is less energy-intensive, but in certain applications, the frequency of using specific applications may contribute to even higher carbon emissions than that of the training stage [1]. Aside from energy consumption, water consumption is required for data centers hosting LLMs to operate complex cooling systems [5].

1.1 Energy Consumption

As a case study, [4] estimates the carbon footprint of training BLOOM, a 176B parameter language model. Given the advancements in hardware and compute availability, it has become common practice for institutions to train models on scale similar to BLOOM in recent years. The researchers in [4] comprehensively evaluate the environmental costs associated with training LLMs end-to-end, including dynamic consumption (training), embodied emissions (hardware manufacturing for GPUs and servers), idle consumption (energy consumed to keep servers powered and cooled when not actively training), and deployment (inference). Referencing Table 1, although BLOOM and GPT-3 are of similar model sizes, the total CO_2eq in the training stage for BLOOM is significantly lower due to the carbon intensity of the energy source being the dominant factor of carbon emissions (BLOOM was trained on a low-carbon grid in France). Taking into account the entire pipeline for training LLMs, the CO_2eq breakdown is referenced in Table 2.

Model	Parameter Count	Carbon intensity of grid	CO_2eq emissions
BLOOM	176B	57 gCO_2eq/kWh	25 tonnes
Gopher	280B	330 gCO_2eq/kWh	352 tonnes
OPT-175B	175B	231 gCO_2eq/kWh	70 tonnes
GPT-3	175B	429 gCO_2eq/kWh	502 tonnes

Table 1: Total carbon emissions for training LLMs, from [4]

Process	CO_2eq emissions	Percentage of total emissions
Embodied Emissions	11.2 tonnes	22.2%
Dynamic Consumption	24.69 tonnes	48.9%
Idle Consumption	14.6 tonnes	28.9%
Total	**50.5 tonnes**	**100.0%**

Table 2: Breakdown of carbon emissions for training BLOOM, from [4]

During the inference stage for BLOOM, a GPU node of 16 A100 GPUs consumed 914 kWh of electricity over the course of 18 days, equating to about 19 kg of CO_2eq per day [4]. A high proportion of this energy consumption was attributed to simply keeping the model loaded (i.e. idle consumption between requests). Over the course of a year, we calculate almost 7 tonnes of CO_2eq emissions, which is a non-trivial amount.

Overall, training and deploying LLMs creates significant energy consumption costs. The Association of Data Scientists estimates that training GPT-3 consumed roughly the same amount of energy as 120 American households over the course of a year. GPT-4 contained 1.8T parameters, almost 10x the size of GPT-3, which contained 175B parameters.

1.2 Water Consumption

Water consumption is an often overlooked cost of training LLMs. [3] states that water usage from training LLMs could reach 4.2–6.6 billion cubic meters by 2027, which is comparable to the annual water use of entire countries. Water consumption is primarily divided into three scopes: (1) Scope 1 (on-site cooling), which involves evaporating water in cooling towers or using other evaporative air-cooling processes (2) Scope 2 (off-site electricity generation), which involves using water to cool power plants themselves and (3) Scope 3 (embodied water in hardware), which references the need for ultrapure water to manufacturing AI chips and servers.

[3] estimates that training GPT-3 could have consumed over 700,000 liters on-site and up to several million liters total (Scope 1 and Scope 2). Depending on the location, training alone could have used over 5 million liters of evaporated water. For inference, the researchers estimate that a single medium-length query (roughly 1000 words of total input and output) can cost several tens of milliliters of water in both cooling and electricity. The footprint for larger models models like GPT-4 could be significantly higher.

2 Methods to Reduce Carbon Emissions

There have been a multitude of recent efforts to mitigate carbon footprints resulting from training and deploying LLMs. Broadly, there are optimizations for training, hardware, choice of data centers, and deployment/inference.

In the training phase, methods like model pruning, quantization, and distillation can reduce computation needed for training. [2] proposes a unified framework that combines all three (pruning, quantization, and knowledge distillation) to compress models from scratch without sacrificing accuracy, thereby helping lower both carbon and hardware costs. Furthermore, certain architectures like sparse transformers can create free gains in energy reduction without compromising model performance. Certain chips specialized for training LLMs like TPUs and GPUs have been used for efficient

computation, which leads to lower energy consumption.

There has also been a shift to transition data centers to use renewable energy sources like solar and wind power, further minimizing carbon emissions. Innovations in cooling solutions have been developed to reduce water and electricity usage.

Finally, inference efficiency can be improved to reduce energy usage. Techniques like batch processing have been proposed to minimize power consumption for AI services in production.

References

[1] B. Everman, T. Villwock, D. Chen, N. Soto, O. Zhang, and Z. Zong. Evaluating the carbon impact of large language models at the inference stage. In *2023 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pages 150–157, 2023.

[2] J. Kim, S. Chang, and N. Kwak. Pqk: Model compression via pruning, quantization, and knowledge distillation, 2021.

[3] P. Li, J. Yang, M. A. Islam, and S. Ren. Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models, 2025.

[4] A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model, 2022.

[5] A. Singh, N. P. Patel, A. Ehtesham, S. Kumar, and T. T. Khoei. A survey of sustainability in large language models: Applications, economics, and challenges, 2025.