Reducing the Electricity Use of Large Language Models

EMIL TODIRASCU, University of Twente, The Netherlands

The use of artificial intelligence (AI) has dramatically increased over the past few years. With the recent surge of Large Language Models (LLMs) and text-to-image generation models, the general public has begun to see the possibilities of artificial intelligence and use them in their personal and professional lives. A large contributor to this is the advancements in both hardware and software technologies. However, these new technologies require careful consideration regarding their energy consumption. As models become more computationally intensive, their training drastically increases their energy consumption.

Balancing the benefits LLMs can bring to our lives and their energy requirements is essential to ensure that this technological progress does not come at the expense of environmental impact. Therefore, researchers should aim to use efficient techniques that lower the electricity use of such artificial intelligence models. This research aims to create a model of the electricity consumption of training LLMs and explore techniques machine learning researchers should use to reduce the electricity use of training LLMs.

Additional Key Words and Phrases: Large Language Models (LLM), Artificial Intelligence, Energy Consumption, Electricity Use, Natural Language Processing (NLP)

1 INTRODUCTION

Artificial intelligence (AI) is a computer science research field that was formally named an academic field in 1956 [33]. Since then, it has gone through several periods of optimism and disappointment, and more recently, several innovations such as deep learning, and the transformer have brought it back to life, making it one of the most researched scientific disciplines [41, 4].

Artificial intelligence applications range from recommendation systems used on social media and superhuman play in strategy games such as chess or Go [36], to autonomous vehicles and creative tools such as AI art. Most recently, tools like ChatGPT and DALL·E have taken over the world, with ChatGPT being the product with the highest growth rate of new users in its first months [5]. This has therefore further increased the interest and funding in the field.

With this much funding, companies have created very big language models, with up to 70 billion parameters, which consume a lot of energy, spending billions of dollars on hardware alone [23, 20, 39]. Companies, such as Google and Meta, usually use over 10,000 graphics processing units (GPU) to train their AI models, which can use up to 3740 kilowatt-hours per year per GPU. This is as much energy as an entire household for one single GPU [23, 21]. This is an alarmingly high number and begins to make researchers ask themselves if it is even worth the cost.

Data center energy requirements is a thoroughly debated topic. Some recent research shows that data center electricity requirements are expected to increase to up to 321 terawatt-hours (or even 752 terawatt-hours in a worst-case scenario) in 2030 [19, 14]. This would put global data center energy consumption at a total of 2% of the global electricity available. It is hard to estimate the exact increase, but researchers generally agree that it has an upward trend. For this reason, it is essential to better understand why data centers require electricity and how it is used.

The insights from this study help researchers and industry professionals get a better understanding of exactly how a change in the model architecture or training process will affect the total required electricity. By seeing exactly how, for example, the choice of data center impacts the electricity needed, researchers will be able to make more informed and therefore better and more sustainable decisions. Furthermore, this study also provides several solutions as to exactly how to tackle the main problems identified. This should help researchers easily make language models more efficient and environmentally friendly.

2 PROBLEM STATEMENT

As seen above, with the improvement in artificial intelligence software and hardware technologies, comes a great increase in energy consumption. Training big models requires large data centers, which have increased electricity consumption, up to 3740 kilowatt-hours per year per GPU [23].

Data center energy consumption is expected to increase drastically [19, 14], especially with the recent demand for large-scale AI training and inference systems. This trend not only has significant financial implications but raises environmental concerns. Consequently, understanding the energy demands of AI systems and how they can be mitigated, has become a critical area of research.

The purpose of this research is to perform a combination of systematic and unsystematic literature review, create a system dynamics model of the energy consumption of training an LLM, and evaluate the existing options for reducing the energy use of artificial intelligence systems and how they can

41st Twente Student Conference on IT July 5th, 2024, Enschede, The Netherlands.

be applied by researchers. Therefore, the main research question is the following.

What methodologies should researchers use to mitigate the electricity use associated with the training process of large language models?

This will be split into two different sub-research questions that, when put together, should answer the main question. Before looking at reducing electricity use, the main parts that use electricity should be established. This can then be modeled using a system dynamics (SD) model as it is a great tool to showcase the interdependencies in complex systems [8].

Lastly, the research will focus on identifying methodologies to reduce electricity in each specified area of the model and explaining how they can be used by researchers and industry professionals; therefore leading to the following sub-research questions.

1. How can the primary contributors to electricity consumption in the training process of large language models be estimated and modeled under a system dynamics approach?

2. How can the electricity consumption of each identified area in the system dynamics model be decreased?

To address these questions, the paper is divided into the following sections. Section 3 shows the most recent and relevant work that has been done in the fields of AI, AI sustainability and LLM energy use modeling. Subsequently, Section 4 describes the research methods used, delineating exact information such as sources, search queries and exclusion criteria. Section 5 identifies and explains the main aspects that affect the electricity consumption of an LLM and explains the model created. This section also runs three simulations validating the model using real world data from well known language models. Section 6 offers ideas and solutions that AI researchers could implement to reduce the energy use of their LLM. Next, Section 7 addresses the limitations and assumptions of the findings as well as suggestions for possible future research directions. Lastly, Section 8 summarizes the findings and contributions of this research.

3 RELATED WORK

This section will highlight some of the most important research that has been done in the general AI field as well as the energy consumption of AI, specifically large language models.

An important recent discovery is deep learning [15]. This technology was extremely revolutionary and was used in a lot of fields, ranging from computer vision to climate science. Another important recent discovery is the architecture of the transformer [41]. Developed primarily for natural language processing tasks, transformers have demonstrated remarkable capabilities in modeling sequential data and capturing long-range dependencies efficiently. This directly led to the current AI boom that started in 2020. Since then, plenty of new companies and tools have launched that leverage the use of AI. Tools like ChatGPT, DALL·E, Bing Copilot, Tesla Autopilot and many others, all make use of neural networks and variations of the transformer architecture.

In the domain of sustainable AI, one of the most relevant and recent research by Patterson et al [31] suggests that energy consumption of artificial intelligence is actually very low; even creating an interesting comparison, by stating that “the portion of the 22,000 people from 68 countries who in 2019 flew to attend the two major ML conferences (NeurIPS and CVPR) collectively had a CO2e impact arguably had ~10x–100x higher than the impact of training of all the ML models in this paper”. They consider the percentage of total energy use of big tech companies, such as Google, to be very low for machine learning tasks, representing less than 15% of the overall electricity consumption. However, it is important to mention that most authors of the paper are Google employees. Nine out of the eleven authors are high ranking employees. This means that they have a vested interest in portraying this is a small problem. This does not nullify their research, but should be kept in mind when considering their measurements. Moreover, the paper only considers the cost of training and not inference; which has been shown to be highly underestimated in recent research [6].

Another piece of important research, done by Schwarz et al, [43] in the field of artificial intelligence sustainability, suggests that the environmental friendliness of artificial intelligence is heavily understudied. They encourage researchers to pay more attention to this topic and incorporate sustainability considerations into AI development. Their research highlights the use of sustainable practices, advocating for “making efficiency an official contribution in major AI conferences” [43].

In terms of understanding the energy use of an LLM, researchers at Massachusetts Institute of Technology and New York University have conducted experiments to understand the “computational and energy utilization of inference with LLMs.” [34]. In their study, they tried to evaluate inference energy consumption of Llama 65B. They found that the energy required to run inference on it was in the range 300 to 1000 watts. They mention that this depends a lot on the hardware used and the number of GPUs. They also found that the energy per output token was about 3-4 Joules. Additionally, they also suggest that “at a minimum, 8 V100 GPUs each with 32 GB of RAM or 4 A100 GPUs each with

80GB of memory are required for any meaningful inferences with the 65B LLaMA model."

Lastly, another very recent paper [12] takes a higher level approach, investigating the carbon footprint of the entire life-cycle of large language model chatbots. It identifies eight main phases in the life cycle of LLM-powered chatbots, from research and development to hardware manufacturing and waste disposal. It analyzes every single phase in detail and shows how the phases influence one another. It concludes by suggesting three strategic pathways to tackle this issue. The first pathway advocates for systematic and dynamic reporting to accurately estimate the carbon footprints of chatbots. The second pathway suggests an overall greener process: designing greener training and fine-tuning processes for LLMs, incentives for end-of-life management and proactive reporting of energy consumption to avoid lag. The last pathway says governments should establish international non-profit organizations, implement emission legislation and promote international collaboration.

4 METHODOLOGY

This study employs a combination of systematic and unsystematic literature review. The research focuses on synthesizing existing research about electricity consumption of training LLMs. Additionally, a system dynamics approach is used to model the energy use of the LLM.

4.1 Systematic Literature Search

Scopus and Google Scholar were selected as the main databases for the systematic review due to their popularity and substantial information content. In order to assure high quality and relevant resources, the following search queries were used:

("Energy" OR "Electricity") AND ("Artificial Intelligence" OR "Machine Learning" OR "Large Language Model")

"Large Language Model" AND ("Training Cost" OR "Electricity Consumption")

Both queries target the intersection of the two main areas of interest, artificial intelligence and energy consumption analysis. The first one is aimed at more general and high level aspects about the intersection of the domains, while the second is directed towards exactly what is being researched.

To assure low-quality and outdated research is filtered out, several exclusion criteria have been applied. Given that AI is a rapidly evolving area of research and there have been several breakthroughs in the recent past, on which most AI work is based, only recent research (2020 and onwards) was considered. As this domain is heavily studied, the papers were thoroughly checked for relevancy in order to assure they are on topic. Additionally, the articles should be freely available online in order to comply with the "Twente Student Conference on IT" guidelines.

After applying all exclusion criteria, the initial search yielded a small number of relevant articles; only five met all criteria. The strictness of the exclusion criteria was intentional, to ensure only relevant and high quality articles are selected. However, it became apparent that this approach was actually too restrictive. In an attempt to address this, the criteria was slightly loosened, by allowing for papers published before 2020. Now, the opposite problem surfaced, most of the new papers identified were either not relevant to the specific topic of this research or were outdated. For example, due to how fast natural language processing has evolved, even papers from 2018 used old techniques that are not very relevant in today's research. Moreover, the few papers found lacked comprehensive information, especially regarding energy consumption or training costs. Consequently, the limited number of useful papers required an alternative approach to gather the required information.

4.2 Unsystematic Literature Search

As the systematic search proved to not have enough relevant information to be able to come to clear conclusions, some unsystematic literature search was also employed. This allowed for a more flexible and adaptable literature exploration. It consisted of three primary methods: papers received from the research supervisor, LLM technical papers and several other relevant books and articles referenced in papers.

First of all, a collection of academic papers and articles provided by the professor served as foundation and technical background knowledge to better understand the landscape of AI energy consumption. This included recent studies regarding the carbon emissions of machine learning, data center energy use and AI sustainability. This provided a curated starting point for further research.

Secondly, to get a more nuanced and technical understanding of large language models, several technical papers, published by authors of the large language models, were also included in the research process. This includes technical papers from some of the best [18] and recent LLMs such as GPT-3, Gemini 1.5 and LLama 2. This served as a comprehensive and up-to-date view of how modern language models are built and what tools and techniques are used in the process of training them.

Lastly, some of the literature from the above mentioned methods referenced several other key research papers that were highly relevant. This proved to be a highly effective method that unveiled crucial information about topics such as what hardware data centers use, hyperparameters of certain language models, and much more.

4.3 System Dynamics Modelling and Simulation

One of the goals of this research is to create a system dynamics model of the energy consumption of training a large language model. Then, several simulations are run using publicly available data, to validate it.

System Dynamics Modelling is a modeling and simulation tool useful for both experienced researchers and people with little experience in the modeling field. It helps to easily showcase and simulate different scenarios and understand the behavior of a system. Using it will help visualize and explain exactly how energy is used when training LLMs.

To create the model, the online tool Insight Maker is used. It is a free “web-based, general-purpose simulation and modeling tool” [8]. Even though InsightMaker offers three modeling approaches: System Dynamics, Agent-Based Modeling, and imperative programming, for the purpose of this research only the System Dynamics tool is used. It offers several primitives such as stocks, variables, and converters; however, to keep the diagram simple and understandable, only variables, and links are used. Variables are depicted as light orange ovals and are used to show how aspects of the LLM interact with each other, through links (shown as gray dotted lines), and arrive at the final result.

This model is then used to run several simulations using publicly available data to check its validity and accuracy. Section 5.1 dives deeper into the required input data, the calculation of each intermediary variable of the model, and how the final result is calculated.

5 LLM ENERGY USE MODEL

This section is split into two parts. The first subsection will explain the system dynamics model created to estimate the power consumption of training a large language model. Then, the second subsection will validate the model using publicly available data from some of the biggest, open source, and well-known LLMs.

5.1 Explanation of the Model

In order to better understand exactly what requires energy when training large language models, the following system dynamics model was created (Figure 1).

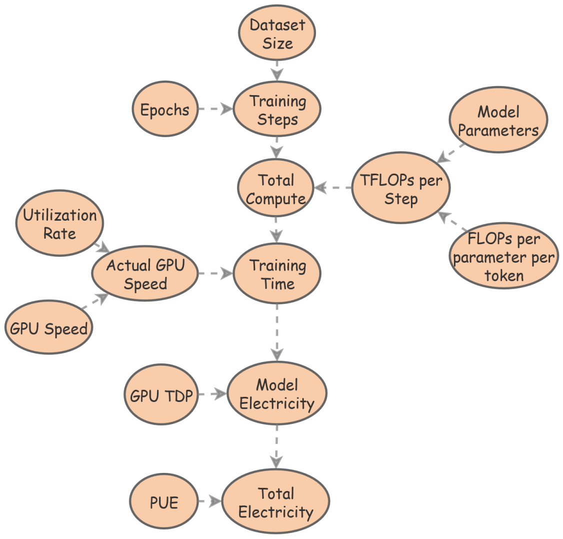

Figure 1. System Dynamics Model of LLM Training Energy Consumption

The goal of the model is to estimate how much electricity a large language model would require to be trained. It takes in 6 input values: PUE, GPU TDP, GPU Speed, Epochs, Dataset Size, and Model Parameters. The final result is the Total Electricity. A full list of the formulas and the corresponding variable can be found in Table 1.

Variable	Formula
Total Electricity	$Model\ Electricity\ \times\ PUE$
Model Electricity	$GPU\ TDP\ \times\ Training\ Time$
Training Time	$Total\ Compute\ \div\ Actual\ GPU\ Speed\ \div\ 3600$
Actual GPU Speed	$Utilization\ Rate\ \times\ GPU\ Speed$
Total Compute	$Training\ Steps\ \times\ TFLOPs\ per\ Step$
Training Steps	$Dataset\ Size\ \times\ Epochs$
TFLOPs per Step	$\frac{FLOPs\ per\ parameter\ per\ token \times Model\ Parameters}{1000}$

Table 1. Formulas for variables of the System Dynamics Model

Starting from the end, Total Electricity represents how much electricity the training of the LLM is expected to consume and is measured in watt-hours (Wh). It is a direct result of the calculated Model Electricity multiplied by the PUE. PUE stands for Power Usage Effectiveness, measuring the efficiency of the data center used to train the model. For an average data center, this value is close to 1.58 and 1.10 for cloud providers [31].

Next, the Model Electricity shows how much energy is needed to train the LLM before accounting for the inefficiency of the data center and is also measured in watt-hours. Multiplying the GPU TDP by the Training Time, results in this value. GPU TDP represents the thermal design power of the graphics processing unit. A graphics processing unit is the device used by data centers to train AI models. Some data centers still use CPUs (central processing unit) but it has been shown that GPUs are drastically more efficient, especially for machine learning tasks [7]. TDP is a measure for the amount of heat that the GPU is designed to produce under heavy use and is a decent approximation for the power required to use it when training AI models [21, 40]. Table 2 shows details about the most commonly used GPUs in big tech data centers.

GPU	FP64	Tensor Core	TDP
Nvidia V100	7 TFLOPS	112 TFLOPS	250 watts
Nvidia A100	9.7 TFLOPS	312 TFLOPS	300 watts
Nvidia H100	25.6 TFLOPS	756 TFLOPS	700 watts

Table 2. Most common GPUs in big tech data centers

Training Time represents how long it would take to train the model on a single GPU - measured in hours. This is calculated by dividing the Total Compute by the Actual GPU Speed and transforming from seconds to hours by further dividing by 3600. The Actual GPU Speed is the GPU Speeds reported by the manufacturer, Nvidia, multiplied by the Utilization Rate. The Utilization Rate shows how efficiently the GPUs are actually being used. Research shows this value is very hard to approximate. It can be anywhere between 20-90% depending on the model, data center architecture and several other factors [10, 11].

AI tasks usually require floating point operations (FLOPs) at different precision levels (64 bits, 32 bits and 16 bits). Occasionally, when very well optimized, it may also make use of tensor cores: specialized hardware in GPUs that can perform mixed precision calculations, such as combining 16-bit floating point precision (FP16) and FP32 [17]. GPUs have different speeds for the different precision levels; see Table 2 for the FP64 and Tensor Core speeds of popular GPUS, measured in tera floating point operations per second (TFLOPS). In practice, models use all of these speeds at different parts of the training [17]. For the purpose of this research, only FP64 is used as that seems to be the most realistic [10,11].

The Total Compute is the number of FLOPs required to fully train the LLM. It is calculated by multiplying the Training Steps by the number or FLOPs per Step. The number of Training Steps represents the total number of times the weights (parameters) of the model are updated. It is calculated by multiplying the Dataset Size (expressed in tokens) by the number of Epochs. An epoch is a pass of the entire training data set through the algorithm. It shows how many times the dataset is used to train the model. It is a hyperparameter (a setting) of the model.

Finally, the number of FLOPs per Step represents how many operations have to be performed in a Forward Pass and Backward Pass of the model. To get this value, the number of Model Parameters is multiplied by the number FLOPs per parameter per token and divided by 1000 to express it in teraFLOPs. The number of parameters is dictated by the architecture and is an input variable.

The number of FLOPs per parameter per token is a measure of the number of operations required for each parameter in the model to process a single token. It can be approximated using OpenAI's scaling law [13, 3] to a value of 6. This does not account for potential optimizations, but for the purpose of this research, this is kept as a constant, 6.

5.2 Model Validation

In order to assure the system dynamics model is correct, it has been validated using publicly available information about some of the most well known and influential LLMs: GPT-3, Llama 2 and Llama 3. To be able to check the validity of the model, the six input values are necessary: PUE, GPU TDP, GPU Speed, Epochs, Dataset Size and Model Parameters; as well as the total electricity required to train the model. These are gathered from technical papers, the source code of the model and other relevant research that analyzes the model and hardware used. Appendix A shows the data used for validating the model and the output (Calculated Electricity).

GPT-3 (Generative Pre-trained Transformer 3) is an LLM developed by OpenAI. It was released by OpenAI in 2020. In 2020, OpenAI also released a research paper where they reveal several key pieces of information about GPT-3 [2]. Using it, as well as several other studies and technical information [31, 25, 28, 17], the needed values were identified. The result of the model was 2.75E+08 watt-hours, while the true value is 1.28E+09.

Llama 2 is an open source and free to use for both research and commercial purposes LLM developed by Meta (formerly Facebook). It was released, along with a technical paper describing its capabilities, specifications and training process, by Meta in July, 2023 [22, 23]. Information gathered from it and details about Meta's data centers and technical hardware details proved enough to use it for validation [20,30]. The result of the model was 6.29E+08 Wh, while the true value is 6.88E+08.

Llama 3, the successor of Llama 2, is also an open source model created by Meta. It was released in April, 2024 in a blog post [23]. Unfortunately, as of writing this paper, Meta has not yet published the technical paper about it, making gathering information about it harder. Nevertheless, sufficient information was found in the source code, information on the data center it was trained in and hardware description [24, 16, 29]. The result of the model was 4.17E+09, while the true value is 4.48E+09.

As can be seen, the system dynamics model's result is very close to the actual electricity consumed in the case of the Llama 2 and Llama 3 models, but around 5 times off in the case of GPT-3. This discrepancy is likely due to factors that are not accounted for; such as model architecture or other aspects of the way the model is trained (i.e. batching or parallelism) [38]. Nonetheless, this model can still be used to approximate the energy use. The goal of the model is not to calculate the electricity exactly, but rather give a fair estimation based on a few parameters.

6 DECREASING ENERGY USE

As the main parts that influence the energy use of training an LLM have been highlighted above, this section will investigate how each identified area of the model can be targeted in order to decrease the resulting energy use.

A big factor that influences the required energy to train the LLM is the power usage effectiveness of the data center where the model is trained. This value is close to 1.58 for the average data center and 1.1 for cloud providers [31, 32]. This means that just by choosing the average cloud provider over the average data center, almost 50% less energy is required. It is a very big difference and therefore highlights the importance of choosing the right data center. As seen in Appendix A, all 3 models used very efficient data centers. GPT-3 used a V100 GPU cluster provided by Microsoft's cloud computing solution Azure [2, 17], while Meta used their own AI Research SuperCluster for Llama 2 [42, 20] and the Meta GenAI Infrastructure for Llama 3 [23, 16, 21]. All of these have excellent PUEs of 1.1, 1.09 and 1.09 respectively.

Another important consideration is to use the best and most recent hardware available for the data center. The most used processors nowadays are Nvidia V100, A100 and H100. It is crucial to choose hardware that is specifically optimized for machine learning tasks as it can improve performance and efficiency by 2 to 5 times [31].

Additionally, to further leverage the potential of the GPU, using mixed precision training to optimize for tensor cores is also a great way to reduce the overall energy use [35, 37, 17, 1]. These can increase the processing power of the GPU by up to 8 times, leading to less time taken to train and therefore less energy consumption.

Also, using the most recent and efficient machine learning models is a good way to further decrease the energy necessary to train the LLM. For example, the Primer architecture is shown to be about 4 times faster than the normal Transformer, without sacrificing quality [31, 37]. Implementing such advanced models not only optimizes performance, but also contributes to the sustainability of AI research.

Moreover, when building the model, the developers should strongly consider the size of the language model, specifically the number of neurons in the network. Recent research shows that "a good portion of neurons are redundant and can be removed to reduce energy consumption without a significant impact on accuracy" [43].

Furthermore, as the budget rises, the model size (parameters number) and the number of training tokens should be scaled equally: "for every doubling of model size the number of training tokens should also be doubled" [25, 9]. This has proven to help both model accuracy and efficiency of training.

7 DISCUSSION

This research has had several assumptions and faced multiple limitations that this section intends to address. The end of this section proposes several future research topics, based on problems faced in this paper.

7.1 Assumptions and Limitations

First of all, as mentioned before, the system dynamics model presented relies on simplified representations of complex interactions between multiple factors. While it helps simplify the model and make it easier to understand, it does omit certain nuances that influence energy consumption. This can also be seen in the validation section; where the numbers do not exactly match with the expected result. Additionally, some values in the model had to be estimated as there is no clear answer in some cases. For example, the GPU processing speed is very dependent on the model architecture and data center configuration and the GPU utilization rate is not a very well researched topic so it's hard to be sure of the estimation. Several different sources use very different values with not much explanation [11, 10].

Second of all, during the validation process, there were several assumptions made. It is assumed that both OpenAI and Meta used the same training practices, which may not be the case as different companies often have proprietary optimization techniques they use. This might overlook specific efficiencies or inefficiencies unique to each company and model.

Next, the model assumes all training is done on one GPU, which in practice is obviously not true and might come with additional costs related to communication and synchronization. This simplification also underestimates the complexity of distributed training. Furthermore, it is also assumed that the model can be scaled up and down infinitely, without any changes in the model architecture or training process. In reality, scaling up introduces new challenges and inefficiencies. "As model size and complexity increase, efficiently scaling training becomes a challenge" [26].

Lastly, this research is constrained by the TScIT requirement that all sources should be openly available to the public. This constraint has lessened the number of sources on top of the fact that, due to the nature of this research domain and how fast it has advanced in the past few years, there is little previous research that is valuable and up to date. Consequently, all of the findings and recommendations are based solely on publicly accessible information, which likely does not capture all current practices and standards.

7.2 Future Work

Future research could focus on perfecting the system dynamics model and creating variations tailored to specific model architectures and training processes. The model created in this research, while not perfectly accurate, serves as a general and comprehensible framework. It should allow other researchers to build upon and customize it to increase its accuracy for different models and training scenarios

Additionally, energy consumption during inference could be another important area to look at. This study only looked at energy costs associated with training; but based on analyzed literature, it became evident that inference is an understudied research area. Given how widespread AI has been deployed and how much people have started using it, understanding and optimizing inference energy consumption is necessary for the development of sustainable AI.

Lastly, another seemingly understudied topic is the utilization rate of GPUs. Most papers analyzed in this research seem to use an arbitrary value for the utilization rate without providing detailed justifications. A more systematic study of GPU utilization rates could provide useful information and help create more accurate energy models. These research directions should provide more in-depth insights into reducing energy demands of AI technologies.

8 CONCLUSION

All in all, the goal of this research was to better understand what aspects of a large language model impact the energy cost when training it and what researchers should do to diminish the overall energy use of training the model. By the use of a system dynamics model, this research has shown a simplified overview of how different factors affect the resulting electricity cost. The model was then validated by the use of publicly available data about some of the biggest and most well-known language models: GPT-3, Llama 2, and Llama 3. Lastly, based on the literature analysis, several methods are proposed to tackle each aspect of the model that contributes to electricity use.

The main contributors to electricity consumption identified in this research are as follows: data center power usage effectiveness, the thermal dynamic power of the GPUs used by the data center, the processing power of the GPUs, epoch count, the dataset size used to train the model and the number of parameters of the model. The way these factors interact with one another and how the final value is calculated can be seen in Figure 1 and Table 1.

To reduce energy use, the following methods have been identified from the literature search. First of all, to get the best possible power usage effectiveness available right now, the use of efficient data centers such as cloud providers, is recommended. It can reduce total electricity by up to 50%. Secondly, choosing the best hardware that is optimized for machine learning tasks is another crucial factor, leading to performance and efficiency increases by up to 5 times. Thirdly, using mixed precision training to optimize for tensor cores can reduce energy consumption by up to 8 times. Lastly, using up-to-date model architectures and scaling the training data set size with the model size equally can further increase model accuracy and efficiency.

This research has highlighted several critical insights into the factors influencing energy consumption in training LLMs. By understanding these contributors, researchers and industry professionals should be able to make more informed decisions, specifically in optimizing data center operations by selecting appropriate hardware and training techniques to minimize energy use. Furthermore, data centers and researchers should be able to better understand and more accurately estimate costs associated with model training, aiding budgeting and resource allocation. Moreover, these insights should encourage the development of more energy-efficient training methodologies, hardware and software, ultimately contributing to more sustainable AI practices.

A Model Validation Values Table

Model	GPU Used	GPU TDP (watts)	GPU Speed (TFLOPS)	PUE	Epochs	Dataset Size (billion tokens)	Model Parameters (billions)	Calculated Electricity (watt-hours)	Actual Electricity (watt-hours)
GPT-3	V100	250	7	1.1	1	300	175	2.75E+08	1.28E+09
LLAMA 2	A100	300	9.7	1.09	1	2000	70	6.29E+08	6.88E+08
LLAMA 3	H100	700	25.6	1.09	1	15000	70	4.17E+09	4.48E+09

Table 3. Values used to validate the System Dynamics model

REFERENCES

During the preparation of this work the author used ChatGPT and Grammarly in order to summarize research and improve writing quality. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

[1] Guangji Bai et al. 2024. Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models. Retrieved June 23, 2024 from http://arxiv.org/abs/2401.00625

[2] Tom B. Brown et al. 2020. Language Models are Few-Shot Learners. Retrieved June 23, 2024 from http://arxiv.org/abs/2005.14165

[3] Adam Casson. 2023. Transformer FLOPs. Retrieved June 23, 2024 from https://www.adamcasson.com/posts/transformer-flops

[4] Helen Crompton and Diane Burke. 2023. Artificial intelligence in higher education: the state of the field. *Int J Educ Technol High Educ* 20, 1 (April 2023), 22. https://doi.org/10.1186/s41239-023-00392-8

[5] Robert Dale. 2020. GPT-3: What's it good for? Retrieved June 23, 2024 from https://towardsdatascience.com/gpt-3-whats-it-good-for-156a445cefc8

[6] Alex De Vries. 2023. The growing energy footprint of artificial intelligence. *Joule* 7, 10 (October 2023), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

[7] Jesse Dodge et al. 2022. Measuring the Carbon Intensity of AI in Cloud Instances. Retrieved June 23, 2024 from http://arxiv.org/abs/2206.05229

[8] Scott Fortmann-Roe. 2014. Insight Maker: A general-purpose tool for web-based modeling & simulation. *Simulation Modelling Practice and Theory* 47, (September 2014), 28–45. https://doi.org/10.1016/j.simpat.2014.03.013

[9] Jordan Hoffmann et al. 2022. Training Compute-Optimal Large Language Models. Retrieved June 23, 2024 from http://arxiv.org/abs/2203.15556

[10] Myeongjae Jeon et al. 2019. Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads. (July 2019). Retrieved June 23, 2024 from https://www.usenix.org/system/files/atc19-jeon.pdf

[11] Shixin Ji et al. 2024. Towards Data-center Level Carbon Modeling and Optimization for Deep Learning Inference. Retrieved June 23, 2024 from http://arxiv.org/abs/2403.04976

[12] Peng Jiang et al. 2024. Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots. *Engineering* (April 2024), S2095809924002315. https://doi.org/10.1016/j.eng.2024.04.002

[13] Jared Kaplan et al. 2020. Scaling Laws for Neural Language Models. Retrieved June 22, 2024 from http://arxiv.org/abs/2001.08361

[14] Martijn Koot and Fons Wijnhoven. 2021. Usage impact on data center electricity needs: A system dynamic forecasting model. *Applied Energy* 291, (June 2021), 116798. https://doi.org/10.1016/j.apenergy.2021.116798

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (May 2015), 436–444. https://doi.org/10.1038/nature14539

[16] Kevin Lee, Adi Gangidi, and Mathew Oldham. Building Meta's GenAI Infrastructure. Retrieved June 23, 2024 from https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

[17] Chuan Li. 2020. OpenAI's GPT-3 Language Model: A Technical Overview. Retrieved June 23, 2024 from https://lambdalabs.com/blog/demystifying-gpt-3

[18] LMSYS. 2024. LMSYS Chatbot Arena Leaderboard. Retrieved June 23, 2024 from https://chat.lmsys.org/?leaderboard

[19] Eric Masanet et al. 2024. Recalibrating global data center energy-use estimates. Retrieved June 23, 2024 from https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf

[20] Meta. 2022. Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research. Retrieved June 23, 2024 from https://ai.meta.com/blog/ai-rsc/

[21] Meta. 2023. Data centers. Retrieved June 23, 2024 from https://sustainability.fb.com/data-centers/

[22] Meta. 2023. Llama 2: open source, free for research and commercial use. Retrieved June 23, 2024 from https://llama.meta.com/llama2/

[23] Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Retrieved June 23, 2024 from https://ai.meta.com/blog/meta-llama-3/

[24] Meta. The official Meta Llama 3 GitHub site. Retrieved June 23, 2024 from https://github.com/meta-llama/llama3

[25] Shervin Minaee et al. 2024. Large Language Models: A Survey. Retrieved June 23, 2024 from http://arxiv.org/abs/2402.06196

[26] Maxim Naumov et al. 2020. Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems. Retrieved June 23, 2024 from http://arxiv.org/abs/2003.09518

[27] Nvidia. 2023. Convolutional Layers User's Guide. Retrieved June 23, 2024 from https://docs.nvidia.com/deeplearning/performance/dl-performance-convolutional/index.html

[28] Nvidia. NVIDIA V100 TENSOR CORE GPU. Retrieved June 23, 2024 from https://www.nvidia.com/en-us/data-center/v100/

[29] Nvidia. NVIDIA H100 Tensor Core GPU. Retrieved June 23, 2024 from https://www.nvidia.com/en-us/data-center/h100/

[30] Nvidia. NVIDIA A100 Tensor Core GPU. Retrieved June 23, 2024 from https://www.nvidia.com/en-us/data-center/a100/

[31] David Patterson et al. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Retrieved June 23, 2024 from http://arxiv.org/abs/2204.05149

[32] David Patterson et al. 2021. Carbon Emissions and Large Neural Network Training. Retrieved from http://arxiv.org/abs/2104.10350

[33] Stuart J. Russell, Peter Norvig, and Ernest Davis. 2010. *Artificial intelligence: a modern approach* (3rd ed ed.). Prentice Hall, Upper Saddle River.

[34] Siddharth Samsi et al. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. Retrieved June 23, 2024 from http://arxiv.org/abs/2310.03003

[35] Valerie Sarge et al. 2019. Tips for Optimizing GPU Performance Using Tensor Cores. Retrieved June 23, 2024 from https://developer.nvidia.com/blog/optimizing-gpu-performance-tensor-cores/

[36] David Silver et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (January 2016), 484–489. https://doi.org/10.1038/nature16961

[37] David R. So et al. 2022. Primer: Searching for Efficient Transformers for Language Modeling. Retrieved June 23, 2024 from http://arxiv.org/abs/2109.08668

[38] Jovan Stojkovic et al. 2024. Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. Retrieved June 23, 2024 from http://arxiv.org/abs/2403.20306

[39] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. Retrieved June 23, 2024 from http://arxiv.org/abs/1906.02243

[40] Yifan Sun et al. 2020. Summarizing CPU and GPU Design Trends with Product Data. Retrieved June 23, 2024 from http://arxiv.org/abs/1911.11313

[41] Rob Toews. 2023. Transformers Revolutionized AI. What Will Replace Them? Retrieved June 23, 2024 from https://www.forbes.com/sites/robtoews/2023/09/03/transformers-revolutionized-ai-what-will-replace-them/

[42] Hugo Touvron et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. Retrieved June 23, 2024 from http://arxiv.org/abs/2307.09288

[43] Roberto Verdecchia, June Sallou, and Luís Cruz. 2023. A systematic review of Green AI. *WIREs Data Min & Knowl* 13, 4 (July 2023), e1507. https://doi.org/10.1002/widm.1507