The Environmental Consequences of Generative AI LLMs

Luke Van Der Veen

1.Introduction

Generative AI (Gen AI) particularly AI chatbots have become a transformative technology across multiple sectors, with applications in education, business and research, allowing for personalised learning, content creation and automated services. However while the benefits of gen AI are widely recognised, its environmental implications are largely unexplored. This report shall highlight the ecological footprint of these gen AI models including its intensive training, resource consumption and hidden sustainability costs.

Luccioni, Strubell and Kreutzer (2023) [1] highlight that general purpose Gen AI models such as these can produce a significant carbon footprint even after its training, with 1000 image generations creating just over 1kg of CO2. Generative Pre-trained Transformers (GPT) are becoming the most well known form of gen AI with OpenAI's ChatGPT being the most well known and widely used with 3.8 Billion visits in the month of January 2025 [2] illustrating the demand for these models. Computational power is required for these models, placing a heavy reliance on large data centers which require cooling and consume a substantial amount of electricity.

With global use growing, the demand for newer, more efficient and more powerful models, there is bound to be a strain on the environment, especially with the high computational demands for these models training and inferencing. This report shall explore the environmental landscape for these models.

2.Literature review

Gen AI chatbots are not a new phenomenon. The earliest chatbot, ELIZA, was invented in the 1960s and was designed to be a command line therapist for psychiatric patients [3]. Over the following decades digital assistants such as Apple's Siri and the Google Assistant advanced the conversational capabilities of chatbots. However the most significant and recent breakthrough in the fields occurred with the introduction of large language models (LLMs), which have significantly enhanced chatbots. This sudden explosion in LLM Gen AI was Google's introduction of the Transformer model, making strides in natural language processing [4] by enabling more efficient training on vast datasets. This innovation laid the foundation for modern models such as ChatGPT, Claude AI and DeepSeek. However as these grow their computational and energy demands escalate raising concerns on their environmental sustainability.

The expansion of Gen AI driven technologies has significantly increased global resource consumption. Its operation requires substantial electricity, cooling, water use, carbon emissions and land allocation for data centers. Running these large Gen AI models requires large data centres to power the majority of the AI workloads, which have high electricity and water demands. US data centers report that AI specialised servers containing special processing units such as Graphical Processing Units (GPU) and Tensor Processing Units (TPU) consumed 176 TWh of power in 2023, accounting for 4.4% of the US's total energy use. This is projected to rise to between 6-12% of the US's total energy usage by 2028. Most AI is run on hyperscale data centers which account for 84% of total data center water usage in the US, the total being 66 billion liters. By 2028 water usage for these centers is projected to grow to between 60-124 billion liters alone[5]. 51% of hyperscale data centers

reside in the US, with China at 16% [6] making the US the majority consumer of resources. With popularity in Gen AI chatbots rising, the demand for these large data centers will only increase, causing these predicted resource usages to be so high. This needs to be mitigated as growth such as this will only increase without proper regulation. However it is difficult to calculate how much of these AI servers are utilised for Gen AI chatbot models. Data centers categorise AI workloads collectively, lacking specificity, making regulation difficult to effectively assign for specific fields. Without clear data on Gen AI energy consumption, policymakers may struggle to develop effective sustainability strategies.

Regulatory efforts have emerged to address the ethical, social and environmental implications of AI [7]. With the European Union (EU) pioneering the charge in regulations for AI with its EU AI Act [8]. This piece of legislation extensively covers AI using a risk based approach, defining levels of risk for social safeties; harmful content generation, cognitive manipulation and biometric categorisation of people. In addition to this it demands transparency of Gen AI, requiring high risk AI systems such as ChatGPT to have logging capabilities to record energy consumption, allowing providers to assess environmental impacts [9]. However, regulations and related penalties for Gen AI will only start being applied in August 2025 and only applies to providers in the European Market [10]. Any model that has already been released and is currently being developed for release before this time will not need to adhere to these environmental requirements. In contrast the US is making strides in sustainable AI legislation, ensuring that energy used to power data centers are fueled by clean energy, put in place in January this year [11]. This Executive Order is especially effective as one of the largest Gen AI LLM providers, Open AI, utilises Microsoft's Azure data centers to train and run their ChatGPT product, some of which reside in the US, ensuring that those which use these data centers utilise clean energy [12]. However Azure datacenters are located all over the world, so there is no guarantee that OpenAI's model training is completely green energy as the Executive Order only applies to data centers in the US. Like the EU AI Act it covers AI as a whole not just the specific LLMs that this report focuses on. Despite efforts from both the EU and US, regulations face enforcement challenges, particularly as AI companies distribute workloads across the globe, for instance OpenAI utilises Azure servers which operate world wide. Without international cooperation and standardised energy frameworks, current policies risk failing to meaningfully reduce AI's global environmental footprint

Training LLMs such as chatbots is an exceptionally resource intensive process, requiring high performance hardware and months of continuous computation. Open AI's ChatGPT 4 for instance, is estimated to have taken up to three months of continuous processing with 25,000 Nvidia GPUs[13], consuming an immense amount of energy, contributing significantly to greenhouse emissions. If all of its training was located in the US, a considerable amount of energy used would be clean energy such as nuclear, geothermal or solar. However as OpenAI's servers are distributed globally the extent of this carbon footprint remains unknown. Furthermore, Open AI's technical report on the model doesn't disclose any of its training and energy costs due to concerns that the industry is highly competitive and therefore revealing this information would create a large advantage for competitors[14]. Beyond training, model inferencing, where AI generates responses based on prior learning, also consumes substantial energy. Luccioni, Strubell and Kreutzer (2023) performed a study analysing the electricity used by inferences. Findings showed that every 1000 text to image inferences, on average 2.907KWh was utilised, that's almost 3000 times more costly than text classification, suggesting that images are actually worth 3000 words[1]. This both demonstrate that chatbot models such as these are energy expensive both during training and post launch. Inferencing with millions of global users, would produce millions of KWh every month as each user will be performing multiple inferences a day. However, reported costs for LLMs remain inconsistent and understated. Deepseek claims to have developed a model comparable to ChatGPT 4 for only $6 million, yet leaked information estimates have set the true training price to be $2.24 Billion factoring in hardware and electricity costs[15]. The current trajectory for LLM training and inference costs is unsustainable, underscoring the urgent need for energy efficient AI architectures and improved Gen AI energy reporting.

Bashir, N. et al. (2024), pointed out that with this unfettered growth in development and rising popularity, the GPT market is becoming its own economic sector, requiring its own monitoring, considerations and restrictions for training methods[15]. Given the prohibitive cost of training AI models from scratch researchers have explored techniques such as fine-tuning. Fine-tuning is an algorithmic method that alters already trained models in order to save resources and improve efficiencies. Fine-tuning allows pre-trained models to be adapted to new tasks and datasets, lowering energy costs[16]. Models such as Meta's LLaMA have utilised fine-tuning techniques to improve model performance. However major architectural changes like the going from ChatGPT 3 to ChatGPT 4 still requires full training from scratch, as they typically require much larger datasets, eliminating any possible recyclability in training costs. As demand for new models increases every year as popularity and demand for development rises, resource inefficiency and electronic waste will accumulate unsustainably, while the older models will be made redundant. Without advancements, the continuous cycle of model replacement will place an ever increasing strain on global resources.

3.Research Gap

Despite the countless commercial chatbot models, there remains a glaring lack of transparency in regards to environmental impacts. Major model companies including OpenAI, Google DeepMind and Anthropic do not publicly disclose metrics such as training datasets, total training time, server use, hardware specifications and energy consumption. This opacity in reporting is largely attributed to the competitive nature of the market, creating disadvantages for releasing proprietary advancements. Secrecy comes with a cost, environmental accountability becomes difficult to enforce due to lack of accurate data, creating a curtain hiding the consequences of these large models.

The absence of standardised reporting has made it difficult to ensure accurate records of resources used to create these models. The introduction of legislation such as the EU AI Act soon to be enforced, will ensure that high risk AI systems such as chatbot models akin to ChatGPT and Claude AI will have logs of energy consumption. Currently model reports and evaluations predominantly focus on the improvements made from the previous models [14, 17]. However this will only affect European markets and we do not yet know how far it can reach in terms of accessing, which servers will be used and where geographically. US based data centers ensure their servers utilise clean energy sources [12], however these globally recognised models utilise servers world wide, so ensuring that logging extends to servers used and in what regions remains unknown. The EU AI Act doesn't cover resource usage post launch, as its main role is to ensure safety of these high risk systems as opposed to exclusively ensuring environmental responsibility. Implying that any resources used during post launch inferencing by millions of users will not be logged, further masking the extent resources are used for these large scale chatbots.

Strubell et al (2019) examined the environmental impact of training large neural networks for natural language processing tasks[18]. This study quantified energy consumption and carbon emissions associated with developing these learning models, demonstrating the level of transparency which can be achieved. However this study doesn't cover the specific area that this report is attempting to highlight, chatbot LLMs. Similarly Belkhir & Elmeligi (2021) broader environmental footprint, further highlighting that data centers are by far the most energy intensive segment of the AI sector. Their findings highlight that this energy use is predicted to rise, further reinforcing the demand for standardised reports on energy use and other environmental logging. While their study provides valuable insights, it looks into AI as a whole and not specifically these LLM chatbots. While there are studies demonstrating estimates for AI related energy usage, there remains no comprehensive, industry verified reports detailing the full environmental costs to an accurate enough standard.

Therefore a gap still remains. There is a need for an end to end environmental report for chatbot models; training, inferencing and maintaining post launch covering the energy used by servers and each inference by every user. Addressing this gap is essential for developing a sustainable future for this sector and ensuring environmental transparency.

4. Future Research

4.1 Proposed Question

Gaps in research suggest there is a need for end to end tracking of resources utilised when training and maintaining a post launch for public use LLM chatbots. This paper proposes this question for future research: Is it possible to completely monitor the resources used when training and maintaining a LLM model pre and post launch? This question aims to cover the electricity and water usage of LLM chatbots during their training on large datasets where they are running in large data centers on GPU enabled servers. Furthermore it will monitor these resources after it has launched for public use so reports can be made on inferences made by users over its public use. Ensuring that a holistic view on utilised resources is reported on.

4.2 Research Design

This question would be expanded into the following study. A LLM chatbot shall be built, trained and launched for public use, and a real time monitoring framework for tracking energy and water at each stage of the models deployment. The model will be able to engage in natural language tasks and general purpose interactions, shall utilise neural networks for learning and shall be trained on large datasets, such as common crawl [20] which contains terabytes of open source web data. Training shall be performed on 1 or more GPU enabled servers, the number of which shall be recorded to track how much power is used. Providers such as Amazon Web Services or Microsoft Azure utilise the same hardware (NVIDIA H100/A100) that companies such as OpenAI utilise for training, ensuring consistency with other models in the industry. Azure servers report a water usage efficiency (water used per KWh) of 0.02 in Ireland, however this does vary based on the country used, so the country the server is located in shall be noted for the study to ensure accurate report on water usage [21]. Training time shall be recorded in hours the server is utilised for training the model, this is so it can easily scale with the KWh measurement. After the model has been trained and is ready for use, it shall be distributed to 1000 randomly selected students for a trial usage period of a month. Students were selected due to their high levels of engagement with LLM models for research and availability, future research could expand this demographic. During this month, server use shall be monitored, KWhs shall be recorded in order to establish how much power it takes for 1000 users performing inferences on a LLM chatbot model in their day to day lives. This can then be scaled to an industrial level. For example if 1000 users over 1 month use 600 KWhs (arbitrary guess) of power and 12Ls of water (if servers are based in Ireland), scale this up to 100 million people (the user base for chatGPT in its first 2 months of use) that would use 60 million KWhs and 1.2 million litres of water. Both of these recordings will allow a baseline to be created in order to gauge more accurate estimates on not only training but post launch inferences which aren't reported on for LLM models. The closest study we have to this metric is Luccioni, Strubell and Kreutzer (2023) which recorded power used for specific inference types over 1000 inferences. This doesn't consider the cumulative impact of continuous engagement which is critical for understanding real world energy usage, not allowing for the industrial scale this study could produce.

5. Lay summary for science communication to the large public

The rapid growth of generative AI, particularly large language model chatbots, in recent years has transformed the sector. They offer remarkable, easy to use capabilities but they also present significant environmental challenges. This report explored the resource consumption of these models and the dominant causes; training on large datasets, lack of model recyclability, lack of documentation on resource consumption, GPU enabled data center energy consumption and lack of legislation coverage. With AI development accelerating the need for sustainable actions is higher than ever, action needs to be taken before it is too late. A key takeaway is the hidden costs of these models, not just during the intensive training process, but post launch, where millions of users utilise these models day in day out. Research has shown that data centers consume a vast amount of power and water already, with predictions showing a steep climb in consumption. With transparency remaining a gap in the sector and companies such as OpenAI and Anthropic not disclosing energy and water footprints, there is no accountability for these consumed resources and therefore no responsibility for sustainability. To address this issue this report proposes the building, training and the deployment of a fully transparent LLM chatbot model, in order to track resources utilised throughout each stage of the models lifespan. This establishes a baseline to compare and scale to larger companies, comparing to global uses of these models. Providing valuable benchmarks against the industry will allow a curtain to be lifted on the realistic amount of resources consumed to establish these models which are used by millions of users every day. This research is crucial for both the scientific community and the public eye. The rise of these models affect everyone, yet few are aware of the sustainability challenges it presents. Raising awareness can push for standardised reporting on end to end environmental costs of these models covering training to launch, more efficient energy consumption and policy changes encouraging transparency and sustainability. If AI companies and data center providers collaborate to provide real time resource monitoring and reporting the world can move to a future where AI innovation comes at a reduced resource cost. AI is here to stay no matter what and it is expensive to run, but how we manage its growth and consumption will determine its long term sustainability. Through rigorous research, transparency and collaborative efforts between companies and governments around the world, these models can be created to be intelligent and sustainable.

References

1. Luccioni, A., Strubell, E. and Kreutzer, J. (2023) ‘The environmental impact of generative AI’, *arXiv preprint*, Available at: https://arxiv.org/pdf/2311.16863 (Accessed: 05/02/2025)
2. chatgpt.com Website Analysis for January 2025 (Similarweb 2025) Available at: https://www.similarweb.com/website/chatgpt.com/#traffic (Accessed: 11/02/2025)
3. LI Academy (2024.) ‘The story of ELIZA: The AI that fooled the world’. Available at: https://liacademy.co.uk/the-story-of-eliza-the-ai-that-fooled-the-world/ (Accessed: 20/02/2025).
4. Vaswani, A. et al. (2017) ‘Attention is all you need’, *arXiv preprint*. Available at: https://arxiv.org/pdf/1706.03762 (Accessed: 11/02/2025).
5. U.S. Department of Energy (2024) 2024 United States Data Center Energy Usage Report. Washington, D.C.: Lawrence Berkeley National Laboratory. Available at: https://escholarship.org/uc/item/32d6m0d1 (Accessed: 20/02/2025).
6. SRG Research (2024) ‘Hyperscale data centers hit the thousand mark: Total capacity is doubling every four years’. Available at: https://www.srgresearch.com/articles/hyperscale-data-centers-hit-the-thousand-mark-total-capacity-is-doubling-every-four-years (Accessed:20/02/2025).
7. Tan, Y. H., et al. (2024) ‘Current landscape of generative AI: Models, applications, regulations and challenges’, *IEEE Xplore*. Available at: https://ieeexplore.ieee.org/document/10696569 (Accessed: 20/02/2025).
8. European Parliament (2023) ‘EU AI Act: first regulation on artificial intelligence’. Available at: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#ai-regulation-in-europe-the-first-comprehensive-framework-4 (Accessed: 20/02/2025).
9. Heinrich Böll Stiftung (2024) ‘EU AI Act: Missed opportunity?’. Available at: https://eu.boell.org/en/2024/04/08/eu-ai-act-missed-opportunity (Accessed: 20/02/2025).
10. Global Campus of Human Rights (2024) ‘How is ChatGPT regulated by the EU AI Act? Reflections on higher education’. Available at: https://www.gchumanrights.org/preparedness/how-is-chatgpt-regulated-by-the-eu-ai-act-reflections-on-higher-education (Accessed: 22/02/2025).
11. The White House (2025) ‘Executive Order on Advancing United States Leadership in Artificial Intelligence Infrastructure’. Available at: https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2025/01/14/executive-order-on-advancing-united-states-leadership-in-artificial-intelligence-infrastructure (Accessed: 23/02/2025).
12. Microsoft (2025) ‘Microsoft and OpenAI evolve partnership to drive the next phase of AI’. *Microsoft Blog*. Available at: https://blogs.microsoft.com/blog/2025/01/21/microsoft-and-openai-evolve-partnership-to-drive-the-next-phase-of-ai/

(Accessed: 23/02/2025).
13. Seifeur (2023) ‘GPT-4 training time’.
    Available at: https://seifeur.com/gpt-4-training-time/
    (Accessed: 23/02/2025).
14. OpenAI (2023) *GPT-4 Technical Report*.
    Available at: https://cdn.openai.com/papers/gpt-4.pdf
    (Accessed: 23/02/2025).
15. Bashir, N. et al. (2024) 'The Climate and Sustainability Implications of Generative AI', An MIT Exploration of Generative AI [Preprint]. doi:10.21428/e4baedd9.9070dfe7.
16. Patterson, D., et al. (2021) ‘Carbon emissions and large neural network training’, *arXiv preprint*. Available at: https://arxiv.org/pdf/2104.10350
    (Accessed: 28/02/2025).
17. Metr (2025) ‘Claude 3.5 Sonnet Report’.
    Available at: https://metr.github.io/autonomy-evals-guide/claude-3-5-sonnet-report/
    (Accessed: 02/03/2025).
18. Strubell, E., et al. (2019) ‘Energy and policy considerations for deep learning in NLP’, *Energy Policy*, 37, pp. 1–14.
    Available at: https://www.sciencedirect.com/science/article/pii/S221067072100041X
    (Accessed: 10/03/2025)
19. Microsoft (2021) *Microsoft Conversational AI*.
    Available at:
    https://learning.oreilly.com/library/view/microsoft-conversational-ai/9781484268377/html/495919_1_En_5_Chapter.xhtml
    (Accessed: 10/03/2025)
20. Common Crawl (2025) *Common Crawl Dataset*.
    Available at: https://commoncrawl.org
    (Accessed: 10/03/2025)
21. Microsoft (2025) *Microsoft Data Centers Sustainability and Efficiency*.
    Available at: https://datacenters.microsoft.com/sustainability/efficiency/
    (Accessed: 10/03/2025).