Please cite this article in press as: de Vries, The growing energy footprint of artificial intelligence, Joule (2023), https://doi.org/10.1016/j.joule.2023.09.004

Joule

CellPress

Commentary

The growing energy footprint of artificial intelligence

Alex de Vries[1,2,3,*]

Alex de Vries is a PhD candidate at the VU Amsterdam School of Business and Economics and the founder of Digiconomist, a research company dedicated to exposing the unintended consequences of digital trends. His research focuses on the environmental impact of emerging technologies and has played a major role in the global discussion regarding the sustainability of blockchain technology.

Introduction

Throughout 2022 and 2023, artificial intelligence (AI) has witnessed a period of rapid expansion and extensive, large-scale application. Prominent tech companies such as Alphabet and Microsoft significantly increased their support for AI in 2023, influenced by the successful launch of OpenAI's ChatGPT, a conversational generative AI chatbot that reached 100 million users in an unprecedented 2 months. In response, Microsoft and Alphabet introduced their own chatbots, Bing Chat and Bard, respectively.[1] This accelerated development raises concerns about the electricity consumption and potential environmental impact of AI and data centers. In recent years, data center electricity consumption has accounted for a relatively stable 1% of global electricity use, excluding cryptocurrency mining. Between 2010 and 2018, global data center electricity consumption may have increased by only 6%.[2] There is increasing apprehension that the computational resources necessary to develop and maintain AI models and applications could cause a surge in data centers' contribution to global electricity consumption.

This commentary explores initial research on AI electricity consumption and assesses the potential implications of widespread AI technology adoption on global data center electricity use. The piece discusses both pessimistic and optimistic scenarios and concludes with a cautionary note against embracing either extreme.

AI and energy consumption

AI refers to a range of technologies and methods that enable machines to exhibit intelligent behavior. Within this domain, generative AI, used for creating new content (e.g., text, images, or videos), has prominent examples, such as the text-generating tool ChatGPT and OpenAI's DALL-E, a tool popularized in 2022 that transforms text prompts into images. Both these tools use natural language processing, and although they employ distinct techniques, they share a common process: an initial training phase followed by an inference phase.

The training phase of AI models, often considered the most energy intensive, has been the focus of sustainability research in AI.[3] In this stage, an AI model, such as ChatGPT's, is fed large datasets. The model's initially arbitrary parameters are adjusted to align the predicted output closely with the target output. For large language models (LLMs) such as Generative Pre-trained Transformer 3 (GPT-3), from which ChatGPT was developed as a specialized variant, this process results in learning to predict specific words or sentences based on given context. Once deployed, these parameters guide the model's behavior. Hugging Face reported that its BigScience Large Open-Science Open-Access Multilingual (BLOOM) model consumed 433 MWh of electricity during training.[4] Other LLMs, including GPT-3, Gopher and Open Pre-trained Transformer (OPT), reportedly used 1,287, 1,066, and 324 MWh, respectively, for training. Each of these LLMs, was trained on terabytes of data and has 175 billion or more parameters.

Following training, models are deployed into a production environment and begin the inference phase, where they generate outputs based on new data. For a tool such as ChatGPT, this phase involves creating live responses to user queries. The inference phase has received relatively little attention in literature concerning AI's environmental sustainability. In a systemic literature review on this topic, Verdecchia et al. (2023)[3] found that out of 98 papers since 2015, only 17 papers focused on the inference phase, whereas 49 focused on the training phase. However, there are indications that the inference phase might contribute significantly to an AI model's life-cycle costs. Research firm SemiAnalysis suggested that OpenAI required 3,617 of NVIDIA's HGX A100 servers, with a total of 28,936 graphics processing units (GPUs), to support ChatGPT, implying an energy demand of 564 MWh per day.[5] Compared to the estimated 1,287 MWh used in GPT-3's training phase, the inference phase's energy

[1]School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[2]Digiconomist, Almere, the Netherlands

[3]De Nederlandsche Bank, Amsterdam, the Netherlands

*Correspondence: alex@digiconomist.net

https://doi.org/10.1016/j.joule.2023.09.004

Please cite this article in press as: de Vries, The growing energy footprint of artificial intelligence, Joule (2023), https://doi.org/10.1016/j.joule.2023.09.004

demand appears considerably higher. Furthermore, Google reported that 60% of AI-related energy consumption from 2019 to 2021 stemmed from inference.[2] Google's parent company, Alphabet, also expressed concern regarding the costs of inference compared to the costs of training.[6] However, contrasting data from Hugging Face indicates that the BLOOM model consumed significantly less energy during inference compared to the training phase.[4] Various factors, such as AI models' retraining frequency and the trade-off between model performance and energy consumption ultimately influence this ratio. It remains an open question how the inference phase generally compares to the training phase in terms of electricity consumption, as the current literature offers minimal additional insights into the relative weight of each phase.[3] Future studies should therefore thoroughly examine all life cycle stages of an AI model.

Future energy footprint development

The 2023 AI boom has led to an increased demand for AI chips. In August 2023, chip manufacturer NVIDIA reported a record AI-driven second-quarter revenue of $13.5 billion for the 3 months concluding in July 2023.[7] In particular, the 141% increase in the company's data center segment compared to the previous quarter underscores the burgeoning demand for AI products, potentially leading to a significant rise in AI's energy footprint. For example, companies such as Alphabet's Google could substantially increase their power demand if generative AI is integrated into every Google search. SemiAnalysis estimated that implementing AI similar to ChatGPT in each Google search would require 512,821 of NVIDIA's A100 HGX servers, totaling 4,102,568 GPUs.[5] At a power demand of 6.5 kW per server, this would translate into a daily electricity consumption of 80 GWh and an annual consumption of 29.2 TWh. New Street Research independently arrived at similar estimates, suggesting that Google would need approximately 400,000 servers,[8] which would lead to a daily consumption of 62.4 GWh and an annual consumption of 22.8 TWh. With Google currently processing up to 9 billion searches daily, these scenarios would average to an energy consumption of 6.9–8.9 Wh per request. This estimate aligns with Hugging Face's BLOOM model, which consumed 914 kWh of electricity for 230,768 requests,[4] averaging to 3.96 Wh per request.

Alphabet's chairman indicated in February 2023 that interacting with an LLM could "likely cost 10 times more than a standard keyword search.[6]" As a standard Google search reportedly uses 0.3 Wh of electricity,[9] this suggests an electricity consumption of approximately 3 Wh per LLM interaction. This figure aligns with SemiAnalysis' assessment of ChatGPT's operating costs in early 2023, which estimated that ChatGPT responds to 195 million requests per day, requiring an estimated average electricity consumption of 564 MWh per day, or, at most, 2.9 Wh per request. Figure 1 compares the various estimates for the electricity consumption of interacting with an LLM alongside that of a standard Google search.

These scenarios highlight the potential impact on Google's total electricity consumption if every standard Google search became an LLM interaction, based on current models and technology. In 2021, Google's total electricity consumption was 18.3 TWh, with AI accounting for 10%–15% of this total.[2] The worst-case scenario suggests Google's AI alone could consume as much electricity as a country such as Ireland (29.3 TWh per year), which is a significant increase compared to its historical AI-related energy consumption. However, this scenario assumes full-scale AI adoption utilizing current hardware and software, which is unlikely to happen rapidly. Even though Google Search has a global reach with billions of users, such a steep adoption curve is unlikely. Moreover, NVIDIA does not have the production capacity to promptly deliver 512,821 A100 HGX servers, and, even if it did, the total investment for these servers alone for Google would total to approximately USD 100 billion.[5] Over 3 years, the annual depreciation costs on a USD 100 billion AI server investment would add up to USD 33.33 billion. Such hardware expenses alone would significantly impact Google's operating margin. Google Search generated revenues of USD 162.5 billion in 2022, while Alphabet, Google's parent company, reported an overall operating margin of 26%. For Google Search, this would translate to an operating margin of USD 42.25 billion. The hardware costs, coupled with additional billions in electricity and other costs, could rapidly reduce this operating margin to zero. In summary, while the rapid adoption of AI technology could potentially drastically increase the energy consumption of companies such as Google, there are various resource factors that are likely to prevent such worst-case scenarios from materializing.

A more pragmatic projection of worldwide AI-related electricity consumption could be derived from NVIDIA's sales in this segment. Given its estimated 95% market share in 2023, NVIDIA leads the AI servers market. The company is expected to deliver 100,000 of its AI servers in 2023.[10] If operating at full capacity (i.e., 6.5 kW for NVIDIA's DGX A100 servers and 10.2 kW for DGX H100 servers), these servers would have a combined power demand of 650–1,020 MW. On an annual basis, these servers could consume up to 5.7–8.9 TWh of electricity. Compared to the historical estimated annual electricity consumption of data centers, which was 205 TWh,[2] this is almost negligible. Furthermore, the supply

Please cite this article in press as: de Vries, The growing energy footprint of artificial intelligence, Joule (2023), https://doi.org/10.1016/j.joule.2023.09.004

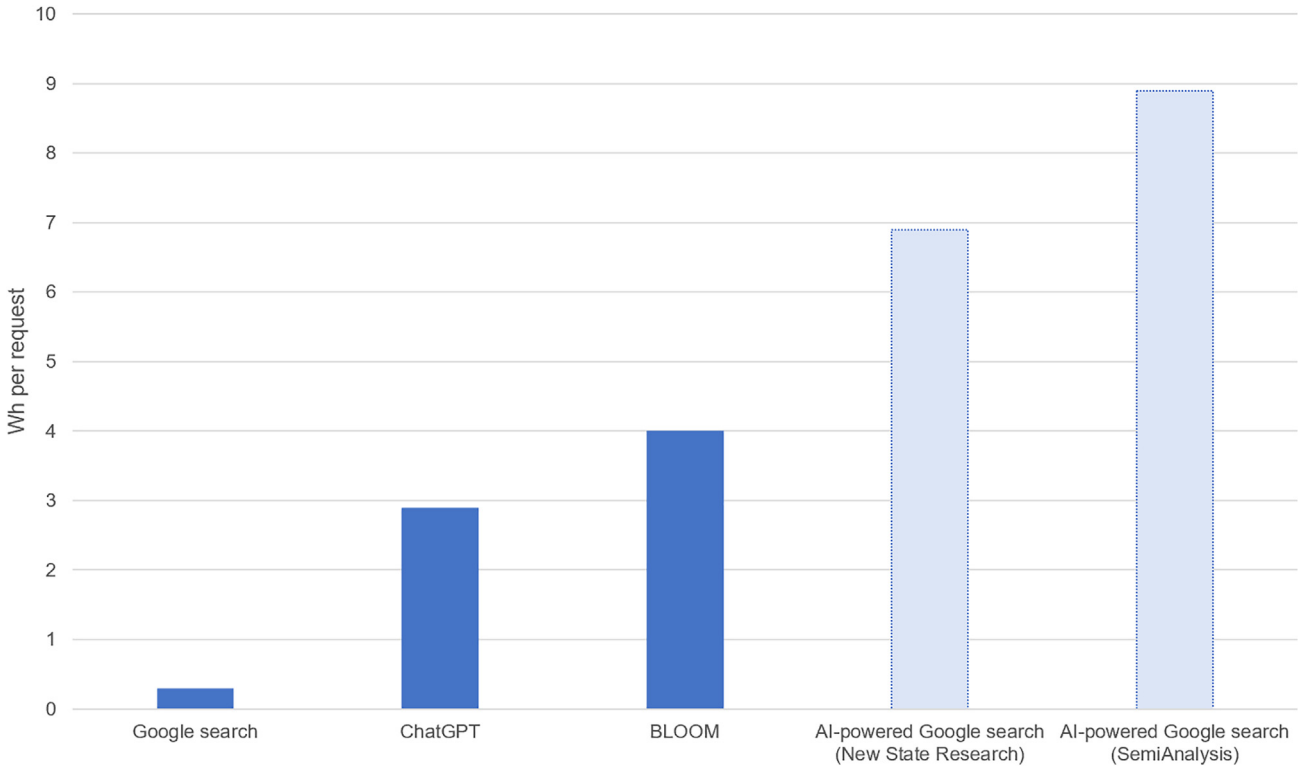

Figure 1. Estimated energy consumption per request for various AI-powered systems compared to a standard Google search

chain of AI servers is likely to remain a bottleneck for several more years. NVIDIA's chip supplier, TSMC, is also struggling to expand its chip-on-wafer-on-substrate (CoWoS) packaging technology, which is essential for the delivery of chips that NVIDIA requires. TSMC is investing in a new CoWoS packaging plant, but this plant is not expected to begin volume production until 2027.[11] By this year, NVIDIA could be shipping 1.5 million of its AI server units, even as its market share is projected to decline.[10] Given a similar electricity consumption profile, these machines could have a combined power demand of 9.75–15.3 GW. Annually, this quantity of servers could consume 85.4–134.0 TWh of electricity. At this stage, these servers could represent a significant contribution to worldwide data center electricity consumption. With NVIDIA strongly outperforming analyst expectations during the first half of 2023,[7] the AI server supply chain is on track to deliver the projected growth. An important caveat to be considered in this scenario is that the utilization rates will likely be less than 100%, which will mitigate part of their potential electricity consumption. Another factor that should be considered is overhead electricity consumption (e.g., server cooling), which increases the total electricity consumption related to the AI server units.

In addition to hardware efficiency improvements, innovations in model architectures and algorithms could help to mitigate or even reduce AI-related electricity consumption in the long term. Google's Generalist Language Model (GLaM) was trained on 7 times the number of parameters included in GPT-3, but it required 2.8 times less energy for this process only 18 months after GPT-3 was trained.[2] However, this perspective overlooks Jevons' Paradox, which was formulated in 1865 and occurs when increasing efficiency results in increased demand, leading to a net increase in resource use. This effect has long been observed in the history of technological change and automation,[12] with recent examples in AI applications.[13] In fact, the sudden surge in interest in generative AI during 2022 and 2023, during which demand began to outstrip supply, could be part of such a rebound effect. Moreover, the improvements in model efficiency now allow single consumer-level GPUs to train AI models. This implies that the growth in AI-related electricity consumption will originate not only from new high-performance GPUs such as NVIDIA's A100 and H100 GPUs but also from more generic GPUs. It is already the case that former cryptocurrency miners using such GPUs have started to repurpose their computing power for AI-related tasks. Many of these GPUs were left redundant in September of 2022, when Ethereum, the second-largest cryptocurrency, replaced its energy-intensive mining algorithm with a more sustainable

Please cite this article in press as: de Vries, The growing energy footprint of artificial intelligence, Joule (2023), https://doi.org/10.1016/j.joule.2023.09.004

alternative. The change was estimated to have reduced Ethereum's total power demand by, at most, 9.21 GW.[14] This equates to 80.7 TWh of annual electricity consumption. It has been suggested that 20% of the GPUs formerly used by Ethereum miners could be repurposed for use in AI, in a trend referred to as "mining 2.0."[15] Based on a power demand of 9.21 GW, this equates to a maximum of 1.84 GW. In turn, this could result in a shift of 16.1 TWh of annual electricity consumption to AI, with more devices potentially following as ongoing improvements in model efficiency continue to broaden the range of hardware suitable for AI purposes.

Lastly, enhancing model efficiency can also affect the trade-off between model performance and electricity costs. AI model performance often reaches a tipping point where even minor accuracy improvements become excessively energy-intensive.[3] By increasing the efficiency of models and reducing their energy costs, efforts to further improve these models may become more viable, thereby negating some of the efficiency gains.

Conclusion

While the exact future of AI-related electricity consumption remains difficult to predict, the scenarios discussed in this commentary underscore the importance of tempering both overly optimistic and overly pessimistic expectations. Integrating AI into applications such as Google Search can significantly boost the electricity consumption of these applications. However, various resource factors are likely to restrain the growth of global AI-related electricity consumption in the near term. Simultaneously, it is probably too optimistic to expect that improvements in hardware and software efficiencies will fully offset any long-term changes in AI-related electricity consumption. These advancements can trigger a rebound effect whereby increasing efficiency leads to increased demand for AI, escalating rather than reducing total resource use. The AI enthusiasm of 2022 and 2023 could be part of such a rebound effect, and this enthusiasm has put the AI server supply chain on track to deliver a more significant contribution to worldwide data center electricity consumption in the coming years. Moreover, enhancing efficiency could also potentially unlock a significant inventory of older and unused GPUs, such as those previously employed in mining cryptocurrency Ethereum, to be repurposed for AI. Therefore, it would be advisable for developers not only to focus on optimizing AI, but also to critically consider the necessity of using AI in the first place, as it is unlikely that all applications will benefit from AI or that the benefits will always outweigh the costs. Information on resource use for cases where AI is being applied is limited, so regulators might consider introducing specific environmental disclosure requirements to enhance transparency across the AI supply chain, fostering a better understanding of the environmental costs of this emerging technological trend.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Heaven, W.D. (2023). Google just launched Bard, its answer to ChatGPT—and it wants you to make it better. MIT Technology Review. March 21, 2023. https://www.technologyreview.com/2023/03/21/1070111/google-bard-chatgpt-openai-microsoft-bing-search/.

2. Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D.R., Texier, M., and Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Computer *55*, 18–28. https://doi.org/10.1109/MC.2022.3148714.

3. Verdecchia, R., Sallou, J., and Cruz, L. (2023). A systematic review of Green AI. WIREs Data Min. &. Knowl. *13*, e1507. https://doi.org/10.1002/widm.1507.

4. Luccioni, A.S., Viguier, S., and Ligozat, A.-L. (2022). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.02001.

5. SemiAnalysis. (2023). The Inference Cost of Search Disruption – Large Language Model Cost Analysis. https://www.semianalysis.com/p/the-inference-cost-of-search-disruption.

6. Dastin, J., and Nellis, S. (2023). Focus: For tech giants, AI like Bing and Bard poses billion-dollar search problem. Reuters. February 22, 2023. https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/.

7. Mehta, C., Cherney, M.A., and Nellis, S. (2023). Nvidia adds jet fuel to AI optimism with record results, $25 billion buyback. Reuters. August 23, 2023. https://www.reuters.com/technology/nvidia-forecasts-third-quarter-revenue-above-wall-street-expectations-2023-08-23/.

8. Leswing, K. (2023). Meet the $10,000 Nvidia chip powering the race for A.I. CNBC. February 23, 2023. https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html.

9. Google (2009). Powering a Google search. https://googleblog.blogspot.com/2009/01/powering-google-search.html.

10. Bary, E. (2023). Nvidia is 'dominating' and could unlock $300 billion in AI revenue by 2027, analyst says. MarketWatch. July 24, 2023. https://www.marketwatch.com/story/nvidia-is-dominating-and-could-unlock-300-billion-in-ai-revenue-by-2027-analyst-says-915935c0.

11. Chen, M., and Chan, R. (2023). TSMC new CoWoS packaging plant to start volume production in mid-2027. DigiTimes. July 26, 2023. https://www.digitimes.com/news/a20230725PD214/cowos-tsmc.html.

12. Autor, D.H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. J. Econ. Perspect. *29*, 3–30. https://doi.org/10.1257/jep.29.3.3.

13. Syed, T. (2023). The Demand Elasticity Paradox: More than Meets the AI. Medium. July 24, 2023. https://ai.plainenglish.io/the-demand-elasticity-paradox-more-than-meets-the-ai-1e87e63a4cfa.

14. De Vries, A. (2023). Cryptocurrencies on the road to sustainability: Ethereum paving the way for Bitcoin. Patterns *4*, 100633. https://doi.org/10.1016/j.patter.2022.100633.

15. Moses, R. (2023). Mining 2.0 Trends as Defunct Crypto Mining Rigs Tap into the AI Boom. Cryptopolitan. July 2, 2023. https://www.cryptopolitan.com/mining-2-0-crypto-mining-rigs-tap-into-ai/.