Will AI workloads consume all the world's energy?

Published 14 June 2023

Christopher Gannatti, CFA
Global Head of Research

On big questions like this, almost nothing stays constant. When we consider a new technology:

- We cannot assume that rates of adoption or usage will remain constant—they may drop, they may even grow.
- We cannot assume that the technology supplying our energy needs will remain constant—there could be breakthroughs in efficiency or changes in the overall energy mix.
- We cannot assume that the efficiency of the specific technology being adopted will remain constant—we have seen numerous examples of areas where an initial version of something in technology or software faces subsequent improvements that may give it greater capabilities with lower energy usage.

We must also recognise that artificial intelligence (AI) itself could suggest improvements in energy efficiency for specific applications—like the heating and cooling of a building. Therefore, any analysis of energy usage and AI must recognise that the one constant will be change.

Environmental impact of select large language models (LLMs)

LLMs have been garnering the lion's share of attention amidst the current excitement around generative AI. It makes sense to consider the amount of carbon emissions generated by some of these systems. The Stanford AI Index Report, published in 2023, provided some data, noting that factors like the number of parameters in a model, the power usage effectiveness1 of a data centre, and the grid carbon intensity all matter.

Figures 1a and 1b provide more detail when considering carbon dioxide intensity across different LLMs:

- Multiple factors drive the result: in Figure 1a, it might be tempting to look at the 'Number of Parameters' and predict that a larger number of parameters always means more emissions. We see that Gopher, at 280 billion parameters, did not have the highest carbon dioxide equivalent emissions x PUE—that distinction went to GPT-3.

- GPT-3, while not the largest model by number of parameters, was used at a grid carbon intensity that was the highest of the models shown and it had the largest power consumption of the models shown. That is certainly a recipe for high emissions, but it is instructive in that it forces us to think about not only the number of parameters, but where the data centre that the training occurs on is located, what the grid intensity of the carbon emissions that the data centre sits in is, and what the power consumption of the model is.
- Figure 1b is designed to help take a relatively abstract activity—that of training an LLM—and relate the resulting carbon emissions to more concrete activities, like flying from New York to San Francisco on a flight. It helps provide a sense of scale to what it means to see that GPT-3 is associated with 502 tonnes of carbon dioxide equivalent emissions.

Figure 1a: Environmental impact of select machine learning models, 2022

Source: Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.

Historical performance is not an indication of future performance and any investments may go down in value.

Figure 1b: Carbon dioxide equivalent emissions (tonnes) by selected machine learning models and real life examples

Source: Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.

Historical performance is not an indication of future performance and any investments may go down in value.

Considering power consumption of an LLM

Those building different LLMs have many levers they can pull in order to influence different characteristics, like energy consumption. Google researchers proposed a family of language models named GLaM (Generalist Language Model), which uses a 'sparsely activated mixture of experts'. While a full discussion of how that type of approach works is beyond the scope of this piece, we note that the largest of the GLaM models has 1.2 trillion parameters. Knowing solely that data point, the assumption would be that this model would consume more energy than any of the models that we saw in Figures 1a or 1b2.

In reality, the GLaM model with 1.2 trillion parameters consumes only one-third of the energy required to train GPT-3 and requires only half of the computation flops for inference operations. A simple way to think of what is going on is that, while the total model has 1.2 trillion parameters, a given input token into the GLaM model is only activating a maximum of 95 billion parameters, that is, the entire model isn't active

across all the parameters. GPT-3, on the other hand, activated all 175 billion parameters on each input token3. It is notable that, even if measuring the performance of AI models occurs on many dimensions, by many measures the GLaM model is able to outperform GPT-3 as well4.

Conclusion

The bottom line is that model design matters, and if model designers want to denote ways to maintain performance but use less energy, they have many options.

1 Power usage effectiveness (PUE) is useful in evaluating the energy efficiency of data centres in a standard way. PUE = (total amount of energy used by a computer data centre facility) / (energy delivered to computer equipment). A higher PUE means that the data centre is less efficient.

2 Source: Du et al. “GLaM: Efficient Scaling of Language Models with Mixture-of-Experts.” ARXIV.org. 1 August 2022.

3 Source: Patterson, David; Gonzalez, Joseph; Hölzle, Urs; Le, Quoc Hung; Liang, Chen; Munguia, Lluis-Miquel; et al. (2022): The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.19139645.v4

4 Source: Du et al, 1 August 2022.

Related blogs

+ 4 takeaways from EmTech Digital's AI conference

+ The end of the SaaSacre and the rise of generative AI

+ Tap into the AI revolution with WisdomTree

Related products

+ WisdomTree Artificial Intelligence UCITS ETF - USD Acc (WTAI/INTL)

Important Risks Related to this Article

Important Information

Marketing communications issued in the European Economic Area ("EEA"): This document has been issued and approved by WisdomTree Ireland Limited, which is authorised and regulated by the Central Bank of Ireland.

Marketing communications issued in jurisdictions outside of the EEA: This document has been issued and approved by WisdomTree UK Limited, which is authorised and regulated by the United Kingdom Financial Conduct Authority.

WisdomTree Ireland Limited and WisdomTree UK Limited are each referred to as "WisdomTree" (as applicable). Our Conflicts of Interest Policy and Inventory are available on request.

For professional clients only. The information contained in this document is for your general information only and is neither an oer for sale nor a solicitation of an oer to buy securities or shares. This document should not be used as the basis for any investment decision. Investments may go up or down in value and you may lose some or all of the amount invested. Past performance is not necessarily a guide to future performance. Any decision to invest should be based on the information contained in the appropriate prospectus and aer seeking independent investment, tax and legal advice.

The application of regulations and tax laws can oen lead to a number of dierent interpretations. Any views or opinions expressed in this communication represent the views of WisdomTree and should not be construed as regulatory, tax or legal advice. WisdomTree makes no warranty or representation as to the accuracy of any of the views or opinions expressed in this communication. Any decision to invest should be based on the information contained in the appropriate prospectus and aer seeking independent investment, tax and legal advice.

This document is not, and under no circumstances is to be construed as, an advertisement or any other step in furtherance of a public oering of shares or securities in the United States or any province or territory thereof. Neither this document nor any copy hereof should be taken, transmitted or distributed (directly or indirectly) into the United States.

Although WisdomTree endeavours to ensure the accuracy of the content in this document, WisdomTree does not warrant or guarantee its accuracy or correctness. Where WisdomTree has expressed its own opinions related to product or market activity, these views may change. Neither WisdomTree, nor any ailiate, nor any of their respective oicers, directors, partners, or employees accepts any liability whatsoever for any direct or consequential loss arising from any use of this document or its contents.