THIRD WAY

MEMO *Published May 1, 2024 · 8 minute read*

How AI Uses Energy

Mike Sexton, *Senior Policy Advisor for Artificial Intelligence and Digital Technology*

When ChatGPT was unveiled to the public in 2022, energy usage was probably not on most people's list of top things to wonder about. But the AI era will use a lot of power. Already, the energy usage of AI is increasing exponentially and by some estimates will surpass the energy used by Bitcoin by 2027. [1] And Bitcoin already consumes more energy than The Netherlands over the course of a year. [2]

There is a role for policy to help avoid a shortage of clean power for AI, but this memo focuses principally on the ways in which AI uses energy and why so much is needed.

Generative AI's energy requirements consist of three parts: manufacturing and maintaining the equipment ("lifecycle"), training the AI model ("training"), and generating an output when prompted by a user ("inference"). This paper will focus on the latter two – training and inference – which now occur primarily in data centers. [3]

Training means exposing an AI model to data to improve its “understanding” in fields like language, biology, or economic statistics. Available data shows the energy usage for training AI models has increased rapidly for decades, beginning in the late 1950s with the “perceptron” artificial neuron, [4] eventually doubling every 3.4 months between 2012 and 2018. [5] Since the publication of open models like Mistral, Llama, and Gemma, that trend is likely not just accelerating but multiplying still further as independent developers train their own tailor-made AIs.

Inference is how most of us use AI: querying it and receiving a result (text, image, etc.). The energy cost of a single inference is minuscule, especially compared to the upfront cost of training the AI model. However, the cost is non-negligible and accumulates quickly. Every year, we perform trillions of Google searches [6] if these searches were all performed with a chatbot like ChatGPT, they would use as much energy as Ireland. [7]

Training an AI model

Training an AI model involves feeding it data to “learn” about its domain of expertise. To better understand this process, consider three high-profile examples of AI training:

- GPT in ChatGPT stands for “Generative **Pre-trained** Transformer,” which tells you how important the training process is. While OpenAI reveals few details about GPT’s training, its data is known to include books, Wikipedia pages, and public websites. [8] There is also extensive post-training testing and fine-tuning to ensure, for example, that ChatGPT does not output training data verbatim – a problem that has led to copyright infringement lawsuits. [9]

- AlphaZero, an AI developed by Google DeepMind to play the board games of chess, Go, and shogi (Japanese chess), was trained strictly by playing the board games against itself on supercomputers. It took nine hours of training to become proficient in chess, 12 hours for shogi, and 13 days for Go. [10] Because it was not trained on human-played games, AlphaZero is renowned for not just playing better than humans, but creatively, deploying strategies humans have never imagined.

- AlphaFold, also by Google DeepMind, [11] was developed to predict the structures of protein and has revolutionized health care in the process. This process, when done experimentally, used to take months or even years. It was first trained for a week on 170,000 structures in the Protein Data Bank, [12] then trained *again* on 350,000 of its own predictions – a process known as self-distillation. AlphaFold has since predicted over 200 *million* protein structures in three years.

The more complex an AI model is, the more energy is needed to train it. In 2018, four years before ChatGPT opened to the public, OpenAI showed that the energy required to train frontier AI models was decupling (10x) each year. [13] This trendline is even more nosebleed-inducing when compared with

Moore's Law, which has accurately projected that the computing power of an integrated circuit merely *doubles* every *two* years.

However, it is not clear whether this trendline will continue indefinitely. Hyperscalers and chip designers like Nvidia [14] are continuously redesigning chips and data centers to improve energy efficiency at impressive rates. Google, for example, leverages its proprietary tensor processing units (TPUs), [15] which are designed specifically for AI, unlike *graphics* processing units (GPUs), which are far more commonly used. With these optimizations, Google was able to train an AI model seven times more powerful than GPT-3 [16] with a third of the energy. [17]

With anxiety growing that AI developers may soon run out of internet content to train their AI on, [18] there's little basis to assume frontier AI training will necessarily grow exponentially more costly for much longer. In five years, AI labs will doubtless still be perfecting their frontier AI models, but they probably won't be driven by exponentially increasing training anymore.

That, however, is only half the equation.

Inference from an AI

The other side of AI energy use is inference: when an AI model processes an input and *infers* the proper output, informed by its training. [19] Even if you have never heard of it, you are familiar with the results:

- YouTube transcribing spoken text from a video into captions
- Outlook flagging incoming email as spam
- Google providing and ranking search results
- DALL-E or Midjourney making an image from a prompt
- Facebook or Instagram ranking your friends' posts in your feed
- Alexa answering questions and fulfilling requests
- iPhone autocorrecting common typos
- Cameras detecting a human face

When you learn what AI inference is and how it is hidden in nearly every facet of our digital lives, you may notice that its energy cost is even *more* hidden. Of the AI inference examples above, all but the last two are performed on cloud servers rather than our own devices. That is why autocorrect still works without internet, but Alexa does not; it is why playing Xbox contributes to one's electric bill but generating images with DALL-E or Midjourney doesn't. When we rely on AI inference, we usually do not pay for it – the electricity comes from server farms, sometimes thousands of miles away.

Just as with training, the energy needed for inference can be thought of as a proxy for its complexity. Detecting a human face is easy but recognizing it is harder. [20] For a computer, processing text is trivial; reading words in an image or hearing them spoken is hard; and "understanding" words and their relationships to each other is very hard.

The energy challenge posed by inference today is that AI is performing *extremely* hard tasks. According to researchers, [21] generative tasks are orders of magnitude more energy intensive than previously common AI applications like text classification (e.g., spam filtering). A Stable Diffusion image generator [22] requires a whole smartphone charge of energy to produce one image; 1,000 images create as much CO2 as driving 4.1 miles. We can only imagine how much energy is needed to power OpenAI's new *video* generator Sora. [23]

In most cases, we can probably say inference is worth the cost. One text generation inference –one chatbot answer – only needs enough energy to power an LED lightbulb for about 17 seconds. [24] AlphaFold used far less energy and *vastly* less labor to infer 200 million protein structures from 2021-2024 than it took to determine 170,000 structures experimentally from 1976-2021.

However, AI appears positioned to outstrip the energy use of crypto soon. One study projected AI to use 0.5% of earth's electricity by 2027 [25] compared to 0.4% currently used for crypto mining, [26] and we can expect this growth will be driven primarily by increased inference. Often, as with AlphaFold, that inference can make the world and our lives vastly better, but we should be mindful that some frivolous or excessive AI usage could be a waste of resources.

What's next?

A paradox of cloud computing is that we don't notice how high its energy cost may be, but we also don't notice how *efficient* it can be.

Google's Tensor Processing Unit (TPU) is a great example of how hyperscalers – the largest cloud providers in the world – innovate to improve the efficiency of AI training and inference. [27] Non-Google developers can rent TPUs, but this specialized hardware cannot perform non-AI tasks well compared to more versatile graphics processing units (GPUs), and so *only* makes sense in a large, optimized data center with a constant stream of AI workloads where its efficiency gains can be maximized.

Lightweight AI is also expected to become more popular and diverse. The breakthrough of ChatGPT-3 was its size: 175 billion parameters, over 100x its predecessor (GPT-2). [28] While many innovators will strive to build bigger, better AI models, many will also build *smaller,* better models, like Google Gemma, available with either 2 billion or 7 billion parameters – small enough to run on a laptop. These models use much less energy and are ideal for narrower applications than larger, more powerful frontier models.

Nonetheless, it is an inescapable fact that significantly more energy will be needed to support the growth of AI. “The electrical grid is the largest and most complex machine humankind has ever built,” Third Way’s Shane Londagin writes in a recent paper. “But it is not prepared for the challenges of the future.”

ENDNOTES

1. AI is projected to use 85-134 Twh globally by 2027, compared to 112.31 Twh used now by Bitcoin.

   Li, Yunqi. “AI vs Bitcoin Mining: Which Consumes More Energy?” *WIRED Middle East*, 21 Mar. 2024, https://wired.me/science/energy/ai-vs-bitcoin-mining-energy/. Accessed 25 March 2024.

2. Bitcoin Energy Consumption Index.” *Digiconomist*, https://digiconomist.net/bitcoin-energy-consumption/. Accessed 30 Apr. 2024.

3. Inference can also take place on-device, a practice known as “edge computing,” which is expected to become more common as it can offer lower latency and greater privacy.

4. Roberts, Eric. “Neural Networks - The Perceptron.” *Stanford University*, 2000, https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html. Accessed 12 March 2024.

5. Hao, Karen. “The Computing Power Needed to Train AI Is Now Rising Seven Times Faster than Ever Before.” *MIT Technology Review*, 11 Nov. 2019, https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/. Accessed 4 March 2024.

6. Robbins, Ron. “How Many Google Searches Per Day Are There?” *Clicta Digital Agency*, 26 Oct. 2023, https://clictadigital.com/how-many-google-searches-per-day-are-there/. Accessed 26 March 2024.

7. Calma, Justine. “The Environmental Impact of the AI Revolution Is Starting to Come into Focus.” *The Verge*, 10 Oct. 2023, https://www.theverge.com/2023/10/10/23911059/ai-climate-impact-google-openai-chatgpt-energy. Accessed 26 March 2024.

8. Johri, Shreya. “The Making of ChatGPT: From Data to Dialogue.” *Science in the News*, 6 June 2023, https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/. Accessed 6 March 2024.

9. Grynbaum, Michael M., and Ryan Mac. “The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work.” *The New York Times*, 27 Dec. 2023. *NYTimes.com*, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. Accessed 6 March 2024.

10. David Silver, et al. “AlphaZero: Shedding New Light on Chess, Shogi, and Go.” *Google DeepMind*, 6 Dec. 2018, https://deepmind.google/discover/blog/alphazero-shedding-new-light-on-chess-shogi-and-go/. Accessed 13 March 2024.

11. “Putting the Power of AlphaFold into the World’s Hands.” *Google DeepMind*, 22 July 2022, https://deepmind.google/discover/blog/putting-the-power-of-alphafold-into-the-worlds-hands/. Accessed 13 March 2024.

12. Jumper, John, et al. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature*, vol. 596, no. 7873, Aug. 2021, pp. 583–89. *www.nature.com*, https://doi.org/10.1038/s41586-021-03819-2. Accessed 13 March 2024.

13. “AI and Compute.” *OpenAI*, 16 May 2018, https://openai.com/research/ai-and-compute#addendum. Accessed 7 March 2024.

14. Narasimhan, Shar. “How Energy-Efficient Computing for AI Is Transforming Industries.” *NVIDIA Blog*, 7 Feb. 2024, https://blogs.nvidia.com/blog/energy-efficient-ai-industries/. Accessed 17 April 2024.

15. “Tensor Processing Units (TPUs).” *Google Cloud*, https://cloud.google.com/tpu. Accessed 12 Mar. 2024.

16. GLaM with 1.162 trillion parameters vs. GPT-3 with 17b5 billion parameters

17. 1287 megawatt hours for GPT-3 vs. 456 megawatt hours for GLaM

18. Seetharaman, Deepa. “Why OpenAI and Other Data-Hungry AI Companies Need a Bigger Internet - WSJ.” *The Wall Street Journal*, 1 Apr. 2024, https://www.wsj.com/tech/ai/ai-training-data-synthetic-openai-anthropic-9230f8d8. Accessed 17 April 2024.

19. “What Is AI Inferencing?” IBM Research Blog, 9 Feb. 2021, https://research.ibm.com/blog/AI-inference-explained. Accessed 11 March 2024.

20. “Luxand.Cloud - Lightning Fast, Accurate, and Stable Face Recognition API.” Accessed April 3, 2024. https://luxand.cloud/face-recognition-blog/luxand.cloud. Accessed 3 April 2024.

21. Luccioni, Alexandra Sasha, et al. *Power Hungry Processing: Watts Driving the Cost of AI Deployment?* arXiv:2311.16863, arXiv, 28 Nov. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2311.16863. Accessed 5 March 2024.

22. https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

23. “Video Generation Models as World Simulators.” *OpenAI*, 15 Feb. 2024, https://openai.com/research/video-generation-models-as-world-simulators. Accessed 11 March 2024.

24. From Luccioni et al, 1,000 text generation queries used on average 0.047 kWh or 47 watt-hours. This averages out to 0.047 watt-hours per inference, enough to power a 10 watt LED for 0.0047 hours or 17 seconds.

25. Erdenesanaa, Delger. “A.I. Could Soon Need as Much Electricity as an Entire Country.” *The New York Times*, 10 Oct. 2023. *NYTimes.com*, https://www.nytimes.com/2023/10/10/climate/ai-could-soon-need-as-much-electricity-as-an-entire-country.html. Accessed 15 Marc 2024.

26. “Data Centres & Networks.” *IEA*, https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks. Accessed 5 Mar. 2024.

27. “Tensor Processing Units (TPUs).” *Google Cloud*, https://cloud.google.com/tpu. Accessed 15 Mar. 2024.

28. Alarcon, Nefi. “OpenAI Presents GPT-3, a 175 Billion Parameters Language Model.” *NVIDIA Technical Blog*, 7 July 2020, https://developer.nvidia.com/blog/openai-presents-gpt-3-a-175-billion-parameters-language-model/. Accessed 18 March 2024.