AI's Environmental Armageddon: The Countdown Has Begun

Pete Bernard – EDGECELSIOR

KEY THEME – GAIN COMPETITIVE ADVANTAGE WITH EDGE COMPUTING AND A SUSTAINABLE ENERGY STRATEGY

Artificial Intelligence has reached a new level of public consciousness (and perhaps its own?) with the advent of easily accessible tools, portals and applications and breathless forecasts of utopia and doom that are washing over us like high tide on a Cape Cod Beach. The C-suite of the Fortune 500 are rapidly conducting workshops and offsites, startups are quickly editing their VC pitch decks and we're all becoming a bit more educated on what AI could do, would do, and should do.

AI is the latest in intense compute workloads that have been made possible by the combination of innovative silicon, next generation networks and virtualized and elastic computing. AI has been around in many forms for many years, but it wasn't until high performance cloud computing reached a tipping point that made it feasible to train AI models on millions and billions of data points and execute those models with accelerated silicon to the point that it was a useful workload. Arguably, the "brute force" method of cleaning, labelling and training these massive models on useful data may not even be the best way to train them[1].

Like many other workloads and functions, we're also seeing a migration of these high performance AI workloads to edge devices that live outside the data center, much "closer to the action." We can even see LLMs and generative AI compute happening on semiconductors designed for cell phones[2]

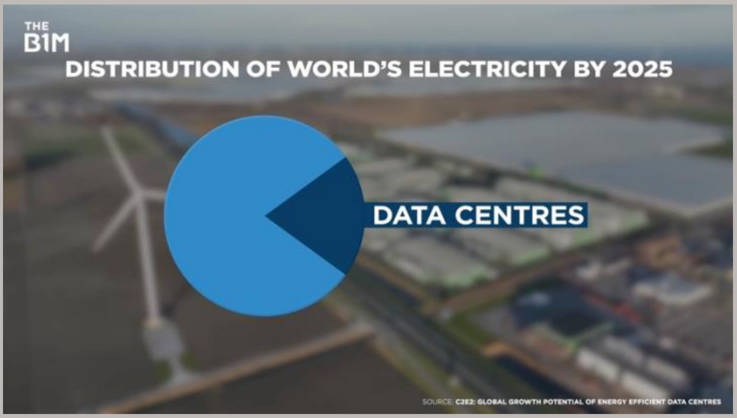


Figure 1 - By 2025, Data Centers will use 25% of the world's electricity.

Although some companies are dipping their toes into the cold ocean of AI pricing[3], many of us are experiencing AI "for free." We can generate images, generate stories, poems, get our

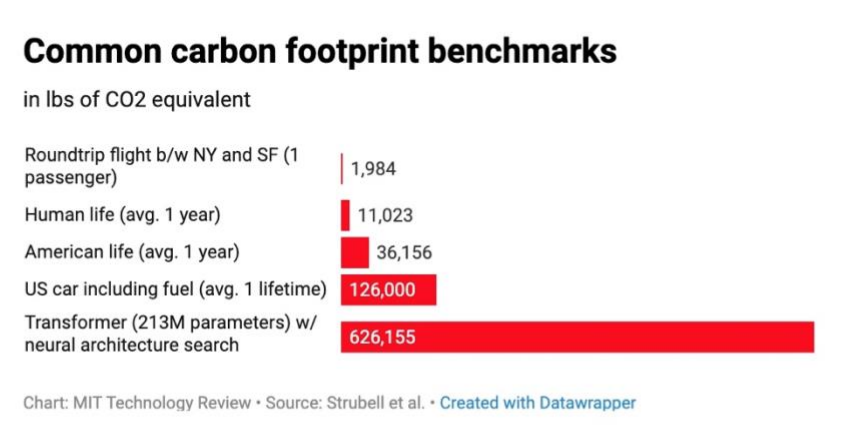
[1] The way we train AI is fundamentally flawed | MIT Technology Review

[2] Qualcomm Taps Meta for On-Device Generative AI (Phone Scoop)

[3] Microsoft 365 Copilot: Release date, features and pricing (androidauthority.com)

questions answered, no matter how arcane. We can also recognize faces in our bottomless collection of photos. It feels free – but is it?

Like other high performance compute workloads before it (e.g., crypto mining), these AI scenarios require a tremendous amount of computing power, and that computing power consumes massive amounts of electricity, water, and carbon impact. The “free” AI we experience is costing us environmental resources, and as we grapple with the ethical issues around AI, we also need to develop an approach to make high performance compute workloads, like AI, sustainable. Not only is this a moral imperative, it’s also a practical one – we will increasingly see the lack of power and other resources as a limiting factor for AI – which will drive up cost and limit deployments and applicability.


Common carbon footprint benchmarks

in lbs of CO2 equivalent

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

This awareness and work is already underway. “Way back” in 2018 the World Economic Forum[4] made a comprehensive set of proposal to address this challenge, including embedding environmental considerations into their design principles.

Let’s look at some examples so that you more easily wrap your head around these issues.

- Imagine the carbon impact of five cars’ emissions over their lifetime – 600,000 lbs. - that is roughly what it costs to train a single NLP (natural language processing model. [5]
- Training a single AI model can gobble up more electricity than 100 US homes use in an entire year[6]
- ChatGPT needs to 'drink' a 500ml bottle of water for a simple conversation of roughly 20-50 questions and answers, depending on when and where ChatGPT is deployed[7]
- A data center training GPT-3 is estimated to use 700,000 liters — or about 185,000 gallons — of fresh water, enough to fill a nuclear reactor's cooling tower or produce 370 BMW cars[8]

The scarcity of water resources is well documented – our groundwater is being consumed far beyond it’s capacity. Things are getting so desperate that Arizona is working on a plan to

[4] Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf (weforum.org)

[5] [1906.02243] Energy and Policy Considerations for Deep Learning in NLP (arxiv.org)

[6] How Much Energy Do AI and ChatGPT Use? No One Knows For Sure - Bloomberg

[7] How much water does ChatGPT ‘drink’ for every 20 questions it answers? (govtech.com)

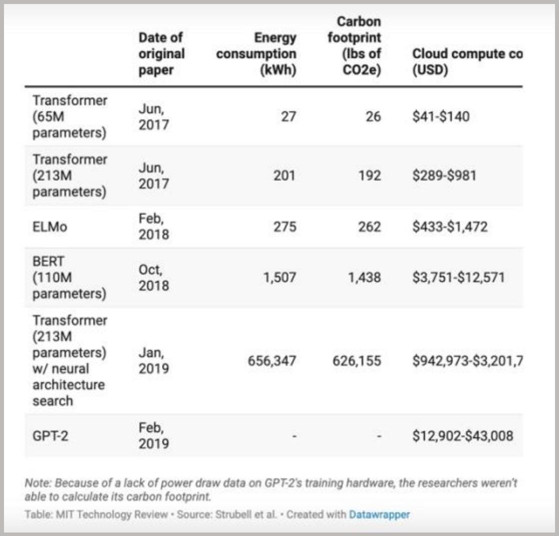
[8] ChatGPT Uses a Lot of Water to Train Itself, Study Shows (businessinsider.com)

desalinate water in the Gulf of Mexico and pump it 200 miles (uphill!) – into Maricopa County[9]. In the face of this many tech companies are innovating in recycling and other methods to cool the chips that are cranking through their AI workloads. However, it is puzzling why companies are focused on putting data centers in water-scarce geographies, such as Arizona. In 2019 alone, Google requested, or was granted, more than 2.3 billion gallons of water for data centers in three different states, according to public records posted online and legal filings. [10]

But what about the power consumption? The public power grids in the US are struggling to meet consumer demand while shifting to more renewable sources to minimize carbon impact. Most of the nation's transmission and distribution lines were constructed in the 1950s and 1960s, with a 50-year life expectancy, meaning they have reached or surpassed their intended lifespan. As consumer, we have an insatiable appetite for power – from cooling buildings (in increasingly warm weather), charging our devices, powering our cars and heating our homes. The shift to LED lighting a lower power devices including TVs have helped stem the tide, but we keep falling farther behind the grid's capabilities. Both consumers and business also need reliable power at the lowest price.

High performance computing is now becoming a key factor in this critical power infrastructure environment. Businesses cannot simply call up their local utility and ask for 40 megawatts for their data center. The capacity is not there, the newer transmissions lines are not there and there is a long waiting list to get that type of allocation.

The upper bound limit on AI might not end up being semiconductor capability and supply, or software sophistication, or even ethical issues. The upper bound on AI may simply be…electricity.



	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute co (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,7
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

It sounds pretty bleak, but there are several areas being worked on to mitigate the challenge of high-performance computing and scarce environmental resources:

- There is a tremendous amount of capital and government sponsorship to upgrade energy grids, including transmission lines, renewable resources and more sophisticated carbon capture capabilities, including areas like VPPs, or Virtual Power Plants, that can use multiple distributed resources, and software orchestration, to supply power on demand to a larger grid. This effort will take years and faces not only technology hurdles but regulatory and community (NIMBY) hurdles along the way.

[9] Arizona's Pipe Dream - https://pca.st/episode/10935526-1c34-452a-b540-6db608b3c006

[10] Secret Cost of Google's Data Centers: Billions of Gallons of Water | Time

- Alternatives energy sources like hydrogen and nuclear fusion[11] continue to evolve. Hydrogen is having a moment but it still faces serious cost issues, and until those get in line it won't be practical to power data centers, even edge ones, with Hydrogen power plants. Geothermal[12] is also getting "hot" as a potential limitless source of energy to spin turbines – although it lags behind other renewables in funding and awareness.
- Innovation in semiconductors, including new fab processes, are enabling much lower power consumption for traditional high-performance workloads. In addition to the amazing capabilities of the latest NVIDIA and AMD chips in the high end of the market, we're also seeing companies like Qualcomm[13], Alif[14], and Arm supplying capabilities to leverage new 5nm and below process to execute AI inferences right on the edge and this will take the load off of hyper-scalar data centers and backhaul networks. Organizations like TinyML are innovating with tiny AI models for anomaly detection and even AI vision running on MCU-class silicon.
- Green Code[15] is an effort to inject intelligence into code that understands power requirements such that is executes at the right time based on grid capabilities. For example, some workloads may run during the evening when the grid is less challenged.
- We can't improve things we can't measure, and we're starting to see new capabilities in measure power such as the Advanced Metering Infrastructure standard that will help drive visibility of energy usage on deployments and data centers. Ultimately, we need transparency in water usage, power usage and how that applies to cloud to edge solutions. Although we now typically measure bandwidth consumed in a given solution, we do not yet typically measure total power.

The intersection of powerful AI workloads, edge computing, semiconductor innovation, and innovation in power and environmental resources is a critical space to pay attention to, and every company that has an edge strategy, a data strategy, and a security strategy needs to add one more – an energy strategy.

As environmental resources continue to remain scarce and our appetite for workloads increases, our industry will need to get much smarter and transparent on how we get the right amount of power and resources to the workloads that we need – and this awareness will drive even more innovation opportunities in low power, low TCO and earth-friendly methods of running our businesses.

[11] Major breakthrough on nuclear fusion energy - BBC News

[12] Tools Born From Fracking Fuel Geothermal Rush - The New York Times (nytimes.com)

[13] Qualcomm power-efficient AI: Making technology more sustainable [video]

[14] Alif Semiconductor & Onsemi Build An Extremely Efficient AI Camera Design Using AI/ML | Alif Semiconductor

[15] Why Green Coding is a Powerful Catalyst for Sustainability Initiatives - IBM Blog