Artificial intelligence (AI) is expensive. AI is costly for companies and the environment, but this must change if we want to leverage its full potential, using it to fix problems and not create them.

If AI is so expensive, how can ChatGPT be free? Because it's not spending your money. OpenAI, the company behind ChatGPT, lost 5 billion dollars this year (Field, 2024) kindly funded by Microsoft, SoftBank, Fidelity, Thrive Capital and other companies who invested in its various rounds of funding. Costs for companies involved in AI are split into two main areas: hardware and energy.

Powerful hardware is needed to store the massive amounts of data and do the billions of calculations required to train and run powerful AI models. The scale of the cost of this hardware can be demonstrated in a number of ways. Unless you've been living under a rock, you will undoubtedly have heard of the explosion of Nvidia stock by almost 300% over the past year, resulting in it claiming the title of world's most valuable company surpassing 3.5 trillion dollars in valuation. Nvidia produce Graphics Processing Units (GPU) which are a type of chip designed specifically to perform massive numbers of parallel calculations, making them perfect for training and running AI. As the demand for AI grew, so too did the demand for GPU's. Meta themselves aim to own 350,000 Nvidia H100 GPU's (Engineering at Meta, 2024) to add to their already impressive infrastructure. With each H100 reportedly costing around £40,000 (Shilov, 2024), you can do the maths.

To run powerful hardware, you need exactly that: power. In 2022, datacentres used up to 460 terawatt hours of electricity and this figure is expected to rise to 1,000 terawatt hours per year as early as 2026. This is equivalent to the electricity consumption of the entirety of Japan according to the International Energy Agency (Baraniuk, 2024).

The energy demands of AI are further evidenced by recent moves from Google to purchase several Kairos Power modular nuclear reactors with a combined energy output of 500 megawatts (Gaulkin, 2024), equivalent to 50% of a traditional reactor such as Chernobyl. According to a BBC article, Amazon "said it would buy a nuclear-powered data centre in the state of Pennsylvania." (Silva, 2024)

Beyond the monetary value of hardware and energy, AI poses a significant cost in terms of environmental impact. Only a select few companies have started investing in nuclear power due to the high initial cost combined with the sometimes decades it takes for nuclear plants to be fully operational and integrated into existing infrastructure.

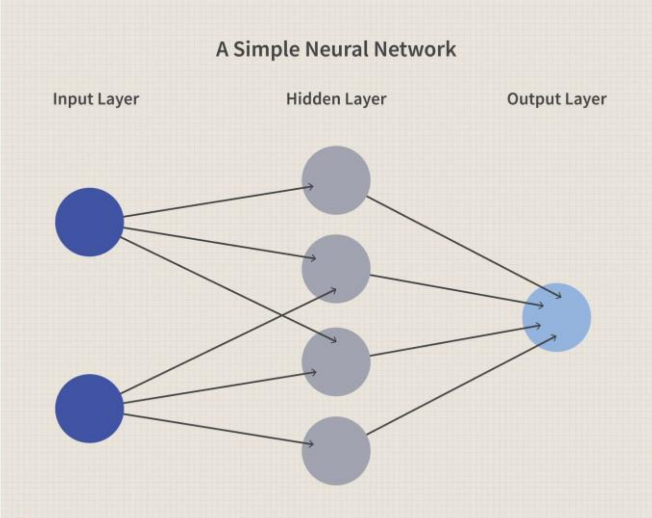
Different models have had different environmental impacts. Hugging Face BLOOM was trained on "a French supercomputer powered mainly by nuclear energy" (Simmons, 2023) with Hugging Face themselves saying training produced around 25 tonnes of carbon dioxide emissions – around 25 flights from London to New York.

As it was trained using modern technology powered by nuclear energy, BLOOM is one of the cleanest AI models (partly proven by its willingness to publicise these numbers) whereas models such as OpenAI's GPT4 may have produced emissions of up to 500 tonnes due to being "trained on older, more energy-intensive hardware based in another country with fewer low-carbon energy sources." (Simmons, 2023). It is important to note however that this is a Hugging Face estimate – itself a rival to OpenAI.

Optimising AI in both its training and operations is key to reducing costs which will reduce energy consumption and therefore harmful gas emissions from dirty energy sources. One way we can do this is by making a certain type of calculation key to neural networks (a popular branch of AI) as efficient as possible.

Neural networks consist of layers of “neurons” or “nodes” connected via different pathways. Every neural network has a set of input neurons and a set of output neurons, with a varying number of hidden neuron layers between the two. Each neuron in the previous layer is connected to each neuron in the next layer via pathways which apply weightings. The value of each non-input layer neuron is found by summing the values of all paths leading to it.

For example, an input neuron might have a value of 1. A signal passes down a pathway with a weighting of 0.5 from this input neuron to a neuron in the hidden layer, resulting in the value being multiplied by the weighting.



A diagram of a simple neural network containing one hidden layer of neurons. Sabrina Jiang © Investopedia 2020

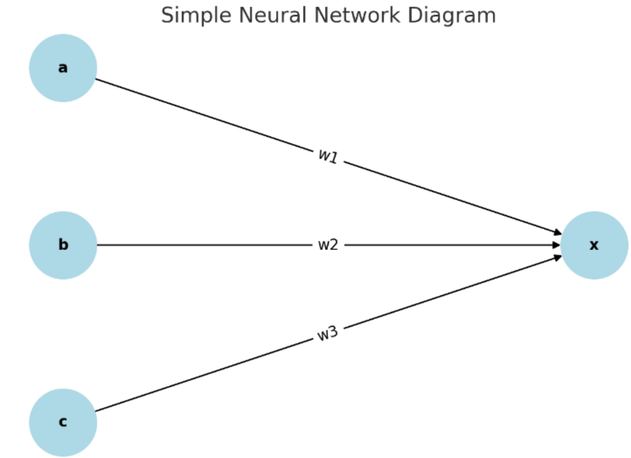
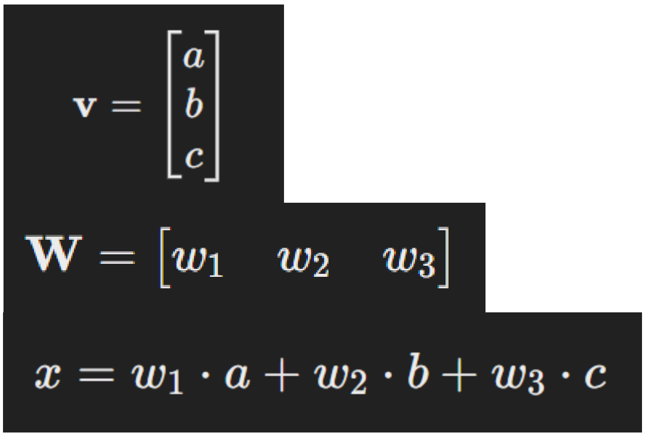


Diagram generated using ChatGPT

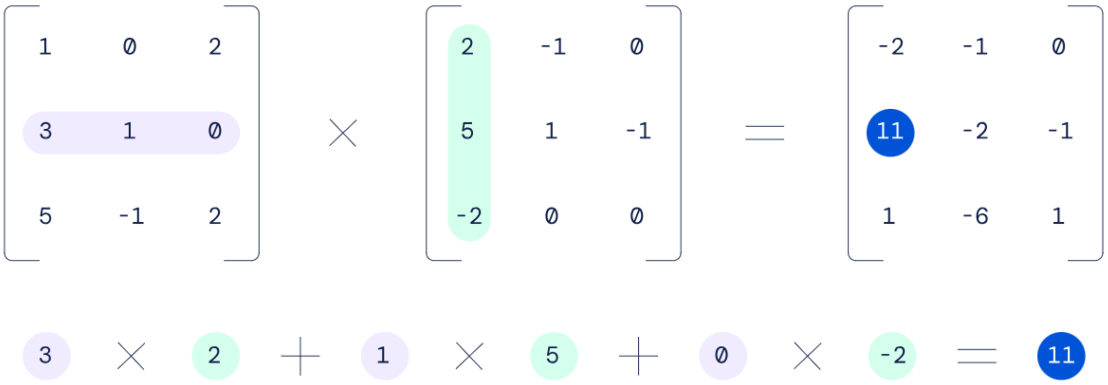


$$\mathbf{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}$$

$$x = w_1 \cdot a + w_2 \cdot b + w_3 \cdot c$$

We can treat the values of each layer of the neural network as a column in a matrix. We can also put the weightings leading to a particular node into a row of a matrix. The value of that node is then found by multiplying these two matrices together. In the neural network above, the values can be represented by the 3x1 matrix “v” and the weightings by the 1x3 matrix “w”. Matrix multiplication can then be used to find the value of x. This simple model can be enlarged to accommodate any number of nodes and weightings, and the same principle of matrix multiplication applies. See below a visual representation (Google DeepMind, 2022).



$$\begin{bmatrix} 1 & 0 & 2 \\ 3 & 1 & 0 \\ 5 & -1 & 2 \end{bmatrix} \times \begin{bmatrix} 2 & -1 & 0 \\ 5 & 1 & -1 \\ -2 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -2 & -1 & 0 \\ 11 & -2 & -1 \\ 1 & -6 & 1 \end{bmatrix}$$

$$3 \times 2 + 1 \times 5 + 0 \times -2 = 11$$

As seen in the previous diagrams, in an extremely simple network with 3 input nodes, 1 output node, and no hidden layers, 5 calculations must be made, and this number gets exponentially larger with networks of growing complexity.

So, surely there must be a way to speed up these matrix multiplications? There is, but it comes with a catch. The German mathematician and computer scientist Volker Strassen published the Strassen algorithm in 1969 proving that there was a more optimised method to multiply matrices than brute forcing them. The Strassen algorithm is a type of divide and conquer algorithm which divides matrices into smaller matrices. When multiplying these smaller matrices, some terms cancel out, resulting in fewer overall multiplications.

But Strassen’s algorithm is also a galactic algorithm – “an algorithm with record-breaking theoretical … performance, but which isn't used due to practical constraints” (Wikipedia Contributors, 2025). Strassen’s algorithm is considered to be a galactic algorithm as the performance increase is only meaningful for figuratively “galactic” matrices – ones which are much larger than are used in neural networks today.

The battle to optimising matrix multiplications didn’t stop there, however. In 2022, Google DeepMind published a paper introducing AlphaTensor, “the first artificial intelligence … system for discovering novel, efficient, and provably correct algorithms” (Google DeepMind, 2022).

Trained via reinforcement learning, the model tries to find the most optimised (aka using the least amount of multiplication steps) way to multiply two matrices using Tensor Decomposition. The example from the blog states: “if the traditional algorithm taught in school multiplies a 4x5 by 5x5 matrix using 100 multiplications, and this number was reduced to 80 with human ingenuity, AlphaTensor has found algorithms that do the same operation using just 76 multiplications.” Interestingly, different algorithms perform differently on different hardware, and AlphaTensor can help with this, finding optimised algorithms for target hardware resulting in “10-20%” performance gains.

AI has become a bit of a buzzword in recent years, so why not combine two buzzwords to make quantum AI. But can quantum computers really help to improve AI? The short answer – nobody knows. There is a famous quote attributed to both Niels Bohr and Richard Feynman that says “if you think you understand quantum mechanics, you don’t understand quantum mechanics” and this applies to quantum computing as well. Every technique in the world of artificial intelligence was created and optimised with classical, not quantum, computing in mind. Asking “how can quantum computers help AI” is like asking “how can the invention of this new racket sport help goalies in football?”. Quantum computers aren’t simply “better computers”, in fact they are significantly worse, even useless, at 99% of tasks. I believe that a more realistic question to ask once quantum computers have become powerful enough would be “what is the goal of AI, and how can we use quantum computers to achieve this goal”. All of the tools which come under the umbrella of AI that were designed and developed within the lens of classical computing cannot simply be ported over and modified slightly but will have to be completely reinvented and reimagined to work with quantum mechanics principles in mind.

So, if physics can’t save us, can biology?

Artificial Intelligence mimics biological intelligence and essentially does what humans’ brains do. During training, they take input, create output and learn based on a response or grading of that output. Babies do the same. They receive an input seeing, hearing, smelling, tasting and touching and produce an output - usually crying. If they knock over a glass, they’ll receive negative feedback in the form of a loud crash and an angry parent, and (hopefully) are less likely to do it again. Babies even watch their peers and parents to see how they respond in unfamiliar situations and copy their behaviour – if you’ve ever seen a child in an unfamiliar situation, it will almost certainly be looking at its parents to see how they are reacting, and then they react accordingly.

So, if our brains and AI function in the same way, how come ChatGPT isn’t ruling the world yet like us and still fails many basic tasks any human could do in an instant. Simply, humans have been at the training stage for longer. Llama 3 70b – meaning 70 billion parameters - required a total of 6.4 million H100 GPU-hours to train (Eassa et Eryilmaz, 2024). This means that if you had one H100 GPU, it would take 6.4 million hours to train that Llama model. Luckily, as we learned earlier, Meta has quite a few and AI can be trained concurrently on thousands of GPUs at a time. For the larger Llama 3 405b model, total training hour estimates are wide ranging – up to 40 million hours or 456 years!

Humans laugh in the face of 456 years. Early hominins – our oldest relatives and "having most recently shared a common ancestor with chimpanzees” – lived between 7 and 4.4 million years ago and a breakthrough discovery in 2017 found the earliest homo sapiens remains, along with stone tools, in Morocco, 315 thousand years ago (Callaway, 2017). Human brains have been training for a long time and won’t stop any time soon.

More importantly to the matter of optimisation and efficiency, the human brain is much more efficient than traditional silicon. Skimming the surface of a complex but deeply interesting study, we find that according to Balasubramanian, V in his 2021 paper “Brain Power”, the human brain has “an energy budget of ~20 W” whereas modern consumer desktop computers can range from 500 to over 1000 watts and a single H100 GPU will consume 700w.

So, why not harness the power and efficiency of mother nature. A group of scientists have been trying to do exactly that, and a whole new area of science – dubbed “biocomputing” – was developed when in 2013 Stanford bioengineers “announced they had created the biological equivalent of a transistor” (Wikipedia Contributors, 2024), completing the final piece of the three-piece puzzle of computing: data storage, information transmission, and logic. Finalspark are one of the many companies attempting to make biocomputing possible. They use live human neurons obtained from stem cells to “store data and perform logic operations using neurons as circuits”.

There are also cases where AI is much better than humans such as simple maths and just sheer memory capacity. But this is because in this respect, AI has better hardware than. We return again to Meta’s 350 thousand Nvidia H100s, each with more than 80gb of memory. That’s around 28 million gigabytes of memory. Whereas a common estimate for the memory of a human brain is only around 10^{14} bits or 125 thousand gigabytes (Trazzi and Yampolskiy, 2020). This is of course an approximate estimate of the computer storage equivalent of the brain as the inner workings of brain synapses are not fully known and like converting between quantum and classical systems, one doesn’t always map well onto the other.

We’ve learned how to optimise AI, but what about the reason why? At the moment, working at the forefront of AI is monopolized by literally the world’s biggest companies. Some argue that this is the same with many areas, but we must put the cost of AI into perspective. With a £1000 laptop – maybe less – you could develop cutting edge software and all you need to be an accomplished artist is paper and a pen. But if you want to make any meaningful contributions to the world of AI, I hope you have billions of dollars to burn. Take the recently announced project Stargate, where a US government backed company plans to invest 500 billion dollars in AI and related infrastructure. This investment is greater than the GDP 156 countries, and if it were a country, would rank as the 34th largest economy in the world. As with many modern technologies, rich nations have access, but poor nations are left behind. When dealing with technology as potentially transformative as artificial intelligence, we must not ignore this divide. Ensuring equitable access to AI by reducing costs via increased efficiency and optimisation, as well as ensuring that while it shapes our future, AI does not jeopardize our present through environmental consequences should be a goal held by all.

Bibliography

Field, H. (2024). *OpenAI sees roughly $5 billion loss this year on $3.7 billion in revenue*. [online] CNBC. Available at: https://www.cnbc.com/2024/09/27/openai-sees-5-billion-loss-this-year-on-3point7-billion-in-revenue.html.

Engineering at Meta. (2024). *Building Meta’s GenAI Infrastructure*. [online] Available at: https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/.

Shilov, A. (2024). *Nvidia's H100 AI GPUs cost up to four times more than AMD's competing MI300X — AMD's chips cost $10 to $15K apiece; Nvidia's H100 has peaked beyond $40,000: Report*. [online] Tom's Hardware. Available at: https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-h100-ai-gpus-cost-up-to-four-times-more-than-amds-competing-mi300x-amds-chips-cost-dollar10-to-dollar15k-apiece-nvidias-h100-has-peaked-beyond-dollar40000.

Baraniuk, C. (2024). *Electricity grids creak as AI demands soar*. [online] BBC News. Available at: https://www.bbc.co.uk/news/articles/cj5ll89dy2mo.

Gaulkin, T. (2024). *AI goes nuclear*. [online] Bulletin of the Atomic Scientists. Available at: https://thebulletin.org/2024/12/ai-goes-nuclear/.

Silva, J. da (2024). Google turns to nuclear to power AI data centres. *BBC News*. [online] 15 Oct. Available at: https://www.bbc.co.uk/news/articles/c748gn94k95o.

Simmons, L. (2023). *The Carbon Cost of AI - The Carbon Literacy Project*. [online] The Carbon Literacy Project. Available at: https://carbonliteracy.com/the-carbon-cost-of-ai/ [Accessed 26 Jan. 2025].

Wikipedia Contributors (2025). *Galactic algorithm*. Wikipedia.

Google DeepMind. (2022). *Discovering novel algorithms with AlphaTensor*. [online] Available at: https://deepmind.google/discover/blog/discovering-novel-algorithms-with-alphatensor/.

Eassa, A et Eryilmaz, S. (2024). *NVIDIA Sets New Generative AI Performance and Scale Records in MLPerf Training v4.0 | NVIDIA Technical Blog*. [online] Available at: https://developer.nvidia.com/blog/nvidia-sets-new-generative-ai-performance-and-scale-records-in-mlperf-training-v4-0/.

Callaway, E. (2017). Oldest Homo sapiens fossil claim rewrites our species' history. *Nature*. [online] doi:https://doi.org/10.1038/nature.2017.22114.

Wikipedia Contributors (2024). *Biological computing*. Wikipedia.

Balasubramanian, V. (2021). Brain power. *Proceedings of the National Academy of Sciences*, 118(32), p.e2107022118. doi:https://doi.org/10.1073/pnas.2107022118

Trazzi, M. and Yampolskiy, R.V. (2020). Artificial Stupidity: Data We Need to Make Machines Our Equals. *Patterns*, 1(2), p.100021. doi:https://doi.org/10.1016/j.patter.2020.100021.

Ewelina Kurtys (2023). *What is biocomputing? - FinalSpark*. [online] Final Spark. Available at: https://finalspark.com/what-is-biocomputing/.