Large Al Models Consume Huge Computational Power and Energy

Shivam Singh[1], Shorya Gupta[2], Divyanshu Panwar[3], Naman Agarwal[4]
[1,2,3,4]*Bennett University*

***Abstract*—Today AI is doing amazing things but often at a price that nobody really seems to notice. Behind the incredibly capable language models today is a massive demand on power and compute. This paper analyses this cost up to a point. It has all done to investigate the systems such as the ones powering models such as GPT 4, PaLM 2, LLaMA 2, and Gemini Ultra. We investigate the relationship between performance and resource scale, the impact on the environment of large training runs and the hardware required. We also consider what can be done specifically including building models that are slimmer and running data centres that are greener. One number in particular stand out. The training of a single top of the line model can consume as many carbon emissions as the lifetime of a handful of cars.**

I. INTRODUCTION

It wasn't long ago that the idea of a machine that could write essays, debug code or have a nuanced conversation seemed like science fiction. This is standard today for models like GPT 4, Gemini Ultra and LLaMA 2. But every great output is the product of a massive infrastructure running at full capacity — and that infrastructure isn't free.

The growth in AI computing has been almost unbelievable. Between 2012 and 2018, the compute needed to train the best models doubled every three and a half months a rate that dwarfs Moore's Law. For example, training GPT 3 took 3.14×10 23 floating point operations. Then the models that followed demanded more. That escalation turns into directly into energy use, and energy use translates into carbon with few estimates suggesting a single large training run can match the lifetime emissions of many cars. And when you factor in the energy cost of running these models at scale, serving millions of queries per day, the picture becomes even more urgent.

The paper addresses four main questions:

1. How much does it cost to run today's big language models?
2. What are the efficiency contrasts between different hardware platforms?
3. What environmental toll does large scale AI training take?
4. What have been few practical methods to make the development for AI more sustainable?

II. BACKGROUND AND SCALING LAWS

A. The Logic Behind Scaling

There are reason AI labs keep building bigger models: it works. If you increase model size, training data and compute, performance tends to improve in a predictable manner. Studies have captured this relationship in what are now called scaling laws.

But bigger is not always better if the scales are wrong. Later research found that many early models were undertrained for their size – they had lots of parameters but not enough data to make effective use of them. The smarter way is to scale them both simultaneously. That is where the real efficiency gains are, as compute budgets grow, the model and dataset should grow together.

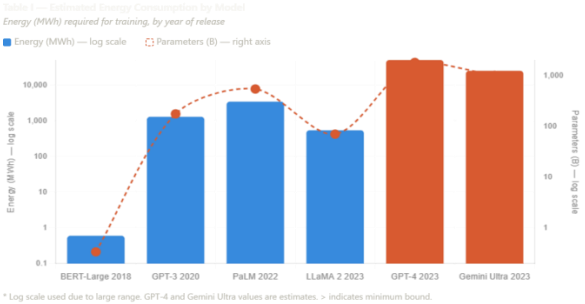
B. Parameter Counts and Compute Requirements

The numbers tell an interesting story. BERT Large, which was large for a model when first introduced in 2018, had nearly 340 million parameters. GPT4 is estimated to run at the trillion-parameter scale just a few years after. That is a growth of several orders of magnitude in a remarkably short time and it explains more than anything else why AI's energy demands have become impossible to ignore.

TABLE I

Model	Year	Params (B)	Train FLOPs	Est. Energy (MWh)
BERT Large	2018	0.34	~1.6 × 10^{21}	~0.6
GPT 3	2020	175	3.14 × 10^{23}	~1,287
PaLM	2022	540	9.9 × 10^{23}	~3,421
LLaMA 2 (70B)	2023	70	~1.7 × 10^{23}	~539
GPT 4 (est.)	2023	~1,800	>10^{25}	>50,000
Gemini Ultra	2023	>1,000	>5 × 10^{24}	>25,000

Comparison of Major AI Models: Parameters and Training Compute



III. ENERGY CONSUMPTION ANALYSIS

A. The Cost of Training

At heart, training a large AI model is a massive, sustained computation, and computation needs energy. The basic formula is not difficult. Work out the total number of floating-point operations needed, divide by the efficiency of your hardware converting electricity into useful work, and scale for the overhead of running a real-world data centre. Modern hardware, like NVIDIA's H100 GPUs, can perform about 312 trillion operations per second in some configurations, and the leading data centers today are surprisingly efficient: they have overhead ratios between 1.1 and 1.5, meaning only a tiny fraction of the energy used for computation is lost to cooling and infrastructure.

But even with efficient hardware, the scale is mind-boggling. Patterson et al. have shown that better hardware selection and smarter architectural choices can lead to a hundredfold reduction in energy consumption compared to a careless setup. Replacing the older GPUs (NVIDIA K80) with Google's TPU v3 hardware for comparable training runs cut energy usage by about a factor of sixteen, a salient reminder that it's as much about how you train as what you train. Just the training run for GPT 3 alone is estimated to have used 1,287 MWh of electricity, which would generate 552 tonnes of CO2 if sourced from the average US power grid. That's it. Beginning in 2023, models will be asked to consume much more.

B. The Hidden Cost of Actually Using These Models

Most of the talk is about the energy footprint of AI, mostly in training, but in the real world running these models is becoming just as significant, if not more so. Picture ChatGPT. A single Google search uses as much energy as 10 ChatGPT searches, says Goldman Sachs. Tens of millions of queries a day times that adds up fast, to about 500 MWh a day, or 182,500 MWh a year. That's not a rounding error, it's a large, growing chunk of the energy budget.

The inference energy is interesting because it has unique character than the training energy. The bottleneck in training is just raw compute throughput -- how many operations you can do per second. The bottleneck at inference time is memory bandwidth. Generating text one token at a time involves loading and processing large amounts of stored context at each step. This distinction is important for hardware design, and is motivating researchers and engineers to build systems that are specifically designed to meet the exact demands of running models efficiently, not just training them.

C. Carbon Footprint Quantification

Carbon Emissions of Prominent AI Training Runs

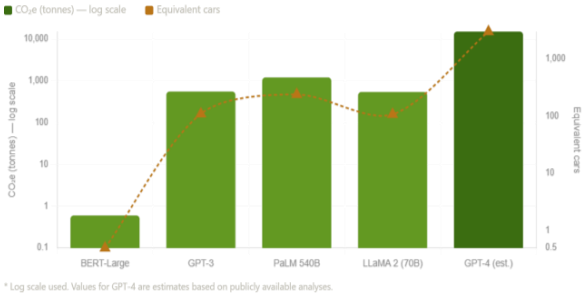


Table II

Model	CO_2e (tonnes)	Equivalent (cars)
BERT Large	~0.6	<1
GPT 3	~552	~110
PaLM 540B	~1,200	~240
LLaMA 2 (70B)	~539	~108
GPT 4 (est.)	>15,000	>3,000

Summarizes carbon emission estimates for prominent training sessions. The trend is clear, as the model size increases, the carbon emissions increase super-linearly unless offset by hardware efficiency improvements or renewable energy sourcing.

IV. HARDWARE ACCELERATORS IN AI PIPELINES

A. GPU Architectures

Despite all the talk about specialized chips, the backbone of large-scale AI training still remains with GPUs. The current gold standard, the H100 from NVIDIA, can do 4,000 trillion operations per second under the right conditions, with memory fast enough to feed that appetite, based on what the company calls its Hopper architecture. String thousands of these together with high-speed interconnects and you have the kind of cluster that makes training trillion parameter models physically possible.

B. Tensor Processing Units (TPUs)

Google took a different route, building its own Tensor Processing Units, or TPUs, based on the kind of math transformer models do all the time. The newest TPU v4 Pods arrange thousands of chips in a three-dimensional mesh, providing more than a quintillion operations per second together. The design philosophy here is tight integration. By keeping compute and memory tightly coupled, Google is reducing the energy wasted just moving data around, which turns out to be one of the larger hidden costs in GPU based training.

C. A New Wave of Challengers

But a growing number of startups and chipmakers are looking at the problem differently. Cerebras avoided the memory bottleneck that plagues traditional chip designs for certain workloads by building its Wafer Scale Engine 2 on a single massive piece of silicon packing 850,000 AI cores. Graph core's IPU goes a different direction, optimizing for the sparse, irregular computations that traditional hardware has a hard time with. Samba Nova's approach is even more flexible, as its chips can dynamically reshape their compute layout to match whatever model they're running. Each of these is a wager that the GPU's reign is not eternal.

V. ENERGY MITIGATION STRATEGIES

A. Model Compression Techniques

One of the most practical ways to reduce energy consumption is to make models smaller without making them dumber. Quantization does this by lowering the numerical precision used to store and process model weights, making a small sacrifice in accuracy for a 2 to 4 times reduction in memory usage and energy per query. Pruning is the scalpel on the model itself, snipping out weights and attention heads that are not pulling their weight. Knowledge distillation is a whole different process: instead of compressing a model directly, you train a smaller model to mimic a larger model, often coming surprisingly close to the original performance at a fraction of the cost.

B. Efficient Architecture Design

Beyond compression, the way models are designed in the first place has a massive impact on efficiency. Mixture of Experts models only activate a portion of their parameters for any given input, letting model capacity grow without a proportional increase in computing cost. Flash Attention rethinks how the attention mechanism accesses memory, cutting bandwidth consumption by 2 to 4 times with corresponding energy savings. Other techniques like Grouped Query Attention shrink the memory footprint during inference, making real world deployment cheaper to run.

C. Green Data Centres and Renewable Energy

Where you put your data centre matters as much as what is inside it. Training a model on Iceland's all geothermal power grid can cut carbon emissions by an order of magnitude compared to running the same job in a region dependent on coal. Even without relocating, data centres can schedule their heaviest workloads for times when renewable energy is most abundant on the grid no extra hardware required. Google, Microsoft,

and Amazon have all committed to matching their electricity consumption with renewable energy purchases, though the details of how those commitments translate into actual emissions reductions remain a subject of debate.

D. Algorithmic Efficiency

Sometimes the biggest wins come not from hardware but from smarter training recipes. Techniques like mixed precision training, gradient checkpointing, and curriculum learning can each shave 20 to 50 percent off the total compute required. And it turns out that training on carefully selected, high quality data as models like Falcon and Mistral have demonstrated can match the performance of models trained on far larger datasets, using 3 to 10 times fewer training steps.

VI. POLICY AND REGULATORY IMPLICATIONS

Regulators are starting to take notice. Under the EU's AI Act, developers of high-impact AI systems must reveal how much compute and energy their models use. Recent US executive orders on AI safety include provisions such as reporting large training runs to federal agencies. While these are important steps toward accountability, they fall short of requiring efficiency improvements.

Some researchers have called for something more systematic a standardized "nutrition label" for AI models that would report parameters, training data, compute used, and estimated carbon footprint in a standard, comparable format. If such a framework were widely adopted, it would be much easier to hold developers accountable and to reward real gains in efficiency.

There is also a better case for shared public infrastructure. As with existing government support for shared facilities for particle physics or genomics research, a national AI compute resource could provide large scale training access to researchers without each lab having to duplicate expensive infrastructure. Both the US and the EU are trialling variants of this idea and while it is early days the direction is promising.

VII. CONCLUSION

The energy cost of building and running advanced AI systems is real, large, and growing. Largest models trained in 2023 alone used tens of thousands of megawatt hours and emitted thousands of tons of carbon dioxide, and that trajectory is increasing.

The good news is that there are levers that matter. Smarter algorithms, better hardware, cleaner energy sources, and more thoughtful infrastructure decisions can all help a lot. The challenge is to ensure that these efficiency gains keep up with the relentless growth in model size and deployment, because if they do not, the net impact continues to get worse even as individual systems improve.

What is needed is a change in how the field thinks about energy. For now, it is an operational concern – something that the infrastructure team cares about. It needs to be a research priority, as significant as accuracy and capability, as a metric that matters. Most importantly, consistency in energy reporting standards across AI research needs to be developed, hardware and software should be designed together for inference efficiency rather than separately, fundamentally different computing paradigms such as sparse or neuromorphic approaches should be explored, and the hard physical limits of what efficient computation can achieve should be better understood.

ACKNOWLEDGMENT

We would like to thank the open-source AI research community for the work of the researchers, developers and volunteers who make their work publicly available and enable studies like this one. Thanks also to everyone who maintains the public benchmarks and datasets that the field relies on; that work is often thankless but it is genuinely invaluable. This research was conducted independently and received no external funding.

REFERENCES

[1] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, Mar. 2023.

[2] Amodei, D., and Hernandez, D., "AI and Compute," *OpenAI Blog*, May 2018. Available: OpenAI Blog

[3] Strubell, E., Ganesh, A., and McCallum, A., "Energy and Policy Considerations for Deep Learning in NLP," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, Jul. 2019, pp. 3645–3650.

[4] Kaplan, J. et al., "Scaling Laws for Neural Language Models," *arXiv preprint arXiv:2001.08361*, Jan. 2020.

[5] Hoffmann, J. et al., "Training Compute-Optimal Large Language Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, Dec. 2022, pp. 30016–30030.

[6] Patterson, D. et al., "Carbon Emissions and Large Neural Network Training," *arXiv preprint arXiv:2104.10350*, Apr. 2021.

[7] Brown, T. et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Dec. 2020, pp. 1877–1901.

[8] NVIDIA Corporation, "NVIDIA H100 Tensor Core GPU Architecture," Technical Whitepaper, Mar. 2022. Available: NVIDIA

[9] Hinton, G., Vinyals, O., and Dean, J., "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, Mar. 2015.

[10] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C., "FlashAttention: Fast and Memory Efficient Exact Attention with IO Awareness," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, Dec. 2022.

[11] Lottick, P. et al., "Energy Usage Reports: Environmental Awareness as Part of Algorithmic Accountability," *arXiv preprint arXiv:1911.08354*, Nov. 2019.