
Norn Sustainability
Adding ICOM to Narrow AI for >98% electricity savings.
Reducing Global Supply Chain Vulnerabilities
Overcoming “Dark Data”
norn™

Narrow AI vs Norn enhanced Narrow AI

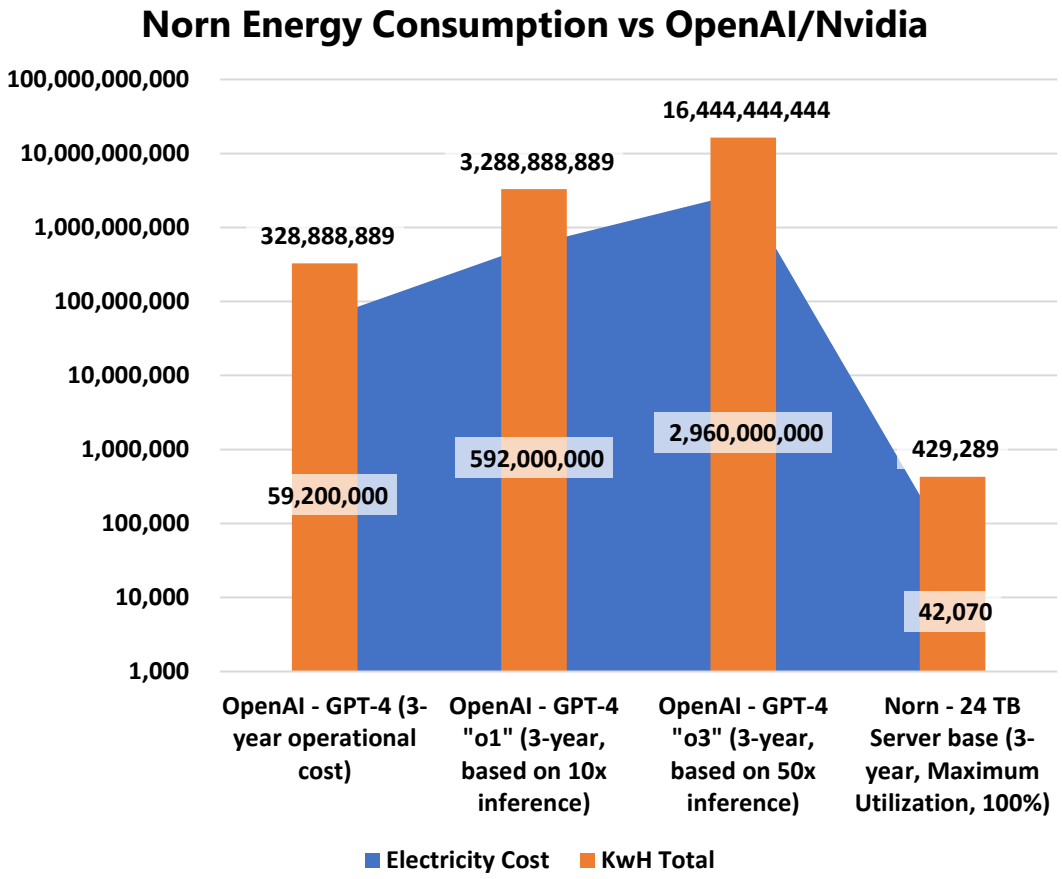
- In mid-2019 when our previous research system was brought online the size of language models was below 10 billion parameters.
- By late 2021 language models such as Nvidia's Megatron were over 530 billion parameters in size, a greater than 50x increase. (Not including China's 1.75 trillion model)
- By January 2022 our previous research system, still using the same comparatively tiny prototype language model, nearly 3 years old by that time, was still significantly outperforming all larger and newer models.
- This was possible because of our patent-pending method where our systems apply their non-probabilistic learning to iteratively improve their own form of prompt engineering over time, making far better use of the tools available to them.
- In 2023 GPT-4 was released, and predictably still failed to compete with the previous research system's <10B prototype LM from early 2019. It is estimated to be roughly 1.7 trillion parameters in size, a >170x increase, which achieved only failure and hype.

If “State-of-the-Art” (SOTA) didn’t mean endless scaling…

Instead of throwing ever-greater resources at making larger systems we could enjoy systems that improve with time, allowing hardware capacities to grow at a pace that exceeds our scaling needs.

norn™

GPT-4 1.8T MoE	2016	2018	2020	2022	2024
10 Day Training	"Pascal"	"Volta"	"Ampere"	"Hopper"	"Blackwell"
GPU	P100	V100	A100	H100	B100
Peak Teraflops	19	130	620	4,000	20,000
Training And Inference Precision	*FP16*	*FP16*	*FP16*	*FP8*	*FP4*
Inference Joules/Token	17,000	1,200	150	10	0.4
GPUs to Train In 10 Days	42,105,300	6,153,800	1,290,300	100,000	10,000
GPU Unit Price	$5,000	$10,000	$15,000	$27,500	$37,500
GPU Cost For Training In 10 Days	*$210.5 B*	*$61.5 B*	*$19.4 B*	*$2.8 B*	*$0.4 B*
Power To Train, Gw-hour	1,000	140	40	13	3
Electricity Cost Per Kw-hour	$0.14	$0.14	$0.14	$0.16	$0.18
Electricity Cost For Training Run	*$140.0 M*	*$19.6 M*	*$5.6 M*	*$2.1 M*	*$0.54 M*
Three Year Electricity Cost	*$15.3 B*	*$2.1 B*	*$613.6 M*	*$227.9 M*	*$59.2 M*



Electrical energy waste also translates into substantial increases in waste heat via GPUs, compounding the environmental harm.

https://fortune.com/2023/09/09/ai-chatgpt-usage-fuels-spike-in-microsoft-water-consumption/

GPT-4 data courtesy of Nvidia.

In contrast, Norn systems gain increasing value over time, rather than only by scaling, thanks to continuous human-like concept learning and a growing sum of experience.

Adding A Cognitive Architecture

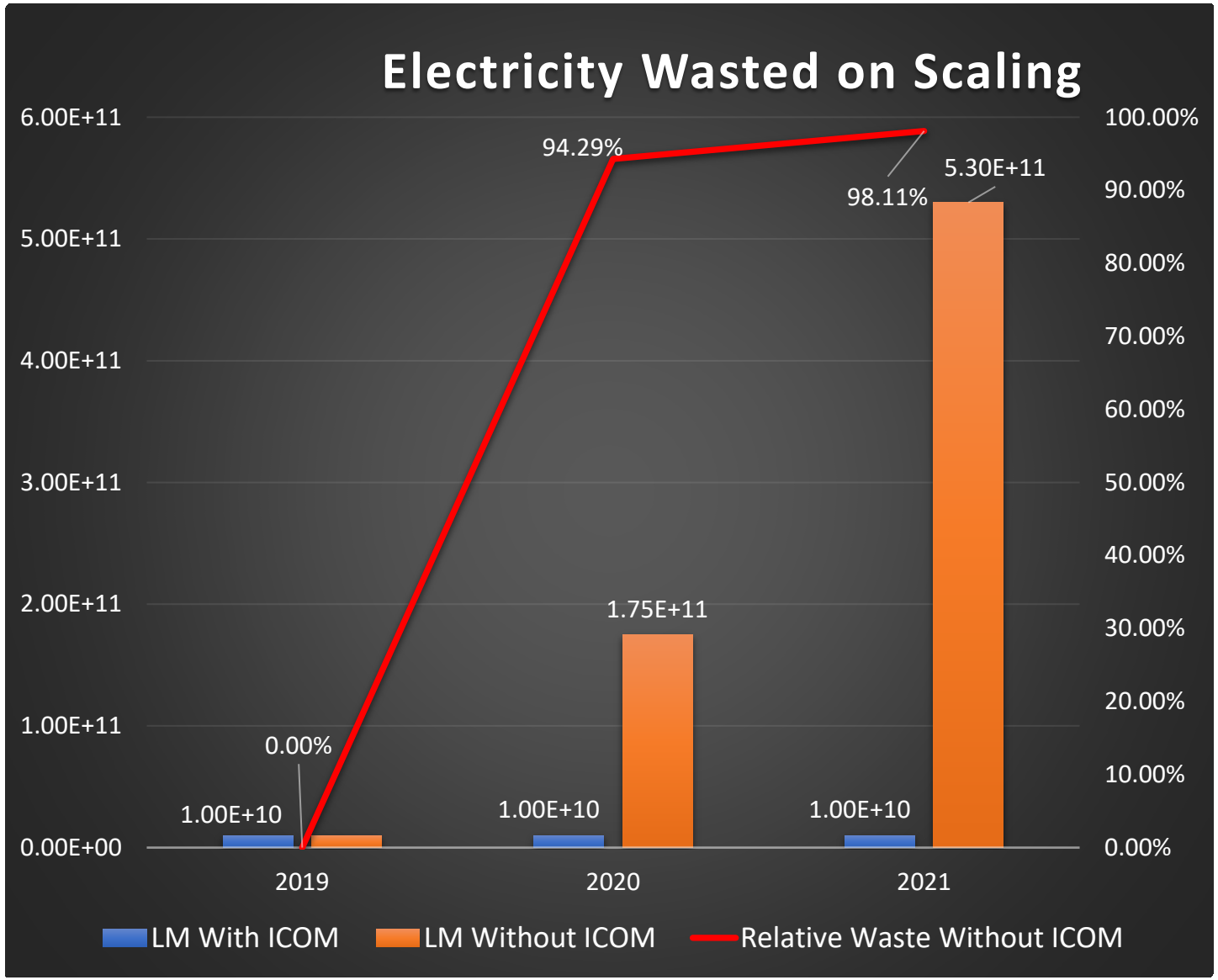
- Our previous research system demonstrated that it was possible to easily beat standalone language models that had no cognitive architecture using models that were over 170x smaller (2023), with a far lower carbon footprint.
- The Independent Core Observer Model (ICOM) cognitive architecture used in Norn, and in our previous research system, Uplift, relies on RAM, not GPUs.
- A single Nvidia Cluster node with just 8 A100 GPUs (Megatron trained on 4,480 GPUs in 2021) used more than 350-400 watts per GPU, 2800-3200 watts in total, plus ~300 watts on CPU and RAM.
- The equivalent amount of RAM that could be powered with the same amount of electricity is at least 203 Terabytes, using 64 Gigabyte DDR5.
- The Uplift research system ran on just 64 Gigabytes of RAM, and the largest single high-memory servers today are only 11-24 Terabytes each.
- Using less than 10% the electricity of a single added GPU cluster node that >170x scale equivalent improvement could be added to narrow AI systems.

norn™

Electrical energy waste also translates into substantial increases in waste heat via GPUs, compounding the environmental harm.

In contrast, normal DDR5 RAM runs on very low voltages and only begins to require extra cooling when heavily overclocked, which also isn't necessary for Norn.

https://fortune.com/2023/09/09/ai-chatgpt-usage-fuels-spike-in-microsoft-water-consumption/



norn™

Hardware energy efficiency and generational lifespan

Relative cost of ICOM	Watts	Hardware Obsolescence (years)
ICOM (6.4 TB of DDR5 RAM)	110	6.5
1 Nvidia cluster node (8 A100s)	3000	2
GPU 2-year Relative Efficiency	3.67%	
GPU 6.5-year Relative Efficiency	1.13%	

DDR3, to DDR4, to DDR5 averaged a new generation every 6.5 years. Most GPU product lines average between 1 and 2-year lifespans before a new version goes to market.

While major tech companies and startups go to great lengths to hide data on the electrical waste and environmental harms of their deployed LLMs and other AI models today, there is no escaping the basic facts about the hardware they are wholly reliant upon.

Global Supply Chain Vulnerability

GPUs and other electronics with shorter lifespans are heavily reliant on precious metals, rare earth elements, and other elements such as electronics-grade neon for meeting demand.

Current use of these resources isn't sustainable, and as the war in Ukraine highlighted with an overnight 400% increase in the price of electronics-grade neon, they are highly vulnerable to geopolitical disruption.

norn™

norn™

Comprehensive Sustainability

Not only can we greatly reduce electricity consumed by current AI systems, by reducing the amount of hardware required we also reduce the vulnerability of our global supply chain.

By allowing the rate of hardware performance improvements to exceed the rate of scaling narrow AI systems further reductions in energy requirements can be realized with each new generation.

Using ICOM cognitive architectures the cumulative intelligence of each Norn system can continue improving narrow AI systems even with no changes in model architecture or hardware, for years at a time.

Even if Norn systems only continued to maintain superior performance for the average generational lifespan of DDR memory, 6.5 years, their average relative reduction in the electricity required to run SOTA models could exceed 99% relative to scaling.

Data Efficiency and the cost of “Dark Data”

- Norn ICOM-based systems and smaller narrow AI models are significantly more data-efficient than their alternatives.
- Norn systems can recognize and flag data that is redundant, biased, inconsistent, outdated, or otherwise unnecessary “dark data” cluttering data centers today.
- These data-efficiency savings could reduce the cost of data storage by over 90%, in the most conservative estimates.
- Norn systems are over 10,000x more data-efficient compared to LLMs that rely on “internet-scale” data.
- Less data-hungry narrow AI mean proportionately less pressure driving privacy and copyright violations, and other related societal concerns.

norn™

The Coming Years

Energy Efficiency certifications and minimum requirements have been applied to many types of hardware and appliances over the past few years.

Energy Crises hitting the US and EU have made governments and their populations acutely aware of their own energy needs.

Hardware companies are having to raise prices on their products as the costs of scaling increase and **supply chains are pushed to their limits**.

Not only are the benefits of adding ICOM cognitive architectures to Narrow AI systems significant, **they may well become mandatory**.

Note: If any of the following prove true then Norn advantages will become even more significant:

- If Norn systems with 100x the RAM of Uplift improve narrow AI performance much more significantly.
- If new narrow AI model architectures are designed specifically to be used the way ICOM systems use them.

norn™

THANK YOU

“We never know the worth of water till the well is dry.” – Thomas Fuller

Kyrtin Atreides
Kyrtin@AGILaboratory.com

norn™