2024 AI Report

MÆc

Written by Michael A.

Everything happening right now revolves around neural networks.

- **Traditional Approach:**
  Neural networks were designed to provide instant answers, optimizing for speed and efficiency.
- **New Insight:**
  Researchers realized that delaying the need for immediate answers could improve **reliability**. By spending more time and computational resources on **post-training** processing, they could achieve better outcomes.
- **Chains of Thought (CoT):**
  By making **multiple queries** to the network and comparing results, it's possible to improve reliability and create **inference chains**. These chains allow for more complex **problem solving** by using the pre-trained network to make **corrective adjustments** and refine its own errors.

Impact:

- **Higher accuracy** in decision-making
- Enables **true problem solving** with AI
- Allows the system to **self-correct** based on stored knowledge

Post-Training Computation for Reliable Problem Solving

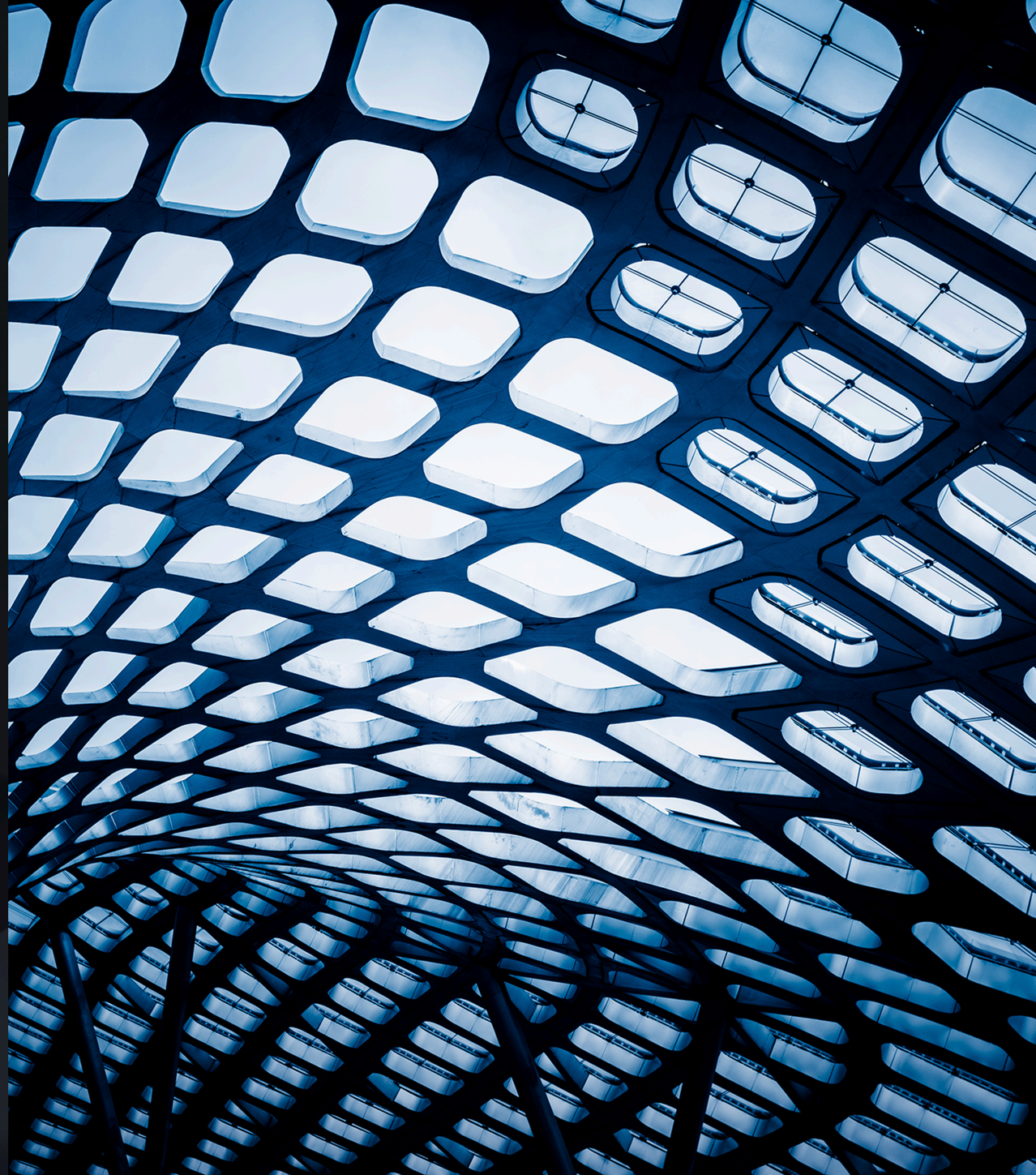
The Power of GPUs in Neural Networks

Neural Networks: Simple Operations at Scale

- Implementation requires **basic multiplication** and **addition** on a massive scale.
- Massive **parallelism** is key to neural network performance.

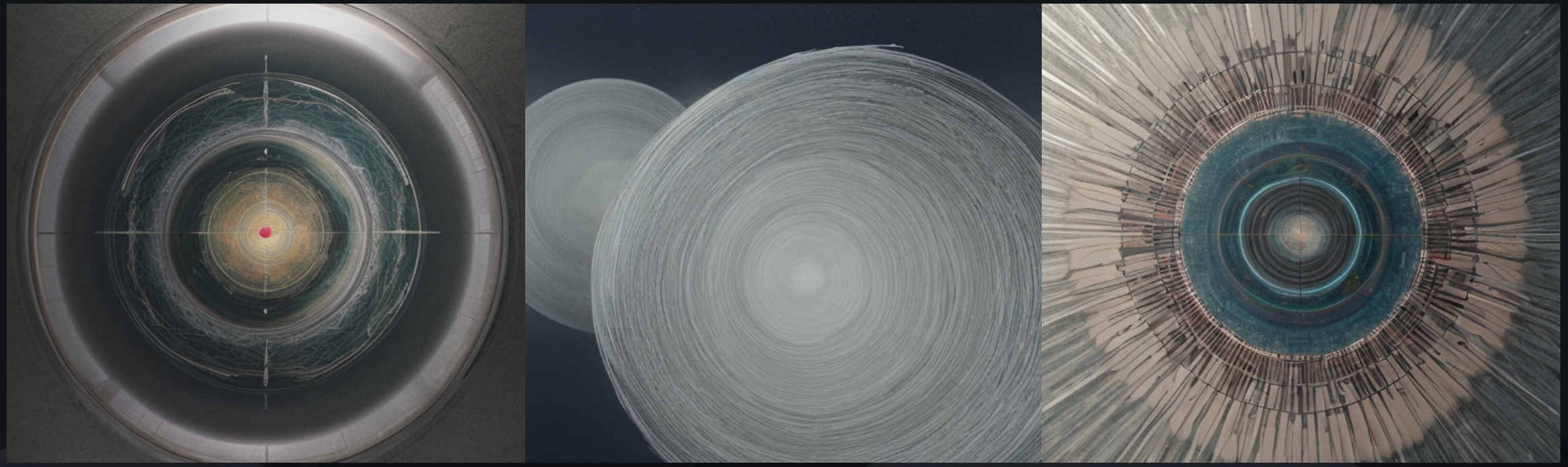
Why GPUs Are Ideal:

- Originally designed for **3D gaming** calculations.
- Later repurposed for **cryptocurrency mining** (solving hash problems).
- Now central to **neural network AI**, handling millions of operations simultaneously.

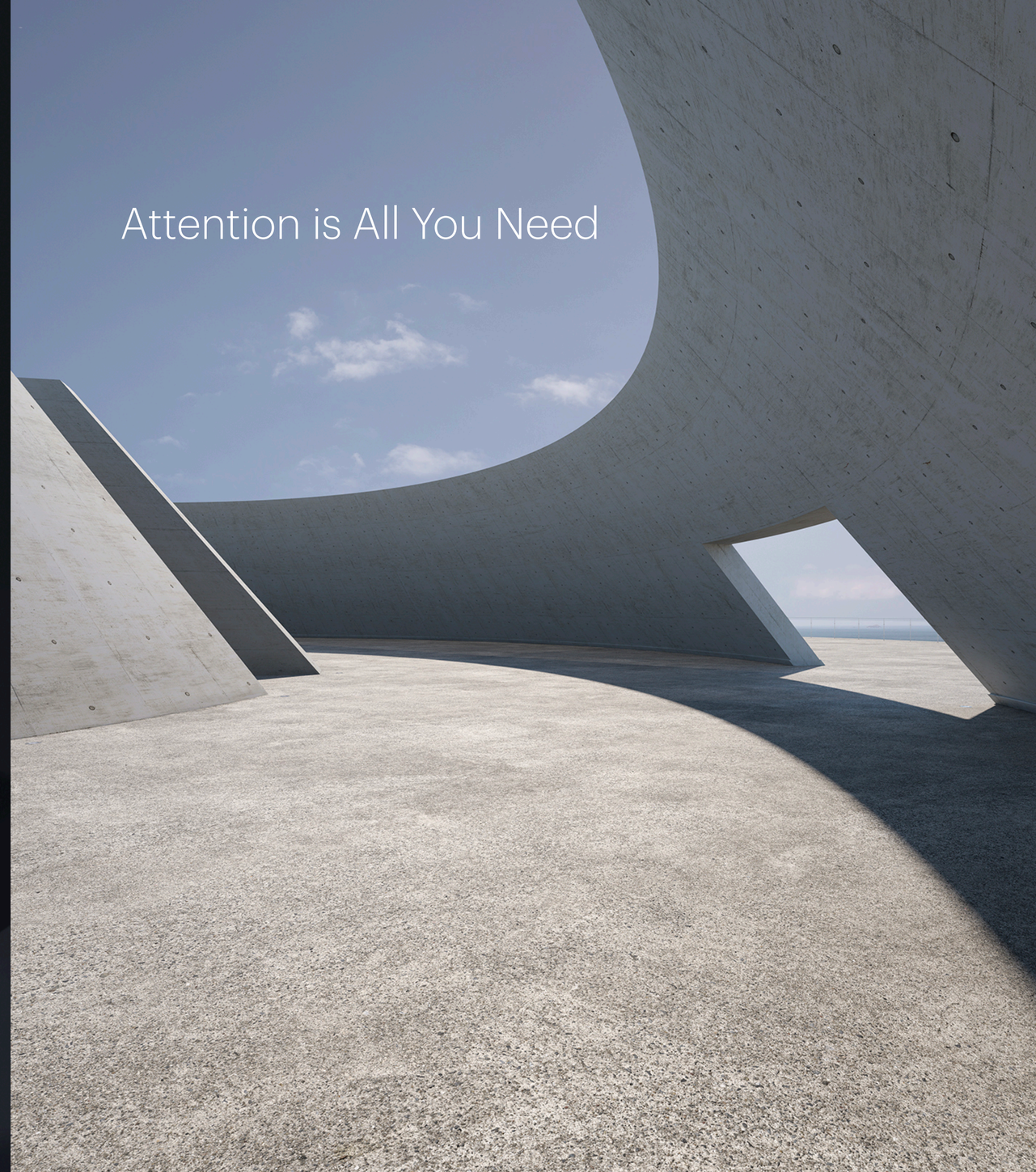


Is it AGI?

OpenAI states that it's not quite there yet, but they're getting closer — and the evidence strongly supports this claim. The independent benchmarks and performance tests referenced in the video above are particularly compelling.



Reading This



AI Computation

AI Computation Stage	Description	Process
Pre-training	The monumental task of feeding information into a massive, untrained neural network.	The network's output is compared to the correct output, and neural parameters are tweaked to improve accuracy.
Test-time (Inference-time)	The phase where the trained model is used to make predictions.	No more training occurs; the model applies learned patterns to new data.

AI computation consists of two key phases: *pre-training* and *test-time* (also known as *inference-time*). Pre-training involves feeding information into an untrained neural network and adjusting its parameters by comparing outputs to correct answers, refining the model iteratively. Modern neural networks can have up to 185 billion parameters, which are fine-tuned millions of times across vast datasets to achieve accurate predictions during test-time.

o1 vs o3

o1 Model	o3 Model
Availability: First widely available inference-chain AI model.	Coming Soon: Not yet available, but anticipated to surpass the o1 model.
Performance: Offers significant improvement over previous ChatGPT 4.0 models.	Performance: Expected to greatly outperform the o1 model.
Cost & Usage: Expensive to run; subscribers are limited to 7 full queries per day due to high computation costs.	Cost & Usage: Unknown; details on availability and cost are forthcoming.
o1-Mini Version: A scaled-down version, better than earlier models but not as powerful as o1, available without query limits.	Anticipation: Expected to revolutionize AI capabilities even further, with more advanced inference chains.

https://www.youtube.com/watch?v=SKBG1sqdyIU