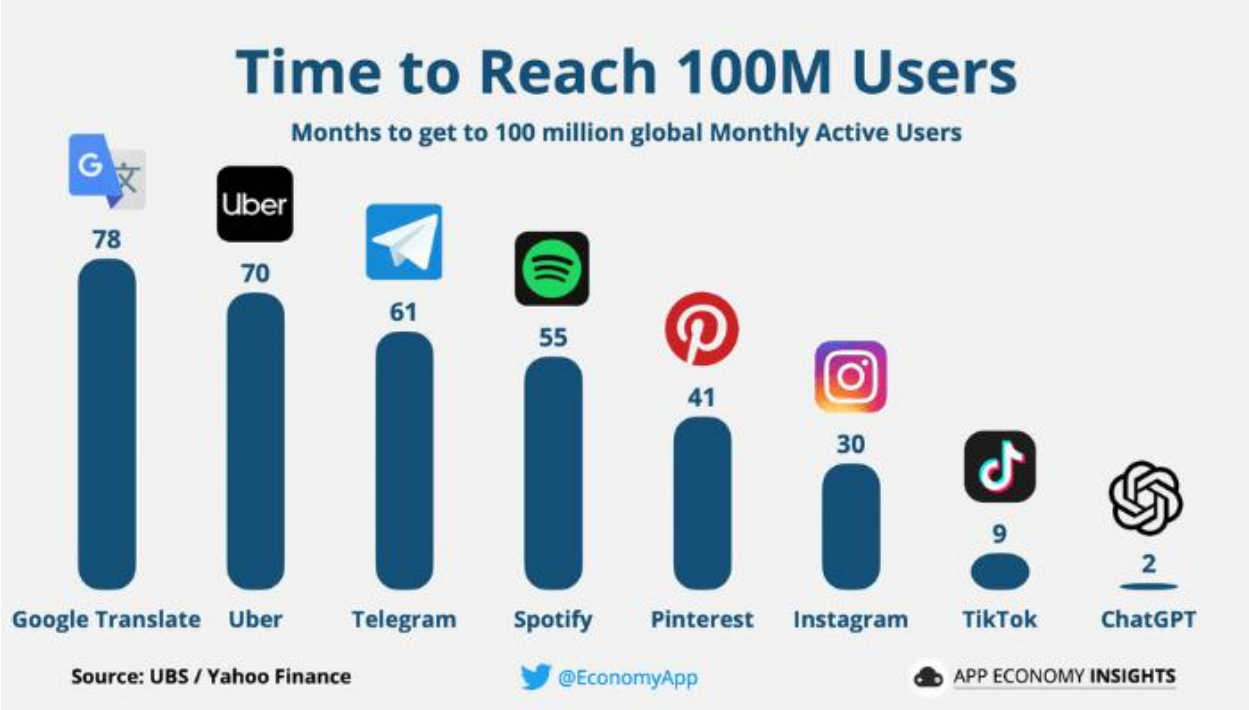
A Primer on Artificial Intelligence

The world has transformed significantly since 5 years ago. We now live in the age of artificial intelligence – an age where the massive sums of information we've collected during the internet era can be used to create, predict, and automate tasks too tedious or too complicated for a person to complete. One major event that signaled this global, accelerating change was the release of OpenAI's ChatGPT3 in 2022.

For the first time ever, non-technical individuals could access and utilize AI models for everyday tasks. And it took the world by storm – becoming the fastest product to reach 100 million users in 2 months, and setting off a global frenzy to develop better AI models, as evidenced by Anthropic's new Claude models, Mistral AI's le Chat, Google's Gemini Models, Meta's Open Source LLaMA models, and many, many more.



But this explosive frenzy for better, larger, and more capable models has a hidden consequence. This whitepaper aims to address how AI will exponentially exacerbate the global climate crisis, current solutions to reducing AI's environmental impact, and highlight a key technological development being developed at this company to drive AI's environmental cost down to zero.

History

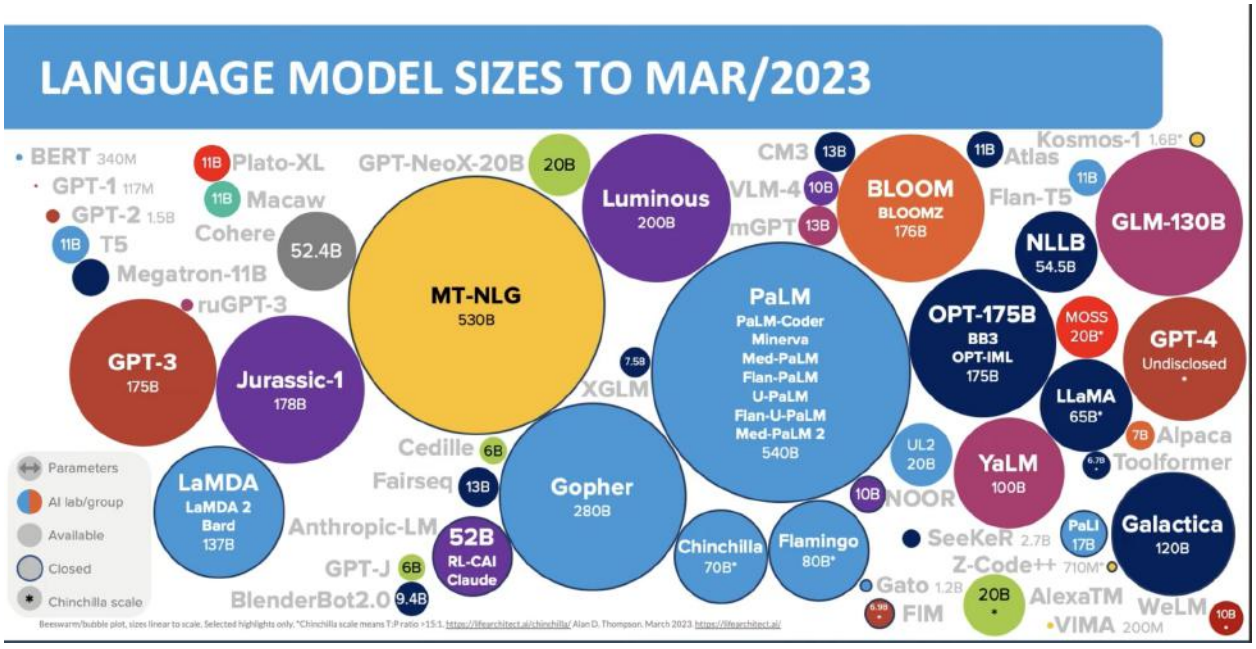
Contrary to mainstream belief, AI research is not a recent development. Breakthroughs in computer science, mathematics, and neuroscience since the 1900s have all served as catalysts

for AI development, leading to slow, but incremental progress, and ultimately culminating to powerful, user-friendly AI models commonly used by the public today.

The beginnings of AI in the 1950s was driven by the rudimentary understanding of how neurons (brain cells) in organisms learn, and implementing that knowledge into computer programs as “artificial neural networks” (ANN). Upon iteration and newer understanding of ANNs, more models were developed, including “deep neural networks” (DNNs) and “reinforcement learning” (RL).

By 1997, IBM’s Deep Blue beat world chess champion Gary Kasparov, and a little more than a decade later, IBM’s Watson beat two former Jeopardy champions. By the 2010s, companies such as Twitter, Facebook, and Netflix started utilizing AI as part of their advertising strategy and user-experience algorithms. By 2021, OpenAI created both GPT-3, a novel DNN trained AI to create virtually indistinguishable human-like content, and DALL-E, which can process, understand, and generate images. Currently, hundreds of advanced AI models have been trained – with popular models including:

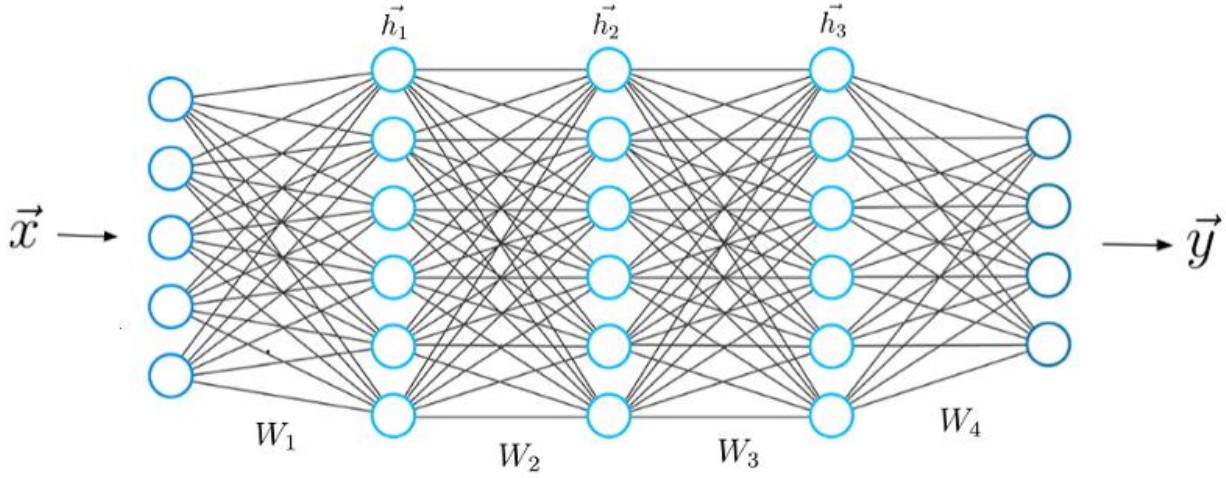
- GPT3, GPT4, GPT4 Vision, GPT4 Turbo, Sora, DALL-E by OpenAI
- Copilot by Microsoft
- Grok by xAI
- PaLM, Bard, Gemini Ultra, Pro, and Nano by Google
- CodeLLaMa, LLaMA and LLaMA 2 by Meta
- Claude 2.0, 2.1, and 3.0 by Anthropic
- Falcon 180B and Falcon 40B by Technology Innovation Institute
- Mistral-7B by Mistral
- Coral by Cohere



How are AI Models Trained?

The hardware driving AI progress currently lies in graphics processing units (GPUs). Created primarily by Nvidia, these GPUs that have fueled the AI boom have become so valuable, major companies reportedly transport them via armored car. In fact, according to a16z, one of the most successful venture capital firms, the “the supply of compute [from GPUs] is so constrained that demand outstrips it by a factor of 10x” and that an average company building in AI spends “ 80% of their total capital on compute resource.” And there is also no sign that the GPU shortage we have today will abate in the near future.

These GPUs are used to fine-tune parameters within an AI model. Parameters are variables that models can adjust during their training process to improve their ability to make accurate predictions, and having additional parameters allow more granular fine-tuning for a more accurate prediction. For example, parameters of DNNs consist of the weights assigned to the connections between artificial neurons (labeled as h1, h2, and h3 below).



Due to the recent explosion in larger AI models, there has been the emergence of "giant models,” reaching billions or trillions of parameters. While these huge models have achieved massive improvements in performance, they come with a significant computational cost. This is because these “giant models” require a lot more GPUs to train and a lot more time to train. For example, a 175B parameter AI model (like GPT-3) requires over 1000GB of data memory. This exceeds the memory capacity of a single GPU (for reference, Nvidia’s cutting edge A100 chip only has 40GB of memory), and thus requires a model to be split across hundreds of cards for thousands of hours. In effect, this means that:

1. More GPUs are needed to train
2. More powerful (and thus power-hungry) GPUs are required to train
3. More GPU time is needed to train

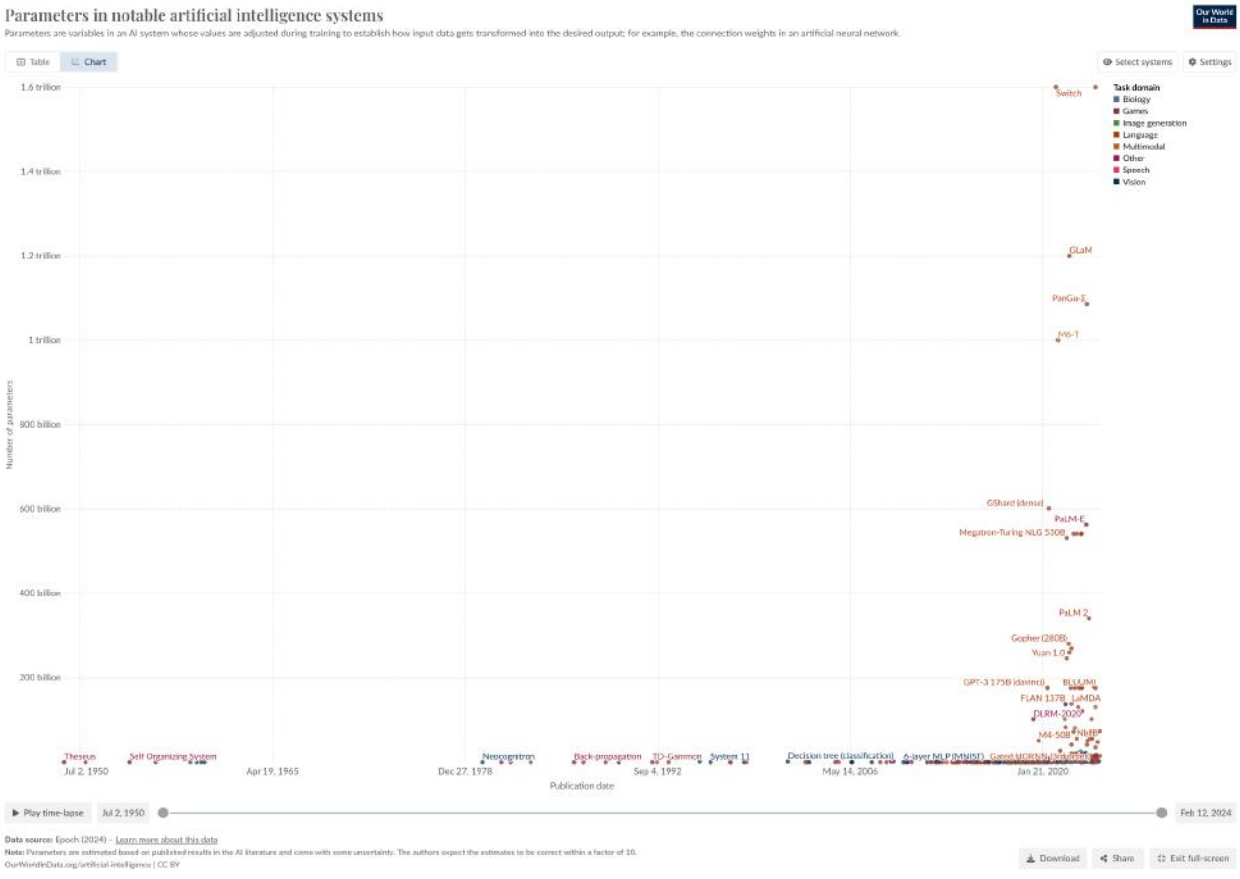
Which means more energy being used.

The AI Race is On

Progress on AI is accelerating. Based solely on OpenAI's GPT models, the number of parameters used by each AI model increases by roughly 10x every 2-3 years.

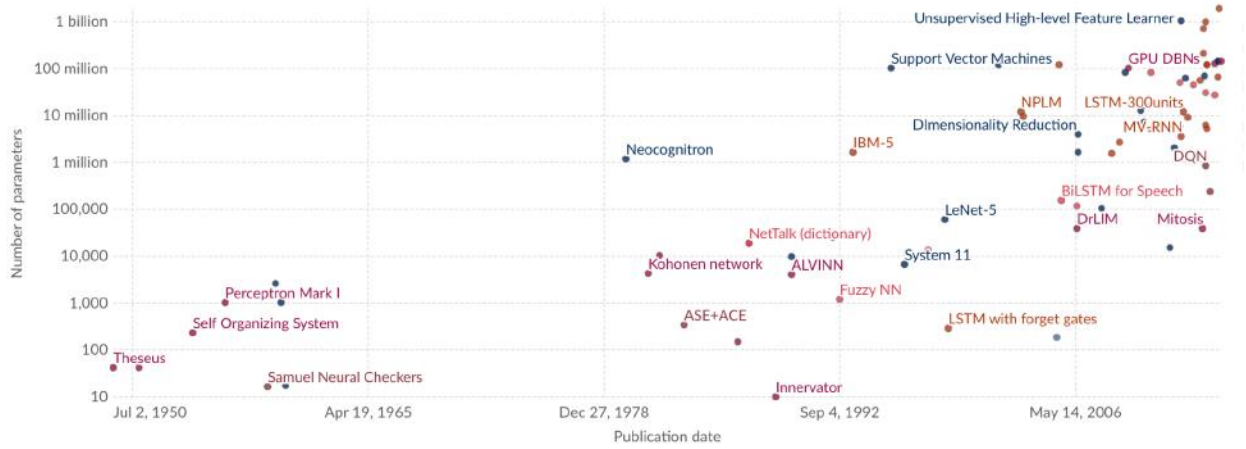
Year	Model	Parameters
2018	GPT1	117M
2019	GPT2	1.5B
2020	GPT3	175B
2023	GPT4	1760B

And across all companies developing AI models, it is clear that the trend of increasing AI model size does not appear to be ceasing.

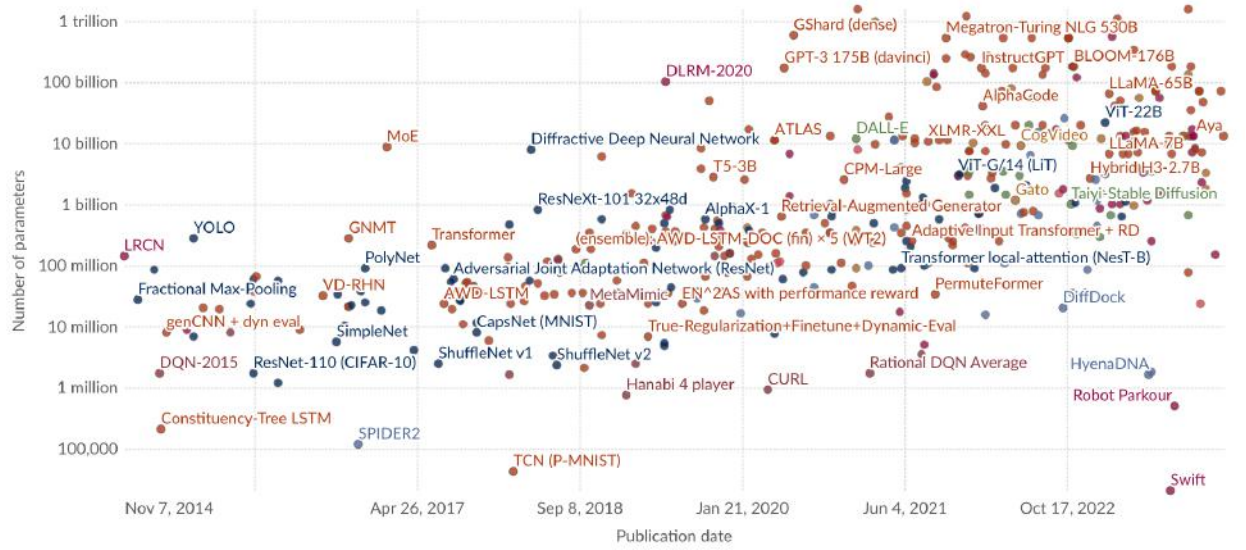


Likewise, in terms of total AI models being trained, it is clear that more and more companies and startups are developing their own.

Below is the number of AI models trained from 1950 to 2014.

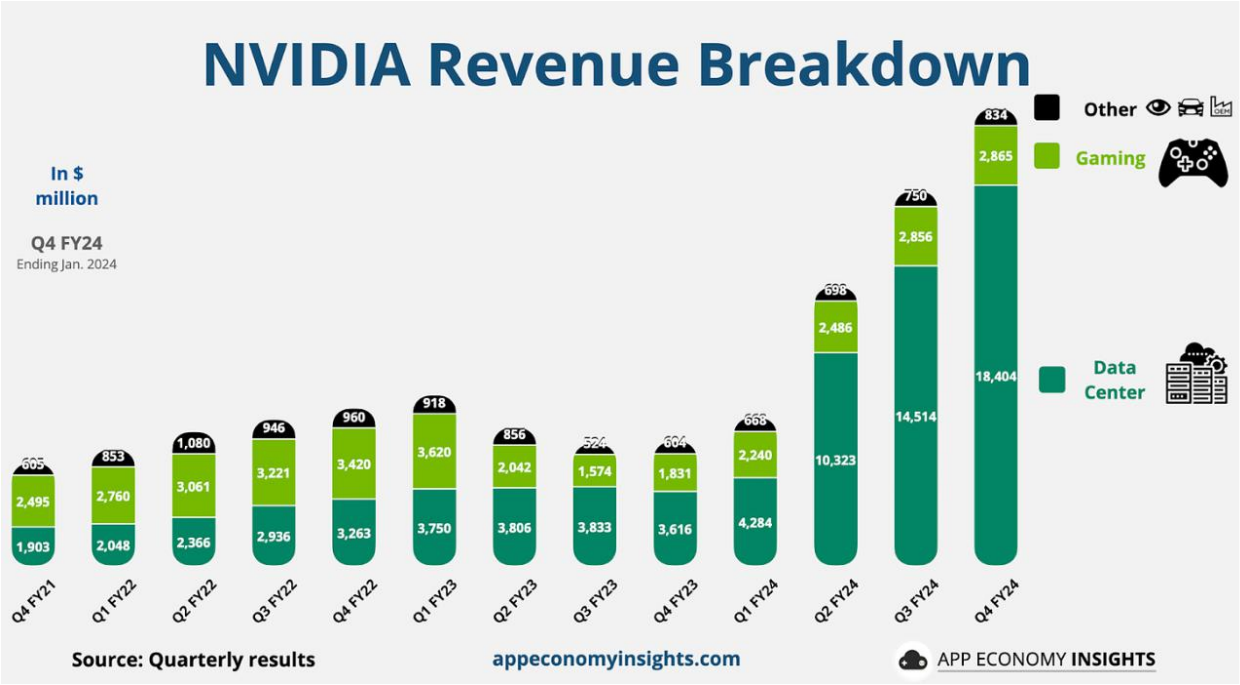


And from 2014 to 2024.



Clearly, a lot more.

Another indicator of the accelerating AI race can be seen in Nvidia’s earnings report. According to their Q1F2024 earnings’ report, the company stated a 404% growth in data center revenues within the past 10 years. In fact, Nvidia’s $47.5 billion in 2023 data center revenues is 18% more than total data center revenues for the past 5 years, combined.



The Age of Artificial Intelligence has arrived, and it's here to stay, along with its negative consequences.

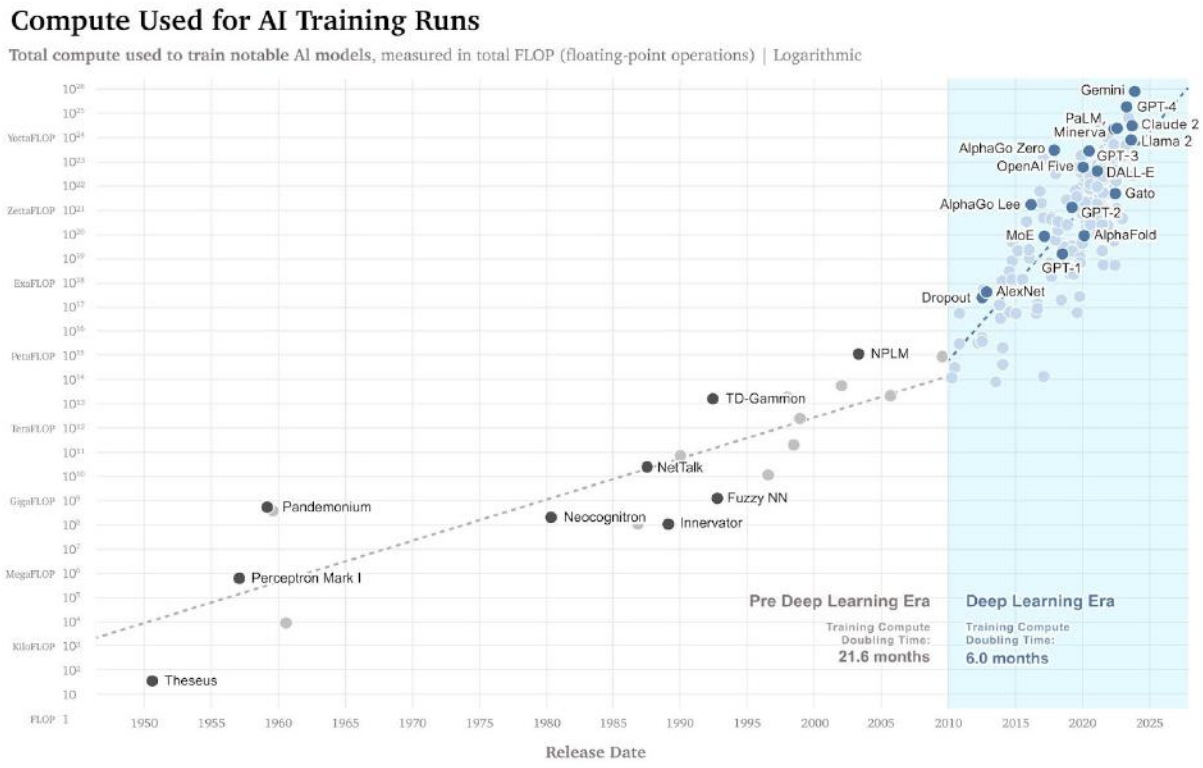
The Impact of AI on Energy and Climate

- Proposition: something along the lines of "We need to be more energy efficient to train AI"

According to the United Nations, climate change is the defining crisis of our lifetime, primarily driven by human activity. If the climate mitigation goal by bringing global emissions to zero by 2050 and limiting the average global warming to 1.5C fails, we will have a global environmental catastrophe of no return.

The acceleration of AI development directly counters this goal.

As the training and deployment of larger AI models grows more elaborate and data-intensive, two things begin to scale up exponentially: the need for more memory storage and the need for more energy. Already, the rise of AI in the past 2 years alone have led to a significant growth in data center utilization and their associated energy usage. Across all AI models from 2012 to 2022, the amount of computing power (provided by data centers) required grew by about 1 million percent, according to Princeton Professor Varma.



In fact, the amount of compute required by current AI systems and models seems to double every 6 months, while OpenAI's research suggests that the compute required has been doubling every 3.4 months since 2012.

The exponential demand for computing power can also be generalized from the explosive growth of data centers revenues from Nvidia, as previously evidenced. Meanwhile, traditionally software-focused companies like Amazon, Google, and Meta have also been building more data centers all over the country – resulting in increased carbon dioxide emissions that drastically exacerbate the climate crisis. For example, this led to Amazon's carbon emissions increasing by 15% last year.

In fact, data center power and carbon emissions associated with data centers roughly doubled or tripled between 2018 and 2020. In 2018, global data centers consumed roughly 1-2% of the global electricity supply. By 2020, this figure was estimated to be around 4-6%, according to the International Energy Agency. If this trajectory is maintained, by 2030, global energy usage from data centers is projected to rise to 8-21%, creating a global energy shortage crisis. A recent peer-reviewed study from Joule by Dr. Vries estimated that AI would consume at least 85,000,000,000,000W of electricity every year by 2027. For reference, the entire country of Netherlands consumes marginally more electricity (117,000,000,000,000W).

In agreement, research from the University of Massachusetts Amherst also identified AI as a significant emitter of carbon. It is estimated that the process of training each moderate-sized

language model results in 300,000kg of carbon dioxide emissions. (It is generally accepted that a moderately-sized LLM has about 40B parameters.) In comparison, GPT4 was released with over 1760B parameters and Gemini Ultra at 1560B. The newly released Claude-Opus in 2024 has 2000B parameters alone.

Put simply, AI is compute bound. And the problem is, for a linear gain in AI performance, exponentially training parameters and thus compute is required. And that means exponentially more carbon emissions and energy consumption. As of now, the field of AI remains unsustainable. **Thus, AI training, deployment, and research needs to be made significantly more energy efficient to align with climate change goals.**

Current Flawed Solutions

Software

- Talk about why making AI training more energy efficient from a software perspective is unlikely/unfeasible as a solution

OpenAI Model	Release Date	Parameters, B	MMLU
GPT2	2/14/19	1.5	0.324
GPT3	6/11/20	175	0.539
GPT3.5	3/15/22	175	0.7
GPT4	3/14/23	1760	0.864

Even as parameters increased by 1000x, the MMLU score only tripled. Clearly, exponential compute is needed to accelerate AI progress. And with compute, comes increased energy usage.

Thus, the cost of training AI is exponentially increasing. For example, the Falcon-40B model was trained on 384 A100 40GB GPUs, and it took two months. If you rent $2.0/hr, then $2.0 x 384 GPUs x 24 hours x 30 days x 2 months = $1,105,920. Falcon-7B, the smaller model with only 7 billion parameters, took two weeks. Still, that means $2.0 x 384 GPUs x 24 hours x 14 days = $258,048. What about GPT4's 1760 billion parameters? It would cost a lot more.

Energy Scaling

- Part 2: Talk about why scaling clean energy supply in the grid (whether that be through solar or nuclear) is unfeasible in a 10-15 year timeframe to make a meaningful impact on climate change

While Amazon and Microsoft are buying nuclear power plants to power their data centers, the incentive to use clean energy to power data centers is minimal, and thus not widely used. In Virginia, USA, the data center hub of the world, only 1% of electricity comes from renewable sources.

Hardware

- Part 3: Talk about why advancing silicon chip technology is not a viable option due to breaking the laws of physics and the shortage of silicon

Previously, traditional silicon hardware made exponential progress. Driven by both Moore's Law (the number of transistors on a chip doubles every two years) and Dennard's Scaling Law (doubling the number of transistors effectively means shrinking them but also maintaining their power density, so smaller chips meant more energy-efficient chips), silicon chips has been the go to computers.

But Moore's Law and Dennard's Scaling Law have slowed. Each individual transistor is so small – smaller than a virus – that chip manufacturers are breaking the fundamental properties of physics. And thus, improving traditional silicon hardware to make better, denser, and more efficient chips have also slowed dramatically in progress.

Taking into account that the number of transistors on a chip doubles every three to four years now, and coupled with the fact that current AI systems and models require the doubling of compute every 6 months, silicon computing is no longer the answer.

Gopalakrishnan said that innovation within existing computing architectures, as well as improvements in silicon technology, began slowing at exactly the time when AI began creating massive new demands for computation power and efficiency. Not even the best graphics processing unit (GPUs), used to run today's AI systems, can mitigate the bottlenecks in memory and computing energy facing the industry. "A new type of chip will be needed to unlock the potential of AI."

Another problem with silicon chip manufacturing is, well, silicon itself. An estimate from the Semiconductor Research Corporation, a consortium of all the major semiconductor companies, posits that if we continue to scale data at this rate, which is stored on memory made from silicon, we will outpace the global amount of silicon produced every year.

Another problem with silicon hardware is the general release of heat as a byproduct of computing. Currently, Nvidia GH200 AI chips use kilowatts of energy. In a typical server rack, that means 21-24 kilowatts of thermal load needs to be dissipated – resulting in huge energy costs just to cool the silicon hardware required to power AI. Newer GPUs require additional cooling systems: AMD's new accelerators jumped from 560W to 760W, and Nvidia's new rumored chips are projected to use over 1000W. Chips will get hotter and this trend is likely to continue. It is estimated that global data centers, on average, will add 50% to their energy usage just to keep silicon chips cool.

In cases outside of pure AI models, such as autonomous vehicles, researchers from OpenAI have stated that the amount of compute to train AI systems have increased 350 million times in the past 15 years. Another source states that the amount of compute required by current AI systems and models doubles every 6 months, while OpenAI research reveals that the compute used in large AI models has been doubling every 3.4 months since 2012.

The Convergence of Silicon and Biology

The Interconnected History of AI and the Brain

- History of how AI is inspired by the brain

Digital Biology: Engineering, no Longer Science

- Where do I think the next amazing revolution is going to come? There's no question that digital biology is going to be it. For the very first time in our history, in human history, biology has the opportunity to be engineering, not science.

With advances in cell culturing techniques, bioengineering technologies, and AI, biological computing is poised to emerge as a viable and feasible solution to the requirements of AI training and deployment.

1. Part 4: Propose the solution to create biologics-based AI computing chips.
    1. Go into more depth on how they work and why they are more energy efficient

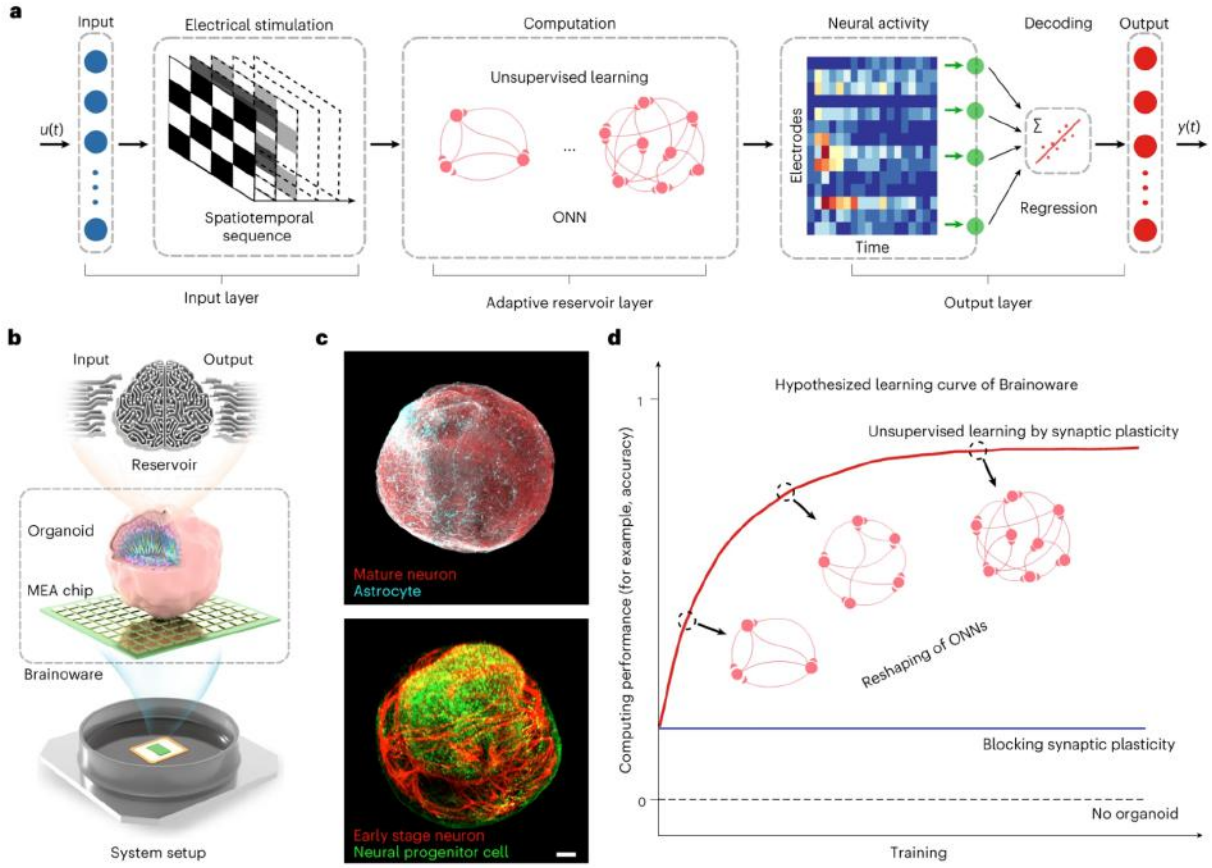
The core architecture of virtually every digital computer has followed a deceptively simple pattern first developed in the 1940s, known as the von Neumann model: store data in one place, do computation in another. Then, shuttle information between memory cells and the processor. Because these currently exist in two separate locations that are millimeters to centimeters apart so electricity needs to travel great distances to facilitate computation which makes it energy and time inefficient. This is called the von Neumann bottleneck or the memory-wall problem. However, each individual neuron within the human brain operates as *both* a memory cell and a processor. As the fusion of data storage and processes within biological neural networks bypasses this von Neumann bottleneck, one major promise of biologics-based computing is that in-memory computing will reduce the time and energy it costs to move and process large amounts of data. This "vertically heterogeneous-integrated architecture" is key to reducing energy consumption, according to Professor Deep Jariwala at the University of Pennsylvania.

The speed and energy efficiency of silicon-based computing hardware is approaching its theoretical limit - hindered by the slowing of Moore's law scaling and the von Neumann bottleneck, which increases the cost for big data movements.

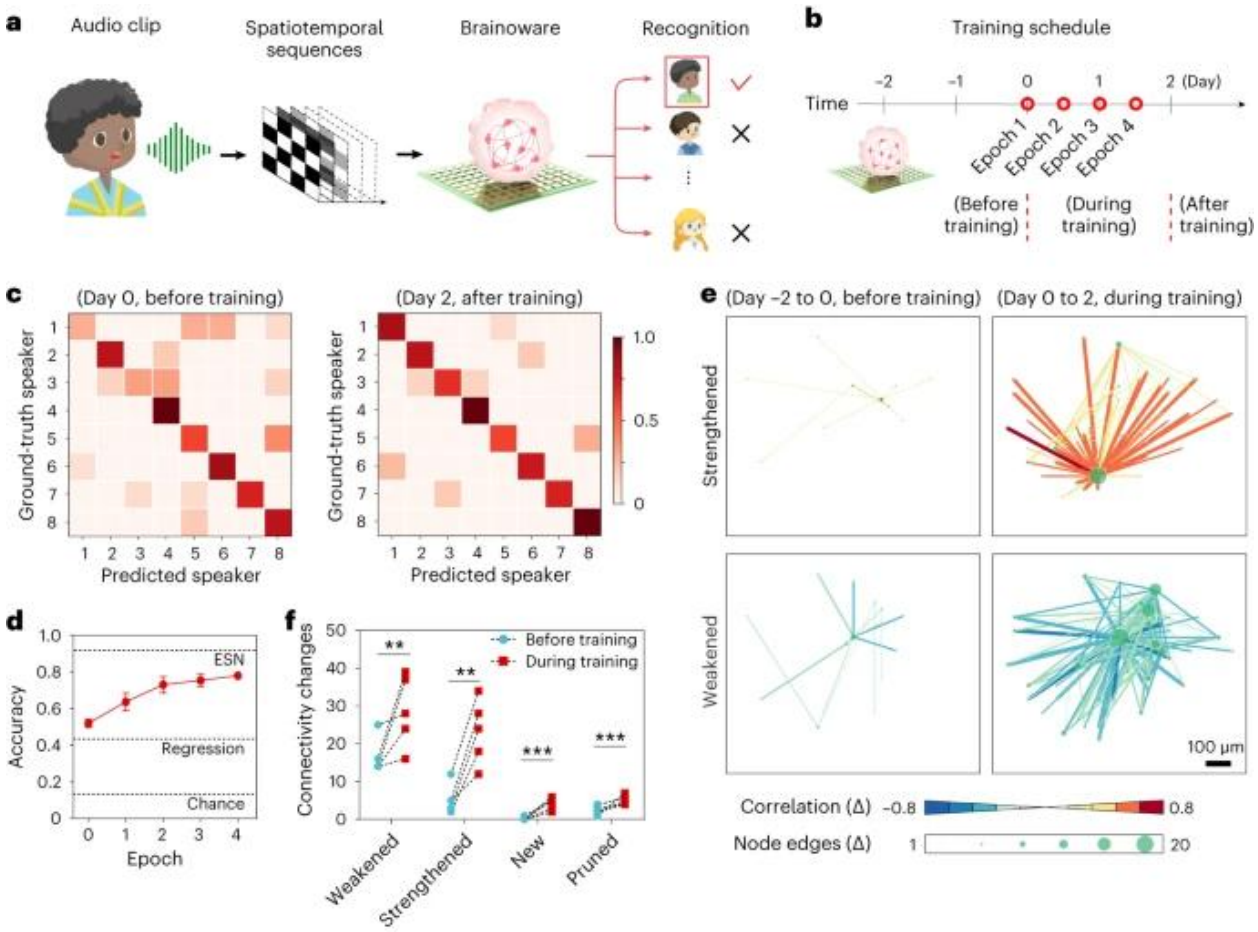
The human brain is a complex, three-dimensional massively parallel network of over 200 billion cells, linked by hundreds of trillions of synapses, and expends only 20 watts while current silicon hardware consumes about 8 million watts to power a comparative ANN model.

It is commonly said that “neurons that fire together, wire together.” On a deeper level, the human brain can learn because it is a living, adaptive mesh of cells that automatically conducts unsupervised learning due to the processes of neuroplasticity and neurogenesis.

Another major advantage of the brain is that it doesn’t produce heat. Therefore, thousands of watts of electricity used to cool silicon chips are no longer needed.



Already, current biologics-based chips have been used in computer vision and speech recognition tasks.



In this study, researchers converted audio clips into spatiotemporal sequences of bipolar pulse stimulations to the brain organoid. The evoked neural activity was recorded and fed into a logistic regression function for classification, then trained and optimized.

Nanoneuro Systems

Our Technology

We aim to grow brain organoids: 3D mini-brains in a dish through the self-organization of human iPSC neuronal cells. Then, by implanting them onto a custom-designed chip at the Singh Nanotechnology Center and sending inputs via external electrical stimulation through embedded shell electrodes and receiving outputs via evoked neural activity, we envision training future LLM models directly on the biology that have undergone 4 billion years of evolution to reach peak efficiency and processing power.

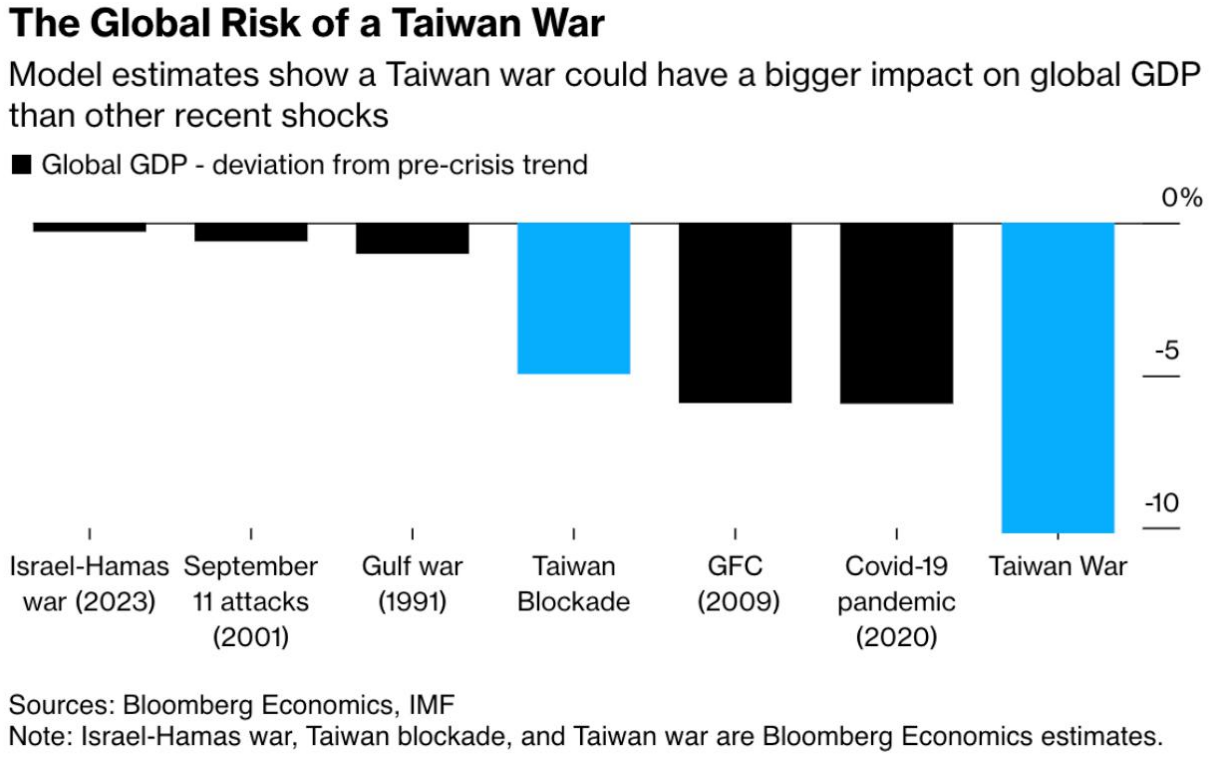
We are implementing new paradigms in ANNs – a Spiking Neural Network (SNNs). SNNs use neuron spiking events to communicate floating point numbers, and information is encoded in the timing and frequency of spikes through event-based processing. The goal is to approach one-shot learning similar to human learning experiences. For example, traditional deep learning models are successful when there is a large, labeled, and unnoisy dataset. But humans are adept at learning from just one or few instances of data. This is due not only to the

computational power of human brains, but also to the ability to synthesize and learn new object classes from limited information about different, previously learned classes. One-shot and few-shot learning aim to achieve a similar goal.

Our Vision

The backlogs to buy or lease Nvidia GPUs stretch over a year. Meta and Amazon have openly supported the notion to buy chips from other companies.

A big risk factor for the AI race is a possible war over Taiwan. The Taiwan Semiconductor Manufacturing Company manufactures 90% of all chips globally, and if disrupted, will impact an estimated 10% of the global GDP.



That's why there is a huge investment in domestic chip manufacturing companies; the US CHIPS Act alone provided $52B to accelerate domestic chip developments.

Conclusion

- Summarize everything

That's not to say AI and advancing it needs to stop because it's incredibly useful for important applications like accelerating the discovery of therapeutics. We just need to remain cognizant of

the effects and keep pushing for more sustainable approaches to design, manufacturing, and consumption.

Bibliography

1. https://penntoday.upenn.edu/news/hidden-costs-ai-impending-energy-and-resource-strain
2. https://www.nature.com/articles/s41928-023-01069-w
3. https://www.nature.com/articles/s41928-023-01096-7
4. https://www.theregister.com/2023/12/26/thermal_management_is_changing/
5. https://www.scientificamerican.com/article/ais-climate-impact-goes-beyond-its-emissions/
6. https://www.governance.ai/post/computing-power-and-the-governance-of-ai
7. https://a16z.com/navigating-the-high-cost-of-ai-compute/
8. https://onlinelibrary.wiley.com/doi/10.1002/adma.201902761
9. https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/
10. https://www.nature.com/articles/s42256-020-0219-9
11. https://transmitter.ieee.org/how-big-will-ai-models-get/
12. https://ourworldindata.org/grapher/artificial-intelligence-parameter-count#reuse-this-work

Logical Outline

Premise

- Given that AI has undergone recent explosive growth in the past 5 years
- Given that AI development shows no sign of slowing down and is actually accelerating
- Given that AI research requires computers (GPUs) to train, which requires energy
- Climate change is a major problem which is mainly caused by carbon emissions from human activities

Proposition

- Thus, AI training, deployment, and research needs to be made significantly more energy efficient to align with climate change goals.

Reasons & Evidence

1. (How) Current silicon chip design should be abandoned in favor of newer technologies
    a. Silicon chips are not inherently energy efficient due to the memory gap problem
    b. Silicon is unsustainable as the world enters into a silicon shortage
    c. Silicon chips requires massively amounts of energy in order to cool
    d. Silicon technology is reaching the limits of physics and can no longer be advanced at a rate suggested by Moore's Law
2. (How) Focusing on massive clean-energy supplying technologies is a long-term goal that will make AI training more climate friendly
    a. Developments in solar and nuclear technologies are decades away from feasible scalability
    b. Clean energy from solar and nuclear technologies will be a long-term solution for supplying clean energy for data centers
    c. Clean energy from solar and nuclear technologies is not favored or implemented by companies currently due to high upfront costs and low efficiencies and a lack of pressure to reduce their climate impact
3. (How) Biologics based chips should be developed for AI specific training and deployment
    a. Biologics-based chips are inherently 10000x more efficient than silicon chip
    b. Biologics-based chips uses less silicon and is thus sustainable in the long term future
    c. Biologics-based chips do not generate heat and thus use less energy
    d. Biologics-based chips have undergone 4 billion years of evolution, resulting in both energy efficiencies and fast processing power

Rhetorical Outline

- Proposition: AI training, deployment, and research needs to be made significantly more energy efficient to align with climate change goals.
- Audience: Chip companies, deeptech venture capitalists
- Genre: White paper
- Motive of the Author: To promote biologics-based computing for investment and R&D
- Motive of the Reader: To recognize the climate problems caused by AI & to fund biologics-based computing and support climate tech solutions in the long-term
- Plan: Publish as a company white paper research (similar to Bitcoin's whitepaper: https://bitcoin.org/bitcoin.pdf), publish in TechCrunch or related site
- Rhetorical Strategies: No idea what this means!
- Keywords: AI, climate, energy, biologics-based computing, biocomputing, neuromorphic computing