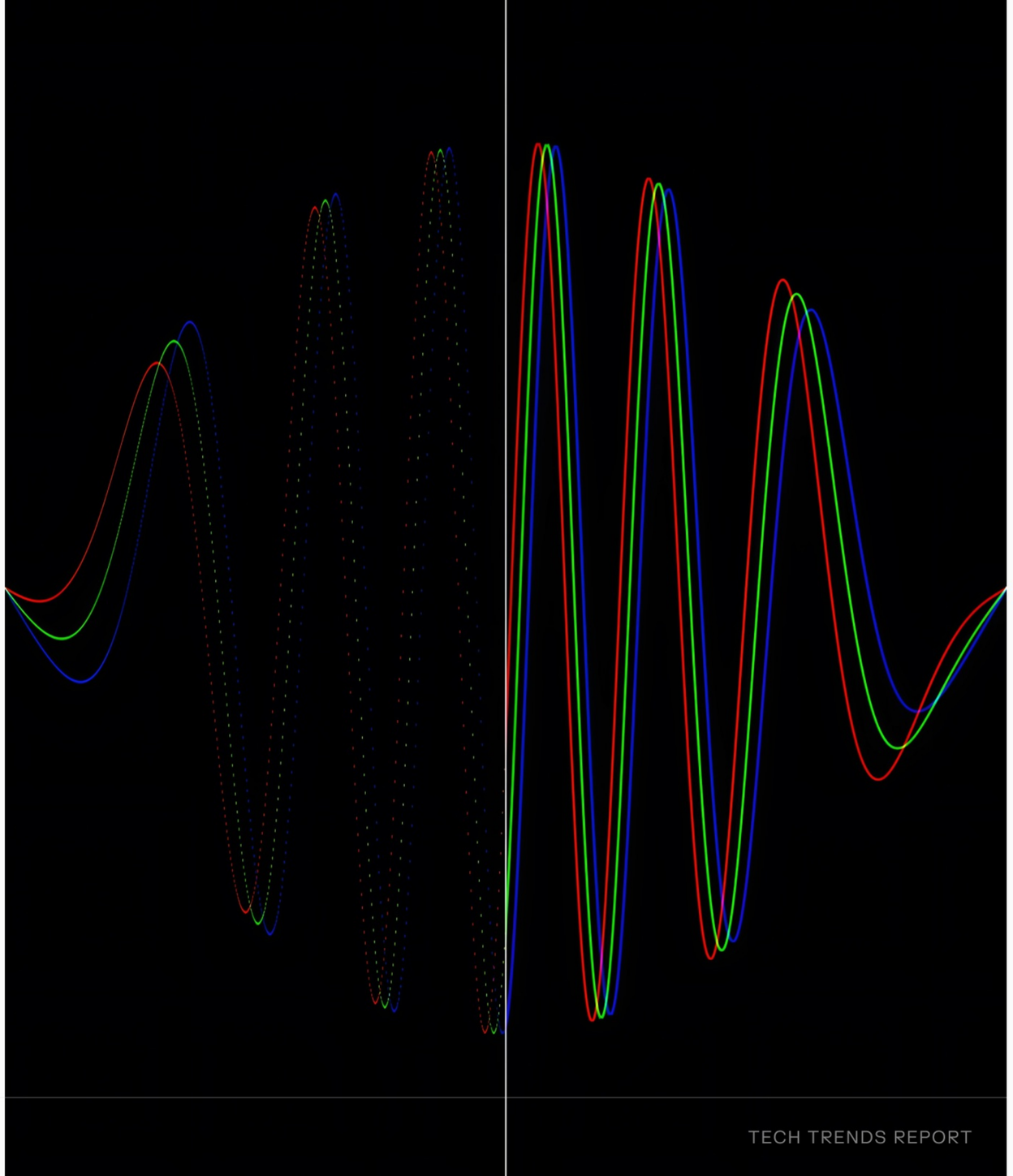
Artificial Intelligence

- Foundation Models
- Frontier Applications
- Scaling

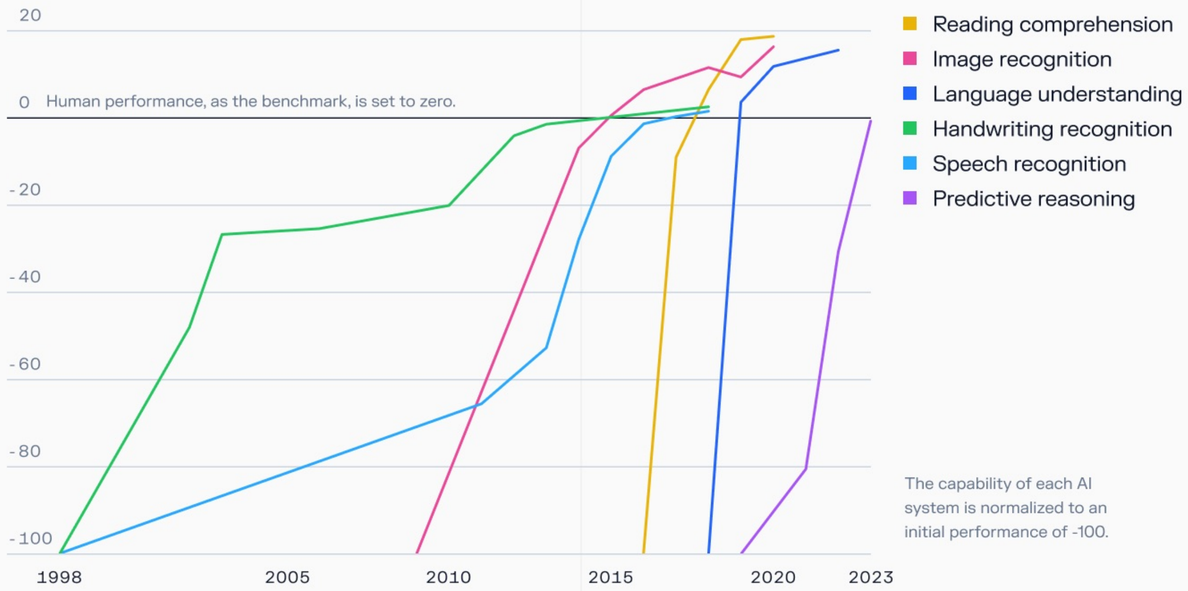
I. ARTIFICIAL INTELLIGENCE

Foundation Models



Large language models, based on the transformer architecture, are the first AI models to understand language, our medium for encoding all human knowledge. The size, performance and wide applicability of these models have led researchers to begin naming them *foundation models.*

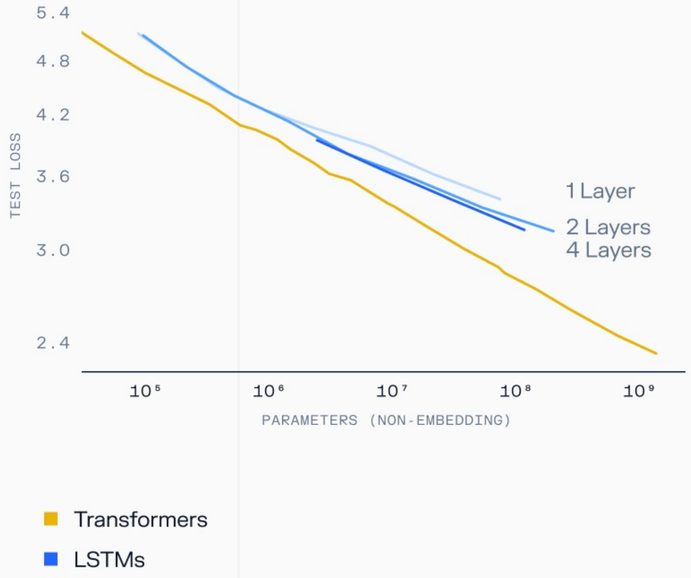
Test scores of AI systems on various capabilities relative to human performance



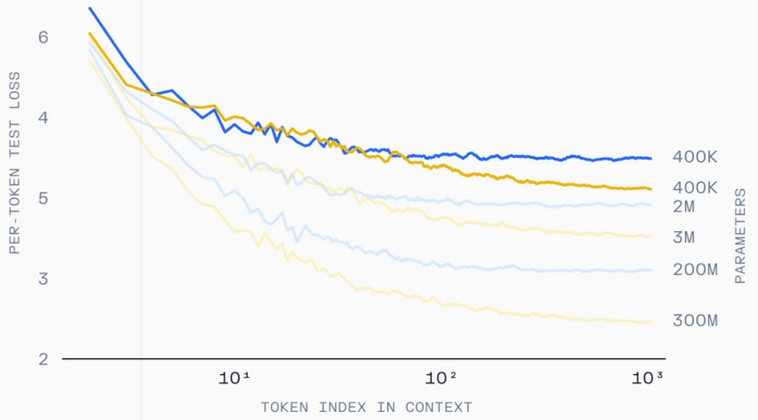
Source: Our World in Data, Kiela et al. (2023). Note: For each capability, the first year always shows a baseline of –100, even if better performance was recorded later that year.

The transformer, a neural network architecture developed in 2017 capable of building contextual awareness as it processed text, far outperformed all other architectures when it came to learning language.

Transformers asymptotically outperform LSTMs due to improved use of long contexts.



LSTM plateaus after <100 tokens.
Transformer improves through the whole context.

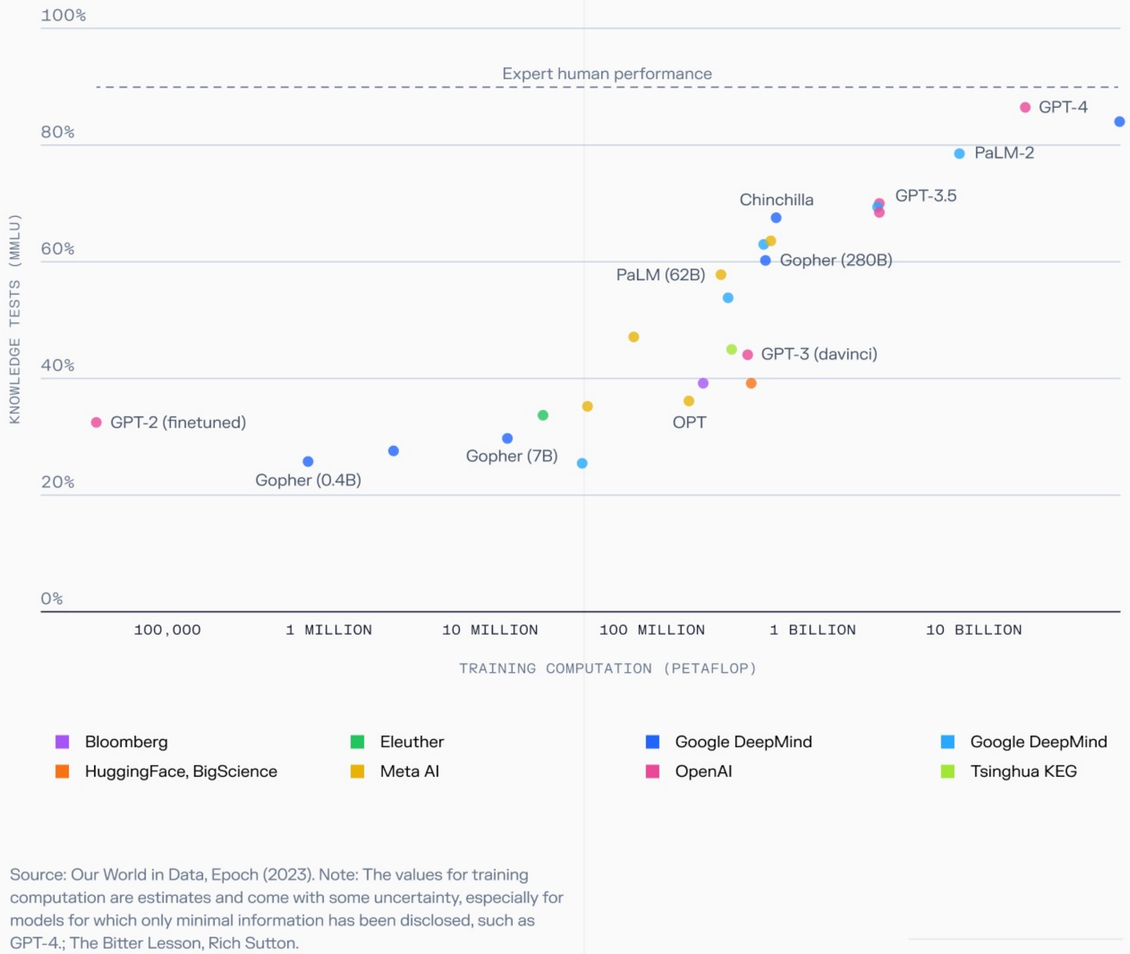


Transformers
LSTMs

Source: Kaplan, Jared et al. "Scaling Laws for Neural Language Models." *ArXiv* abs/2001.08361 (2020).

And yet, the “bitter lesson” of AI research, as articulated by Rich Sutton, is that regardless of architecture, “general methods that leverage computation are ultimately the most effective, and by a large margin.”

Artificial intelligence: Performance on knowledge tests vs. training computation



Source: Our World in Data, Epoch (2023). Note: The values for training computation are estimates and come with some uncertainty, especially for models for which only minimal information has been disclosed, such as GPT-4.; The Bitter Lesson, Rich Sutton.

As a result, models have grown massively in size, increasing by four orders of magnitude between 2018 and 2022.

Number of parameters of notable machine learning models by sector, 2003–23

Source: Artificial Intelligence Index Report 2024, Stanford HAI, Epoch (2023); 4OOM Quote Epoch (2022).

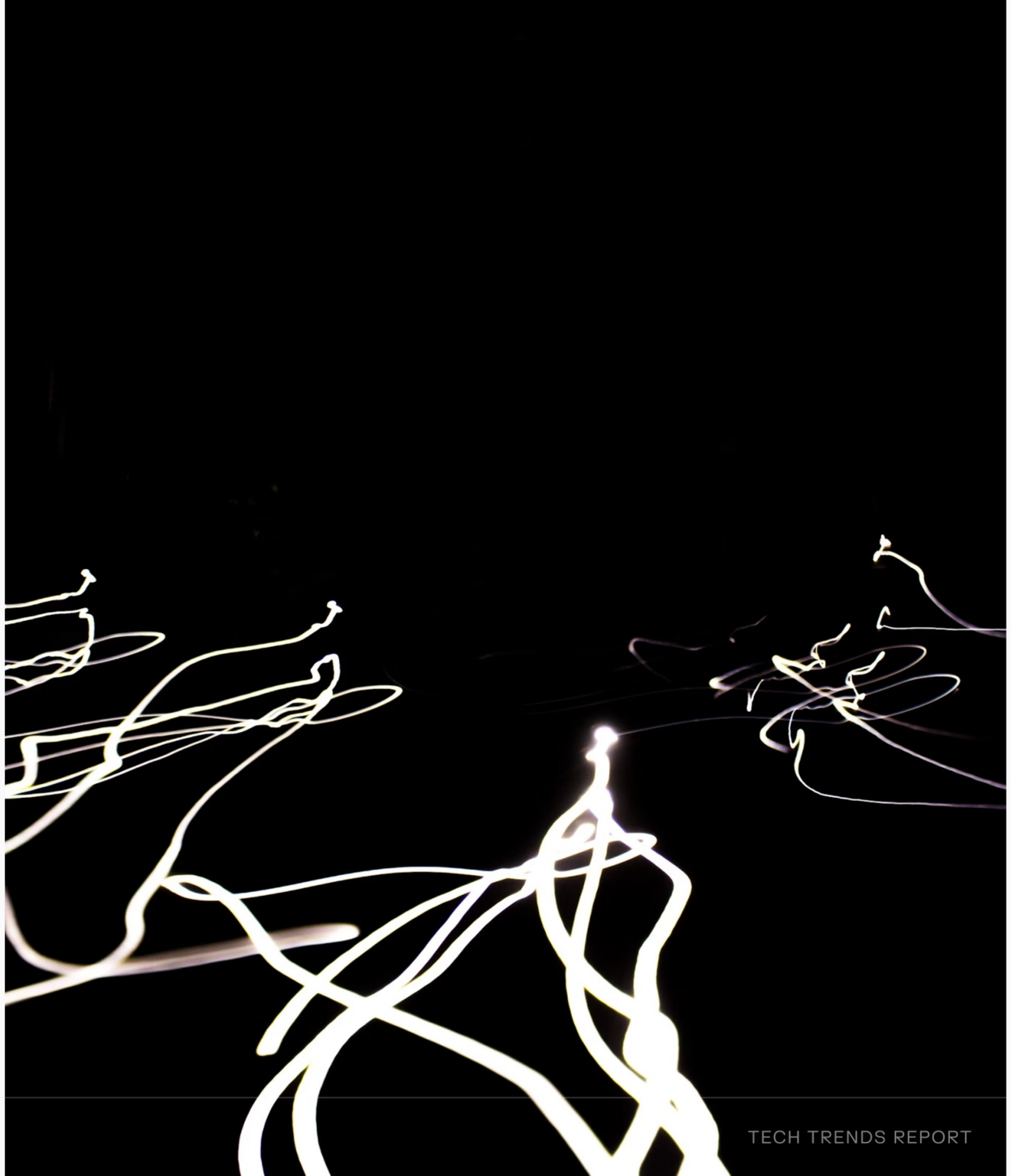
As models got larger, not only did their performance improve, but they began exhibiting emergent behaviors – abilities they were not taught explicitly.

Source: Ganguli, Deep et al. "Predictability and Surprise in Large Generative Models."
Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (2022).

I. ARTIFICIAL INTELLIGENCE

Frontier Applications

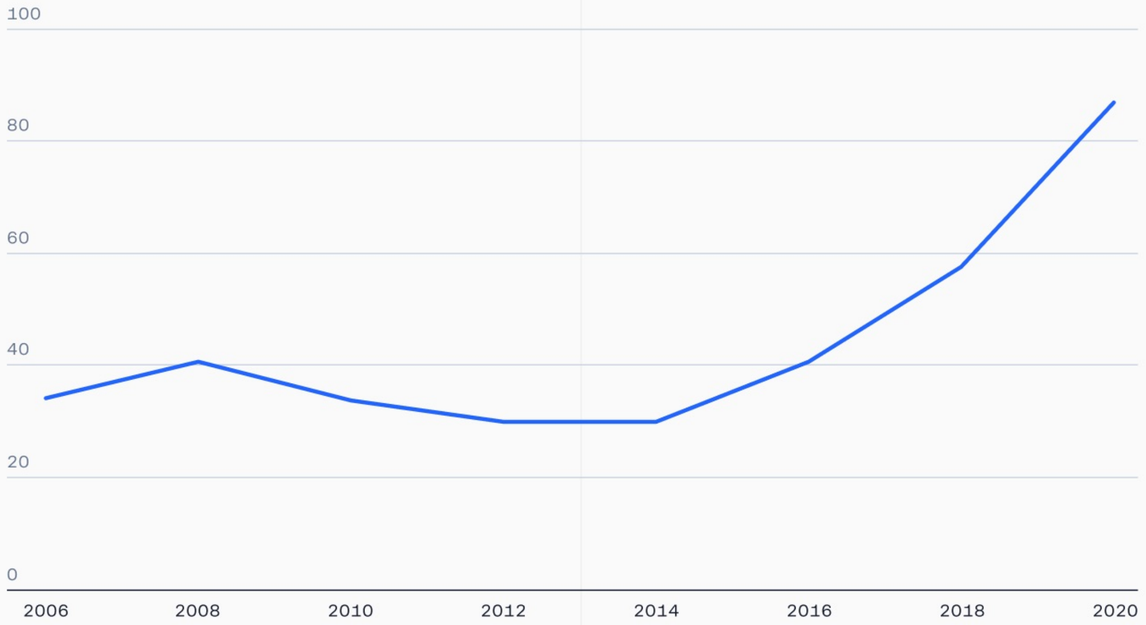
- Foundation Models
- Frontier Applications
- Scaling



Large models are increasingly capable of understanding more than just language, but the patterns underlying a variety of complex domains. In 2021, DeepMind's AlphaFold2 is considered to have solved the protein folding problem.

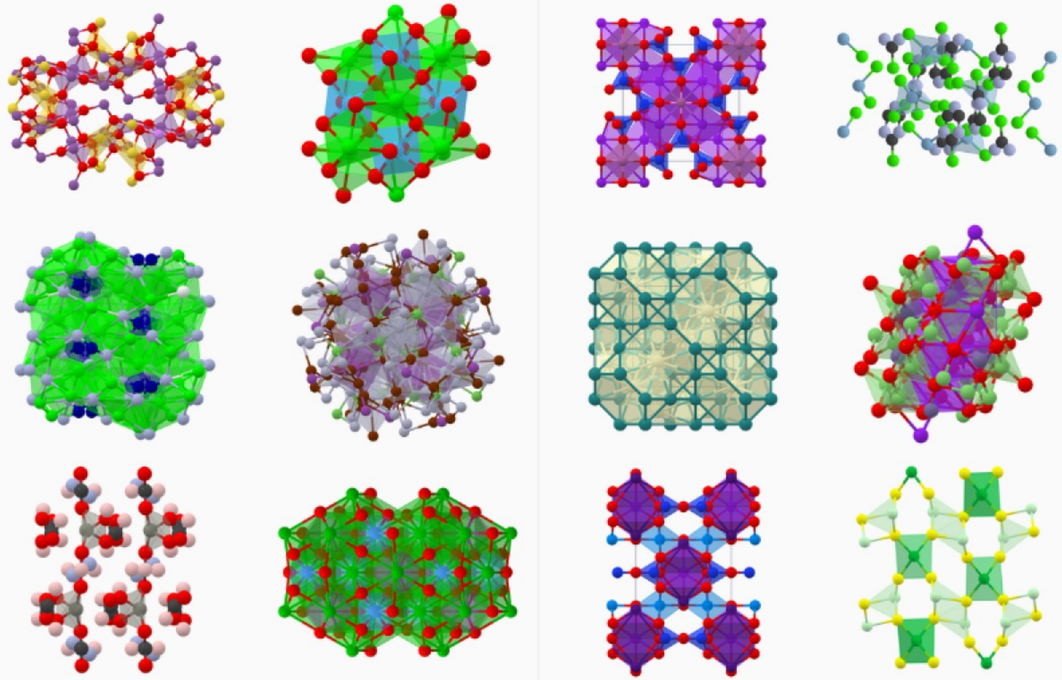
Protein folding prediction accuracy

Median accuracy of predictions in the free modeling category for the best team in each year's Critical Assessment of Protein Structure (CASP) competition. Ranges from 0–100 (100 = best).



Source: Our World in Data, AI Index Report 2021 Note: Measured using best-of-5 Global Distance Test, which scores the similarity between a predicted protein structure and the actual structure.

Recently, new models have also pushed forward our understanding of chemistry. In 2023 DeepMind's GNoME tool produced 2.2 million crystal structures that didn't exist before, including 380K that are predicted to be stable and usable in future technologies.



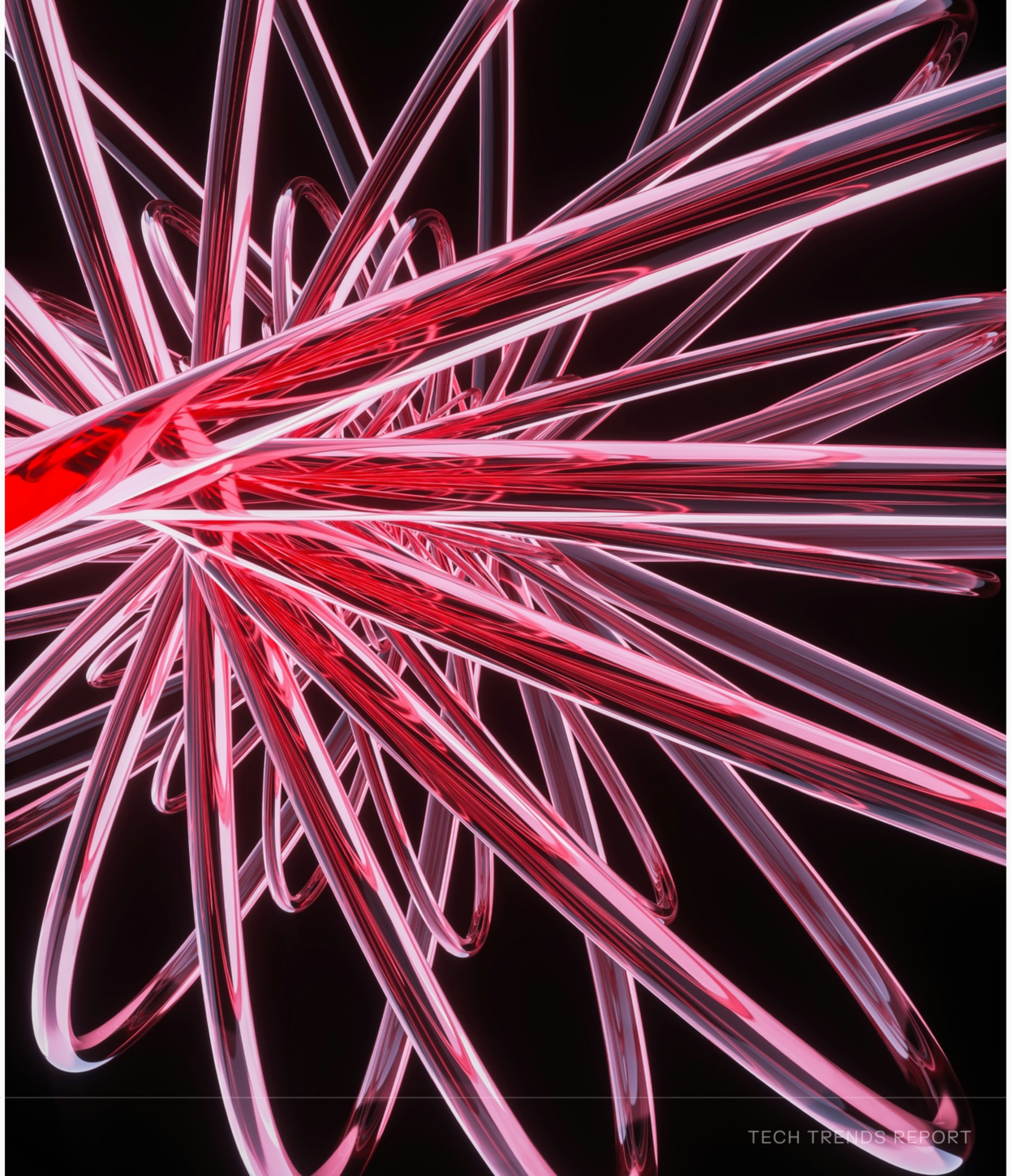
Source: Jenny Nuss/Berkeley Lab Note: Google DeepMind developed a deep learning tool called Graph Networks for Materials Exploration, or GNoME. Researchers trained GNoME using workflows and data that were developed over a decade by the Materials Project, and improved the GNoME algorithm through active learning. GNoME researchers ultimately produced 2.2 million crystal structures, including 380,000 that they are adding to the Materials Project and predict are stable, making them potentially useful in future technologies.

Large language models are allowing robots to build intuition around structuring and automating tasks, as demonstrated by Google's SayCan algorithm here.

Source: Ahn, Michael et al. "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances." *Conference on Robot Learning* (2022).

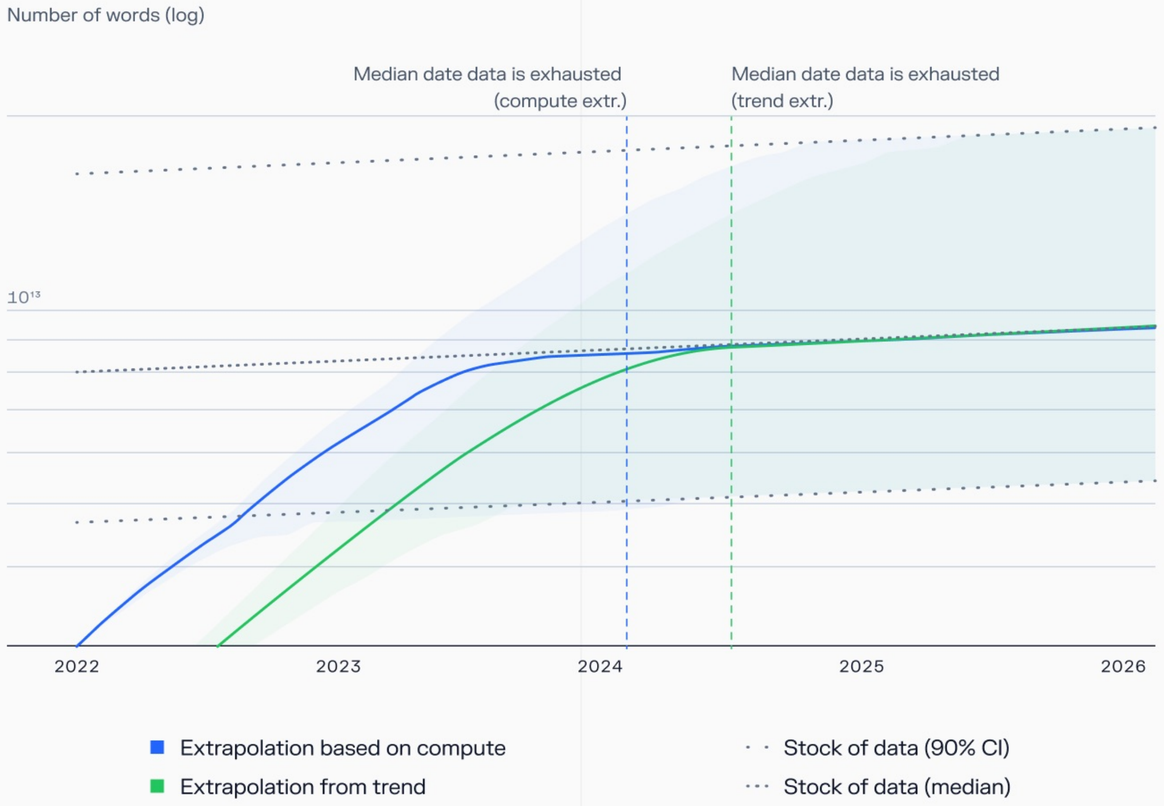
I. ARTIFICIAL INTELLIGENCE

Scaling



However, there are a few major headwinds impeding the continued exponential scaling of large language foundation models. The first is the decreasing availability of high-quality language data.

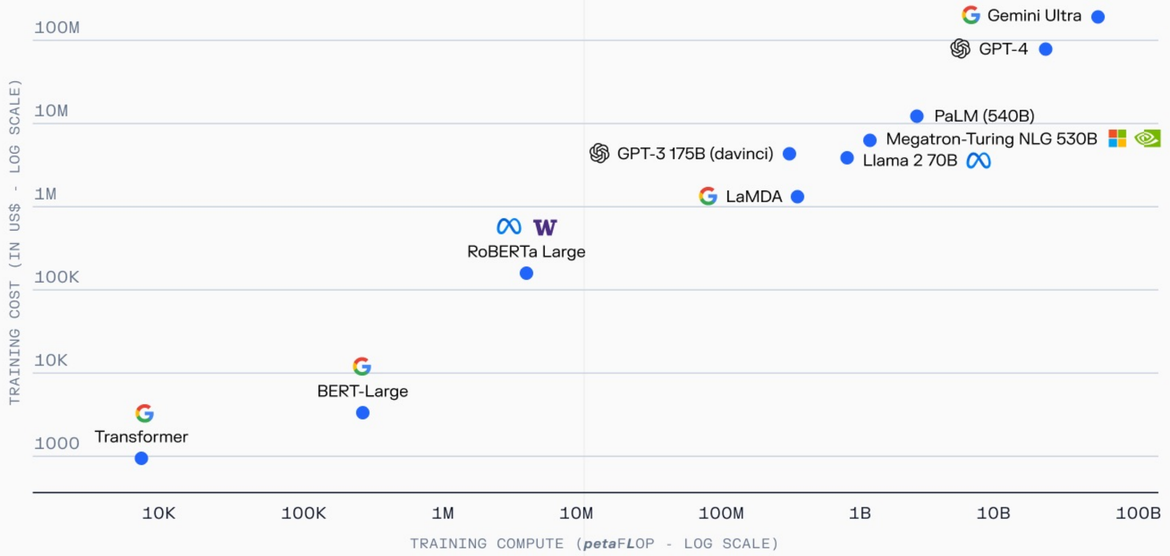
Projections of data usage (for high-quality language data)



Source: Epoch (2022), Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 'Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning'. *ArXiv [Cs.LG]*, 2022. arXiv. http://arxiv.org/abs/2211.04325.

The second is the growing cost of training larger models due to the increasing scale of compute required.

Estimated training cost and compute of select AI models

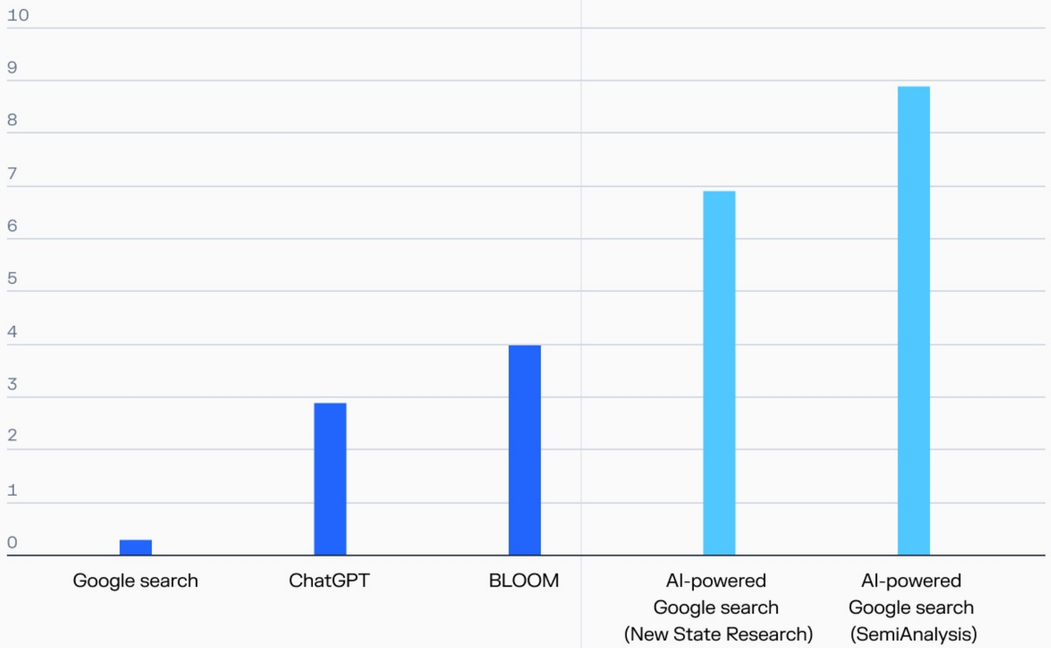


Source: Artificial Intelligence Index Report 2024, Stanford HAI, Epoch (2023).

Third is the increasing amount of energy required to run ever larger models. Already, AI could be on track to consume as much electricity as all of Ireland, i.e. 29.3 terawatt-hours per year.

Estimated energy consumption per request for various AI-powered systems

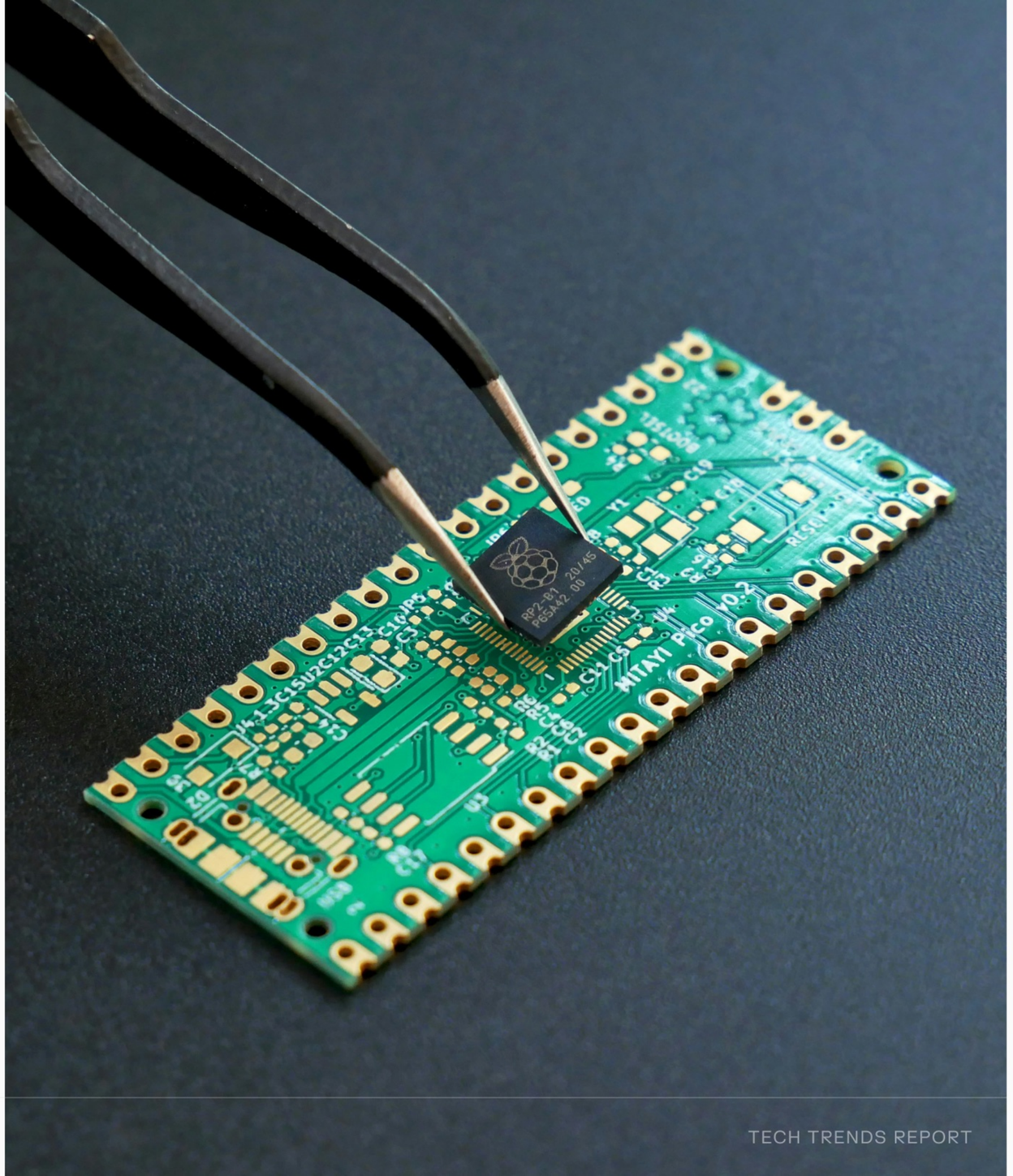
Wh per request, in comparison to a standard Google search



Source: de Vries, The growing energy footprint of artificial intelligence, Joule (2023).

Compute

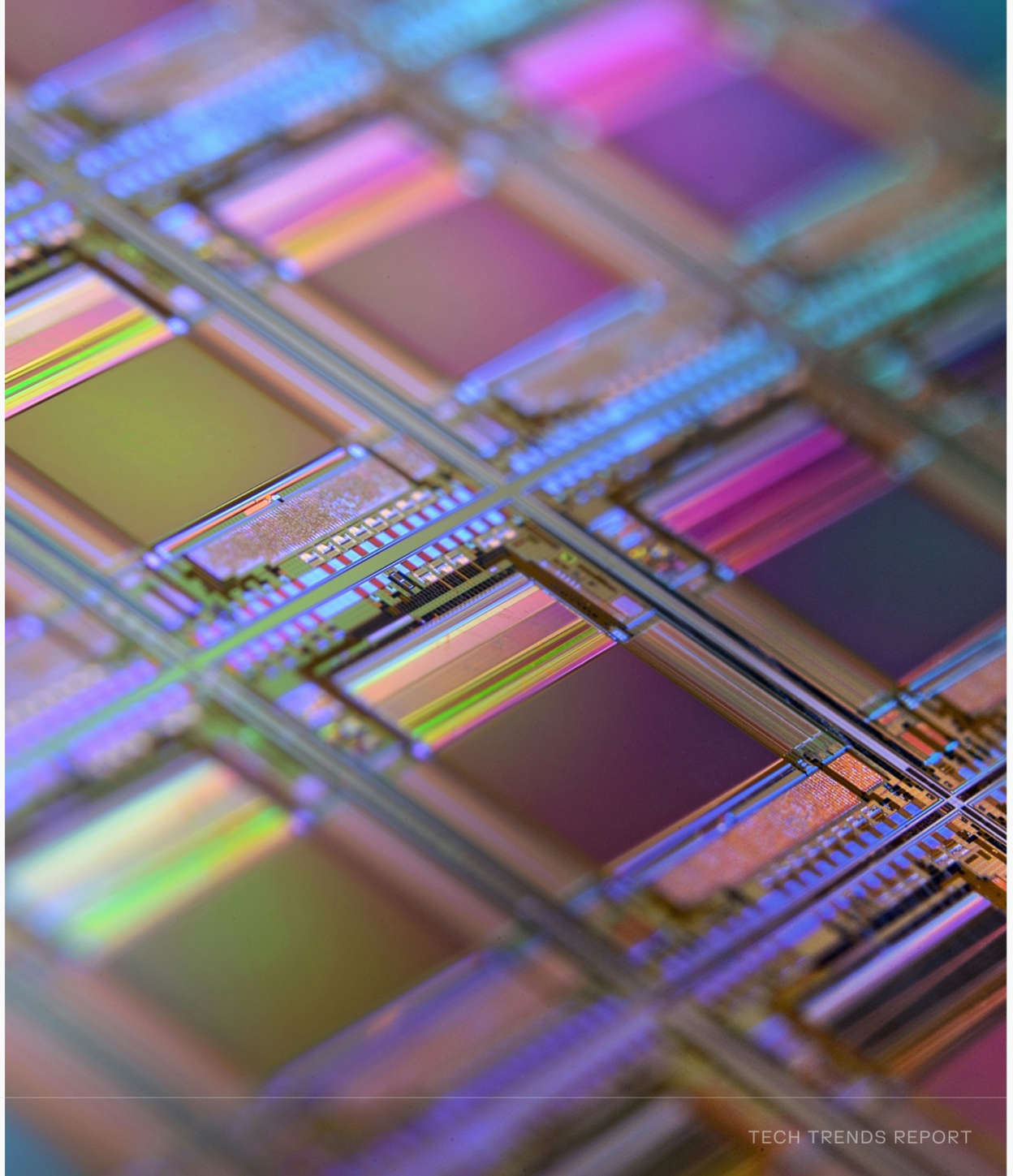
- Moore's Law
- GPUs



II. COMPUTE

Moore’s Law

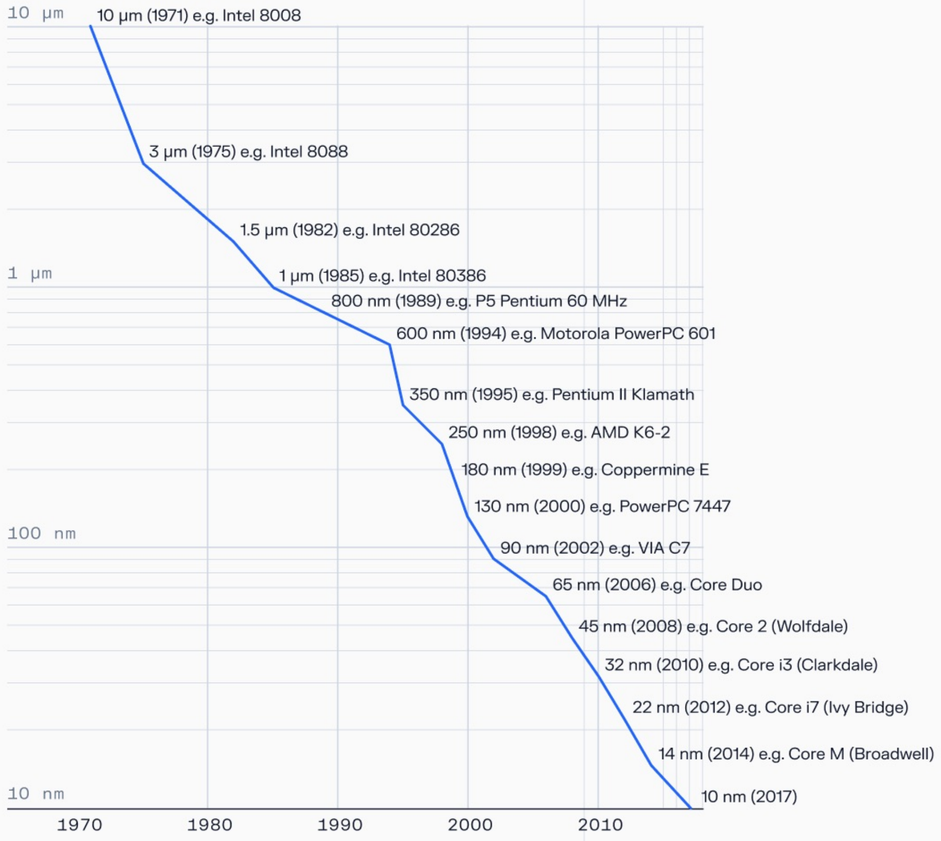
- Moore’s Law
- GPUs



The story of improving computation is a story of miniaturization.

Progress of miniaturisation

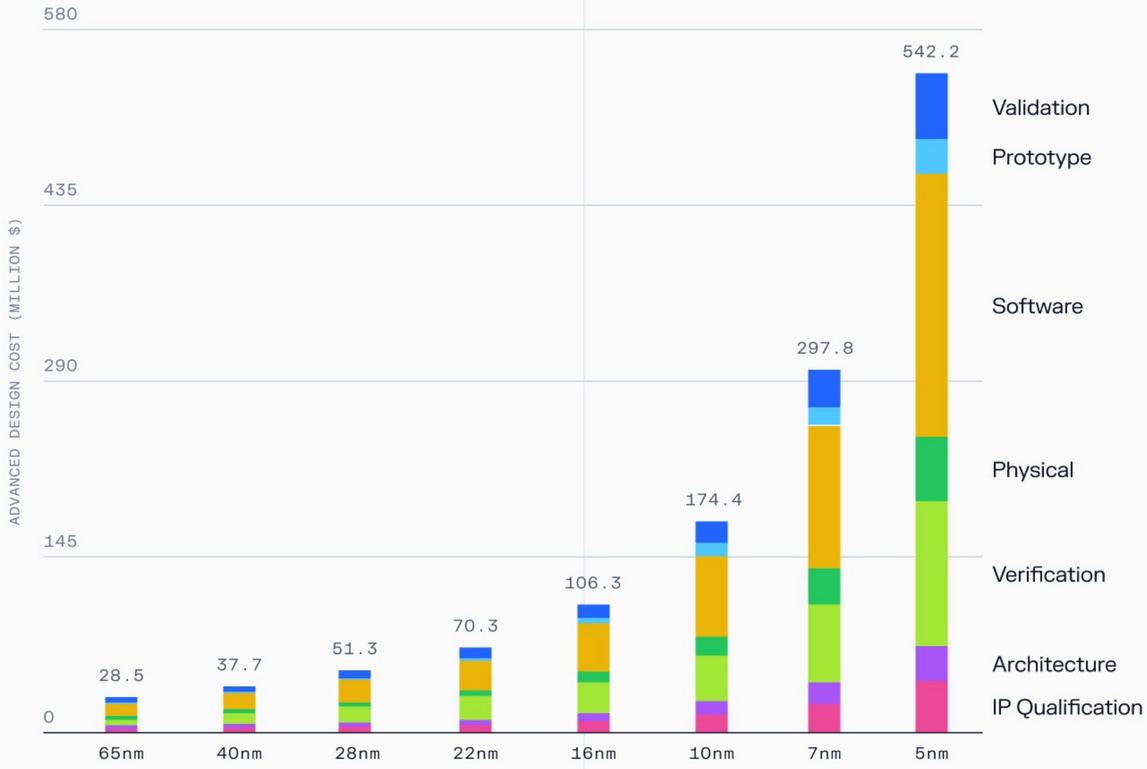
With comparison of sizes of semiconductor manufacturing process nodes with some microscopic objects and visible light wavelengths



Source: Progress of miniaturisation, and comparison of sizes of semiconductor manufacturing process nodes with some microscopic objects and visible light wavelengths; Cmglee, CC BY-SA 3.0, via Wikimedia Common.

However, continued miniaturization is increasingly costly.

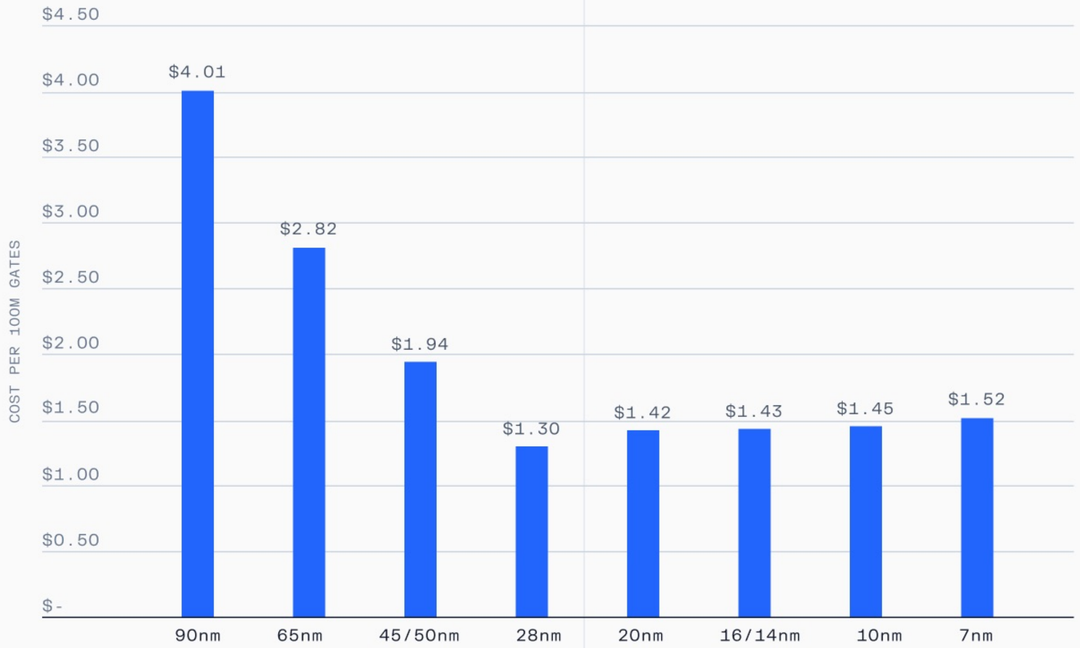
Costs of developing smaller process technologies



Source: F. Martin, European Space Agency, October 2023, using Handel Jones, Cost of Advanced Designs, IBS Jul' 17.

As a result, the cost per transistor has stopped falling. In fact, it has begun to gradually increase. "Moore's Law is dead," said Jensen Huang in 2022, adding "the idea that the chip is going to go down in price is a story of the past."

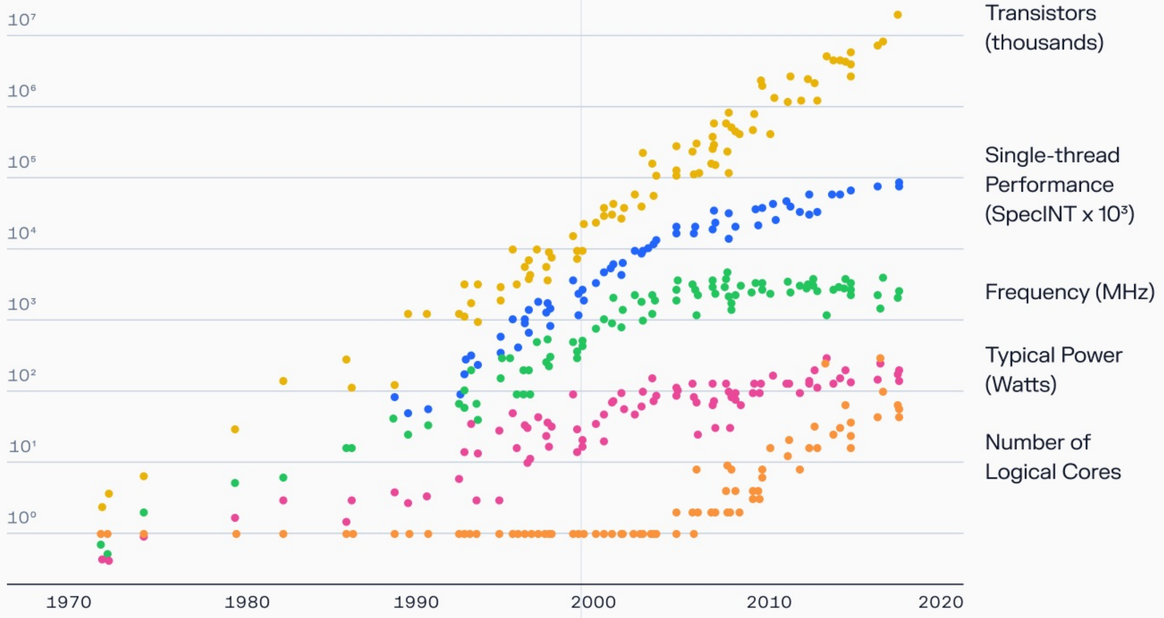
Gate cost trend



Source: Marvell 2020 Investor Day Presentation, via Doug O'Laughlin; Monica J. White, Digital Trends (2022).

As miniaturization reached its limits, everything from clock speeds to energy efficiency and single-thread performance have stalled. Chip architects began adding more cores to improve CPU performance.

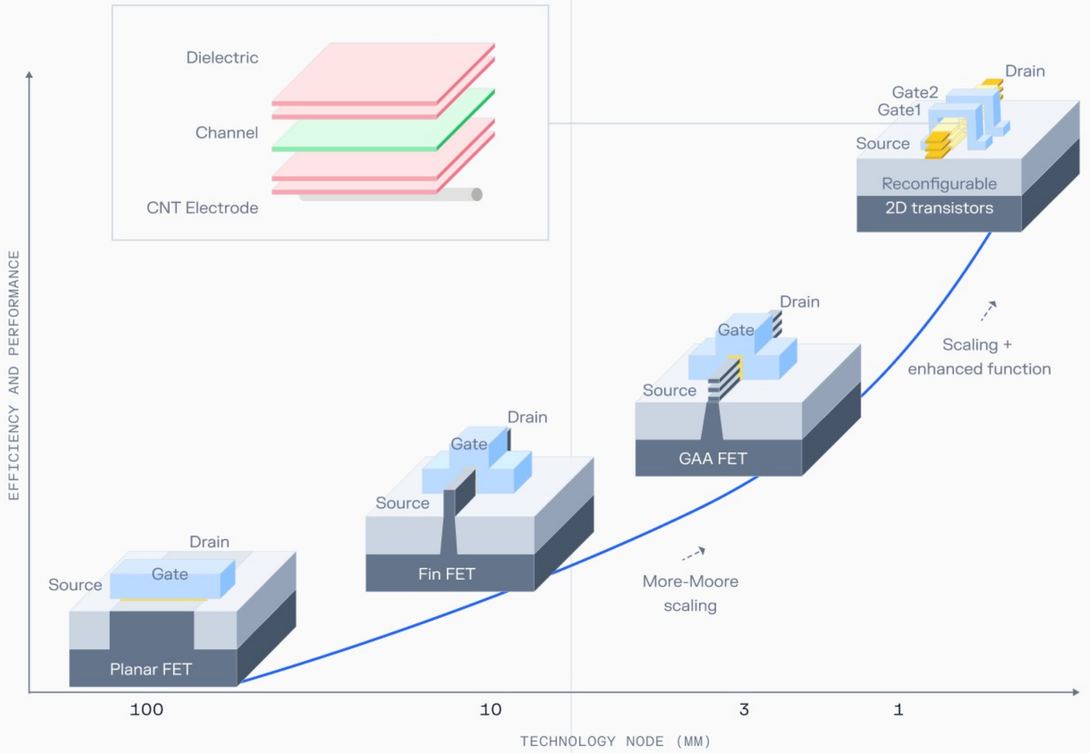
42 years of microprocessor trend data



Source: Karl Rupp Note: New plot and data collected for 2010-2017 by K. Rupp, Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten.

There are still immense efficiency benefits from keeping computation as physically close together as possible. This is why, despite the growing costs, fabs continue to research how to scale down transistors further — with the goal now at 1nm.

Emerging designs for smaller transistors



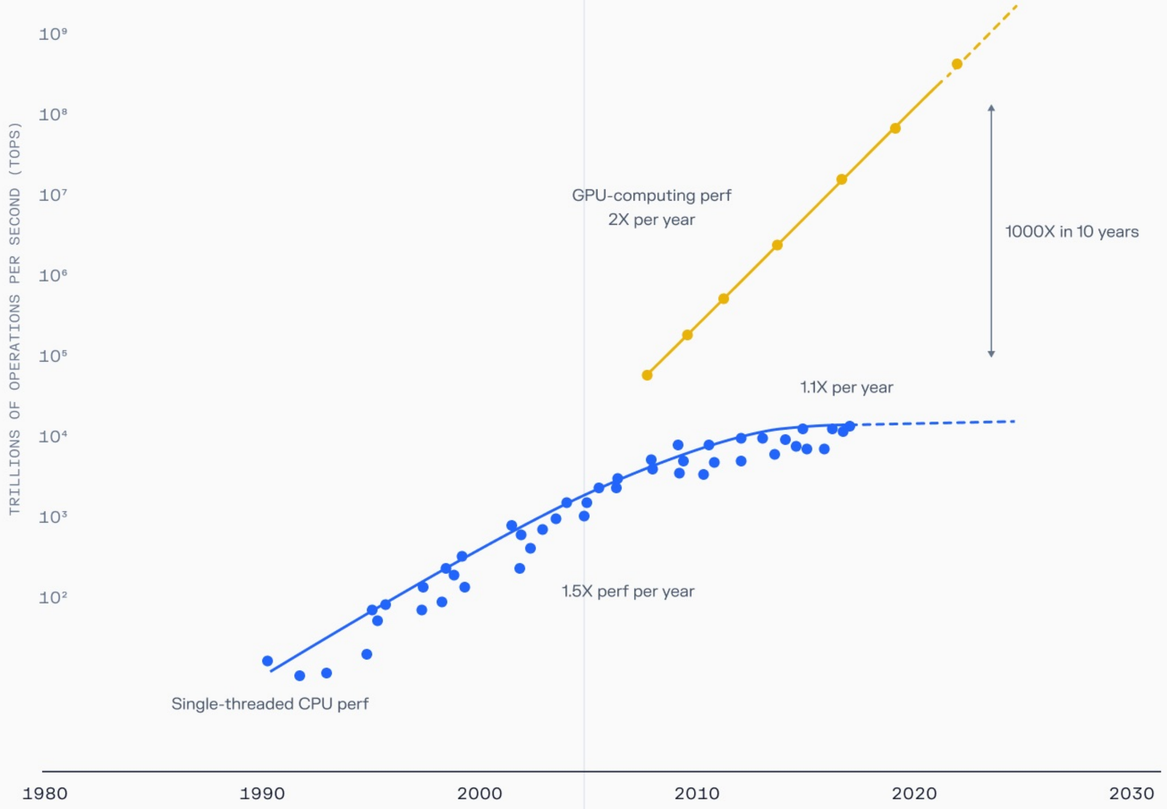
Source: Fei, Wenwen & Trommer, Jens & Lemme, Max & Mikolajick, Thomas & Heinzig, André (2022).

II. COMPUTE

GPUs

- Moore's Law
- GPUs

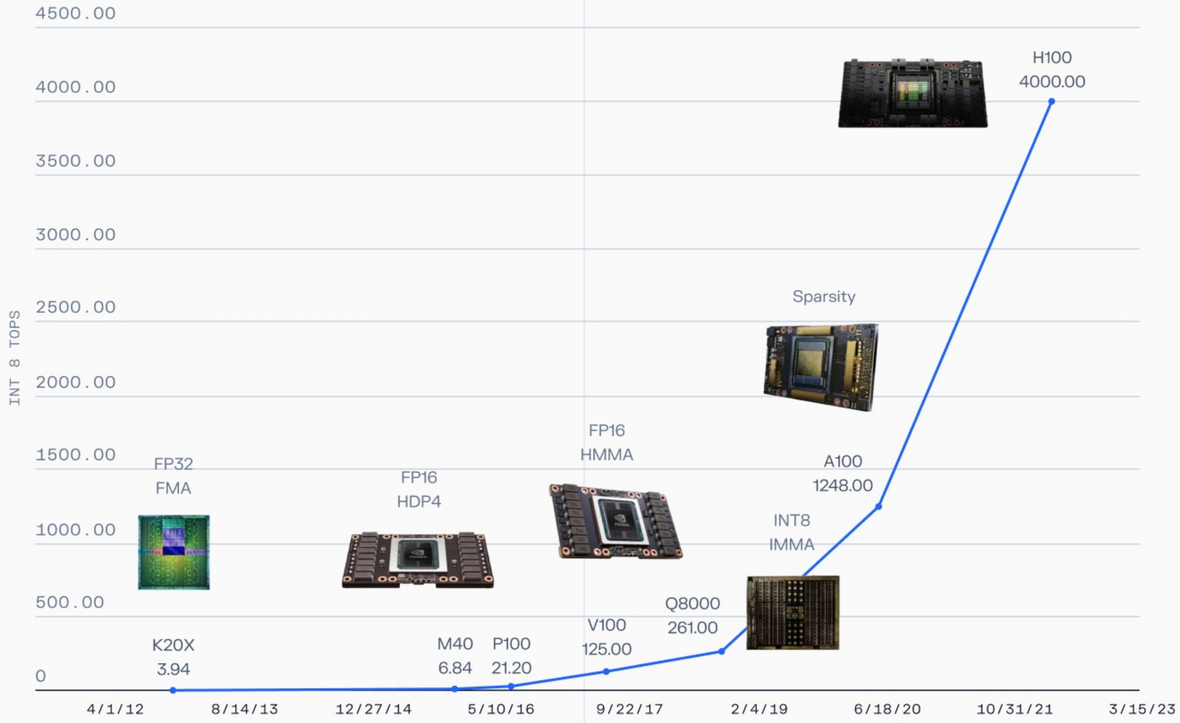
For computation that can be processed in parallel, like matrix multiplication required for AI training, GPUs have delivered a 1000x increase in performance over single-threaded CPUs.



Source: NVIDIA, 2023 Investor Presentation.

Without tailwinds from miniaturization, the industry has looked elsewhere to find improvements in performance, like improved architecture and larger processors.

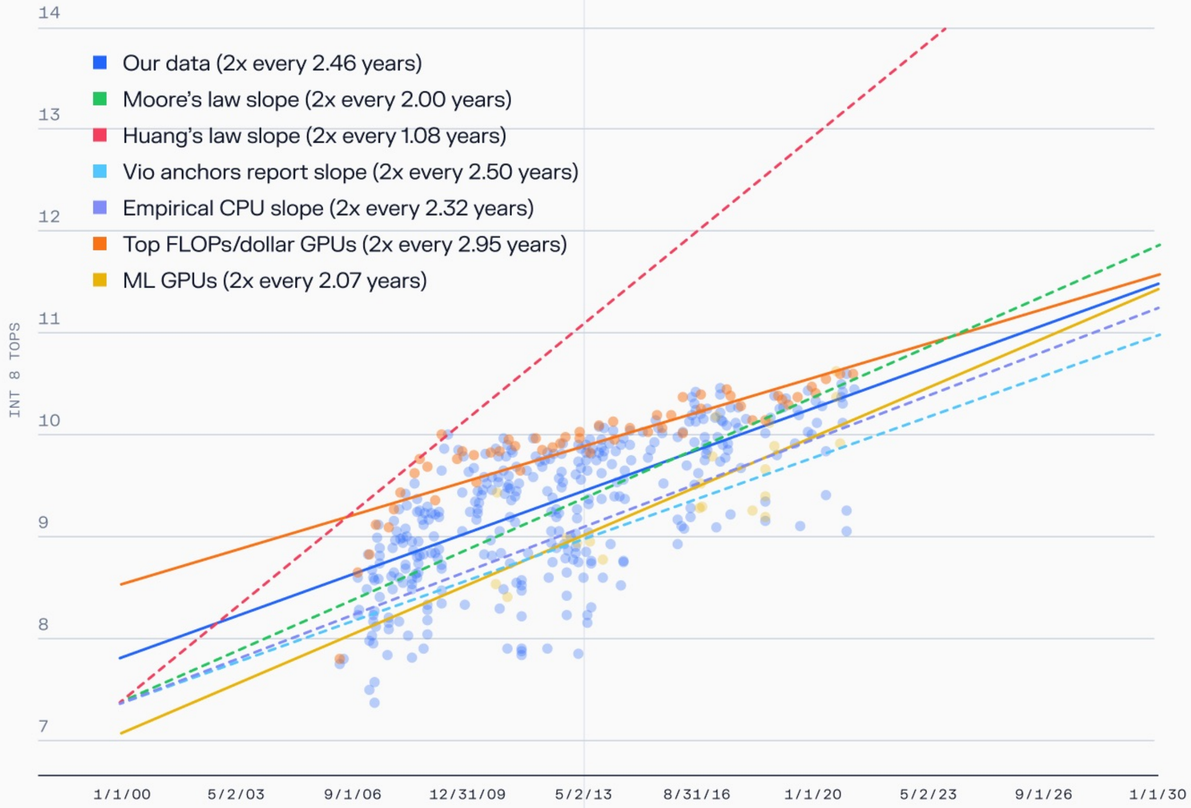
Single-chip inference performance - 1000X in 10 years



Source: NVIDIA, NVIDIA Chief Scientist Bill Dally's Keynote at Hot Chips August 2023.

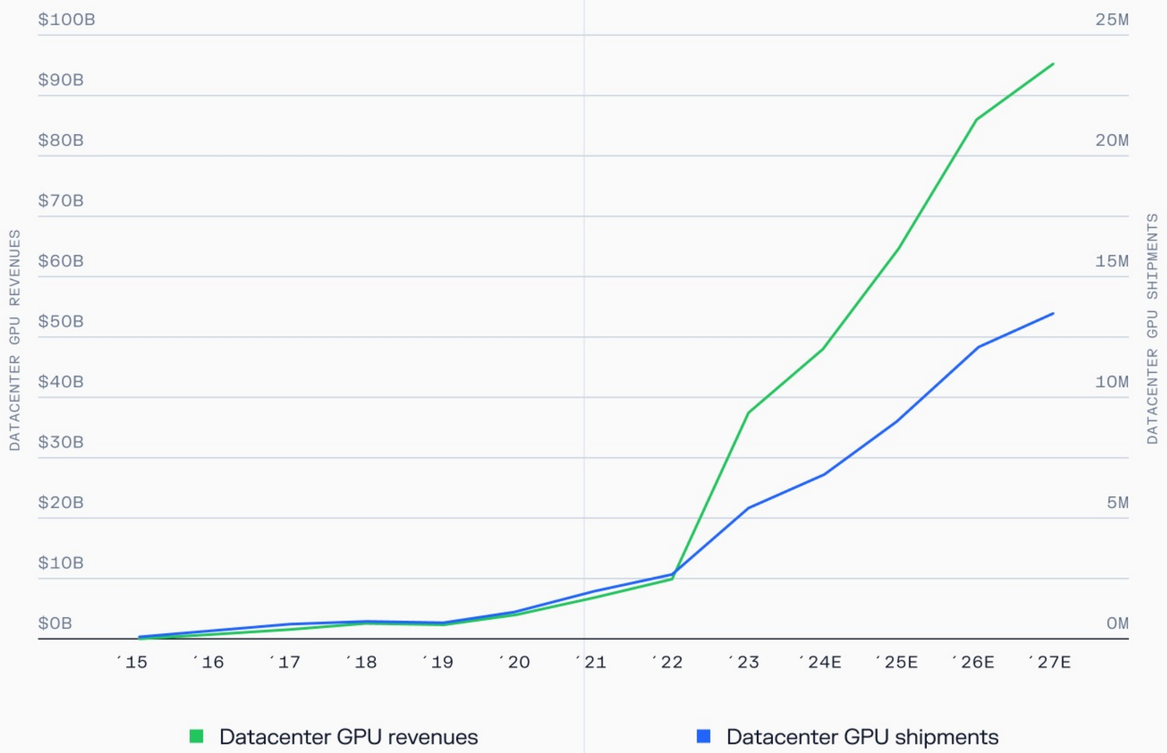
The rate of GPU performance improvement is now on par with Moore's Law, doubling in performance every two years.

Empirical GPU FLOP/s per dollar



Source: Marius Hobbhahn and Tamay Besiroglu (2022), "Trends in GPU Price-Performance".

To supply the needs of AI, data center GPU shipments are expected to nearly triple over the next five years. While Moore's Law might be over, the era of GPUs has only just begun.



Source: Aaron Rackers at Wells Fargo Equity Research via Next Platform.