



Predictable Scaling and Infrastructure

Trevor Cai

Predictable Scaling

The What and How of ChatGPT

- What

- Collect a dataset of text, code, images, audio, math.
- Pre-train a model to predict the next word.
- Post-train it to e.g. follow instructions, be conversational, use tools.

- How

- Synchronous SGD of a transformer on a large cluster of accelerators.
- Combine many forms of data and model parallelism.
- Reinforcement learning from human feedback.

The Sentiment Neuron (2017)

- Trained a character-level neural network on product reviews.
- Observed: there is a neuron encoding sentiment!
- Results in state-of-the-art sentiment analysis.

Team Spirit is maybe made by the best intentions, but it misses the warmth of "All Stars" (1997) by Jean van de Velde. Most scenes are identical, just not that funny and not that well done. The actors repeat the same lines as in "All Stars" but without much feeling.

Prediction is Compression

- Next word prediction models the underlying generative process.
- If the data is the internet, the underlying generative process is the world.
- Grandiose. But theoretically justified via Solomonoff Inductive Inference.

Team Spirit is maybe made by the best intentions, but it misses the warmth of "All Stars" (1997) by Jean van de Velde. Most scenes are identical, just not that funny and not that well done. The actors repeat the same lines as in "All Stars" but without much feeling.

Returns to Scale (2018-2023)

	Release	New Behaviors
GPT-1	June 2018	State-of-the-art language understanding (using task-specific fine tuning).
GPT-2	Feb 2019	Coherent text generation and zero-shot transfer.
GPT-3	Mar 2020	In-context learning.
GPT-4	Mar 2023	Actually being useful.

Scale Works (2023)

- This example required GPT-4 to:
 - Understand both English and French.
 - Interpret a diagram in context of the text.
 - Solve a physics problem!

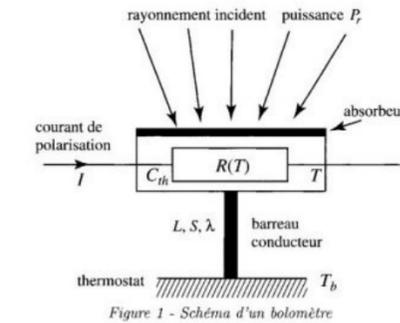
GPT-4 visual input example, École Polytechnique Exam Question:

User Answer question I.1.a. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R} \frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th} .

Un barreau, conducteur thermique, homogène, de longueur L , de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.



I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_0$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

GPT-4 To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x . We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:

$$d^2T/dx^2 = 0$$

Integrating once, we get:

$$dT/dx = C_1$$

Integrating a second time:

$$T(x) = C_1x + C_2$$

To determine the constants C_1 and C_2 , we use the boundary conditions:

$$T(0) = T_0, \text{ which gives } C_2 = T_0$$

$$T(L) = T_b, \text{ which gives } C_1 = (T_b - T_0)/L$$

Finally, the expression for $T(x)$ is:

$$T(x) = (T_b - T_0) * (x/L) + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0)/L$.

Predictable Scaling

OpenAI codebase next word prediction

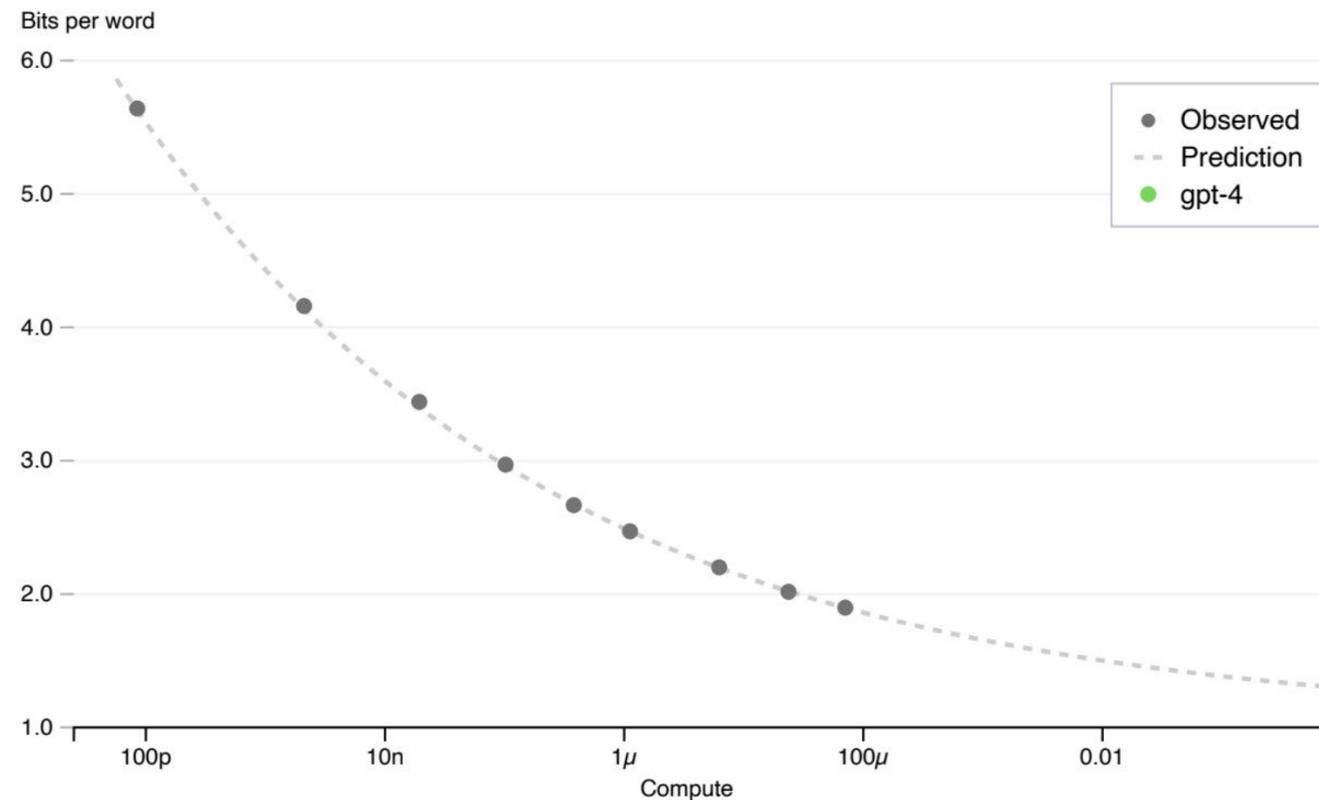
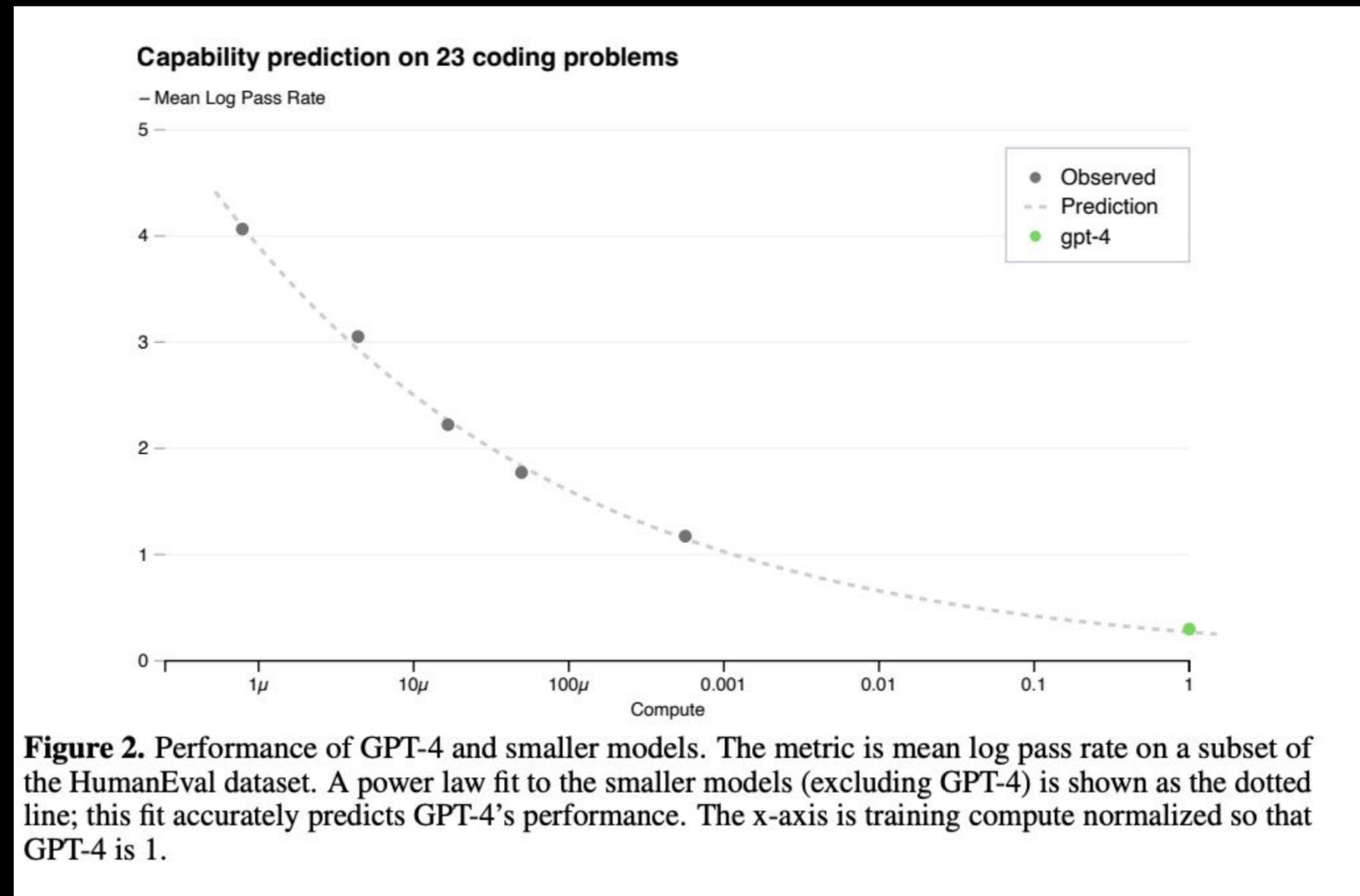
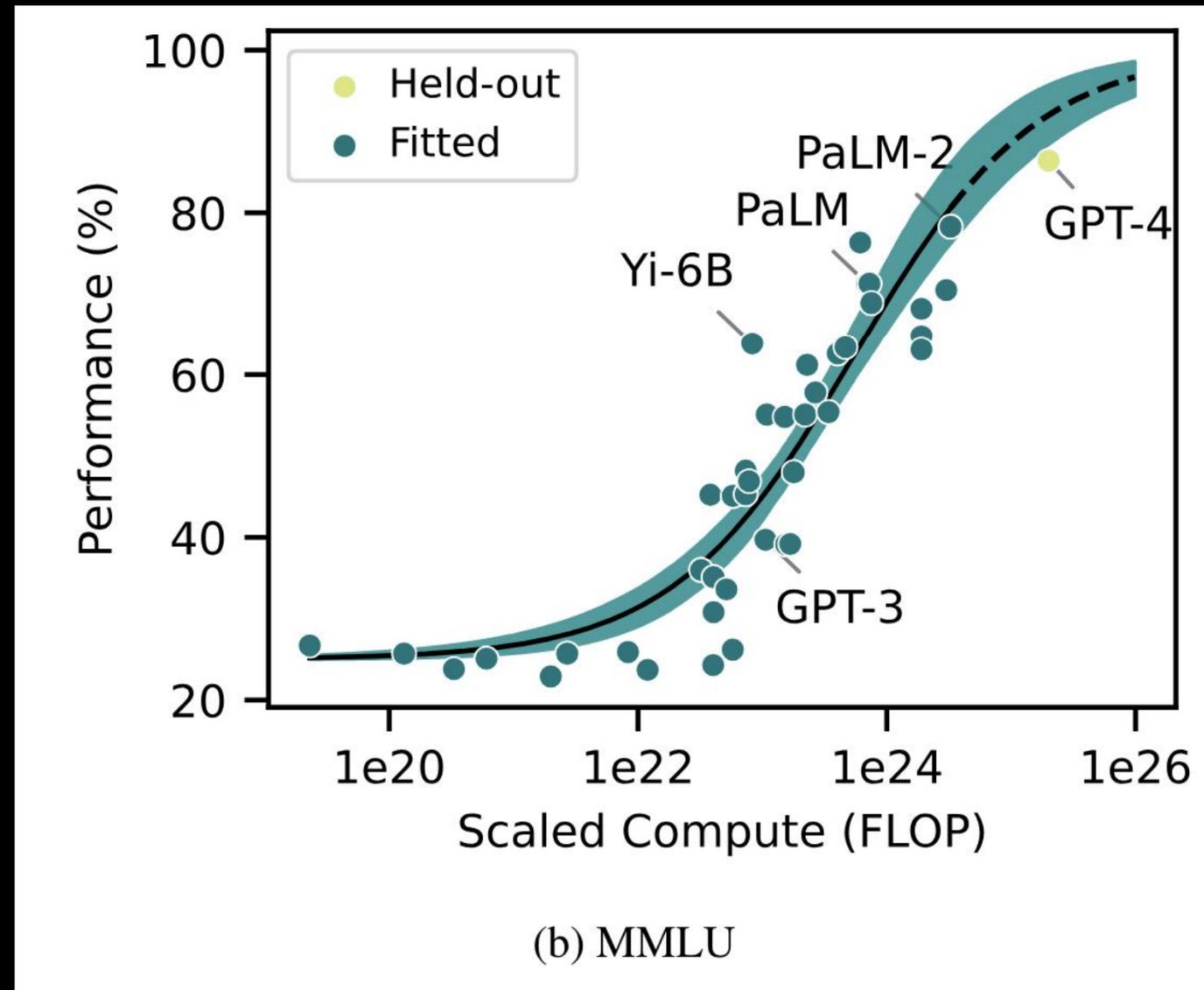


Figure 1. Performance of GPT-4 and smaller models. The metric is final loss on a dataset derived from our internal codebase. This is a convenient, large dataset of code tokens which is not contained in the training set. We chose to look at loss because it tends to be less noisy than other measures across different amounts of training compute. A power law fit to the smaller models (excluding GPT-4) is shown as the dotted line; this fit accurately predicts GPT-4’s final loss. The x-axis is training compute normalized so that GPT-4 is 1.

Predictable Scaling of Practical Capabilities



What Log-Log Plots Obscure



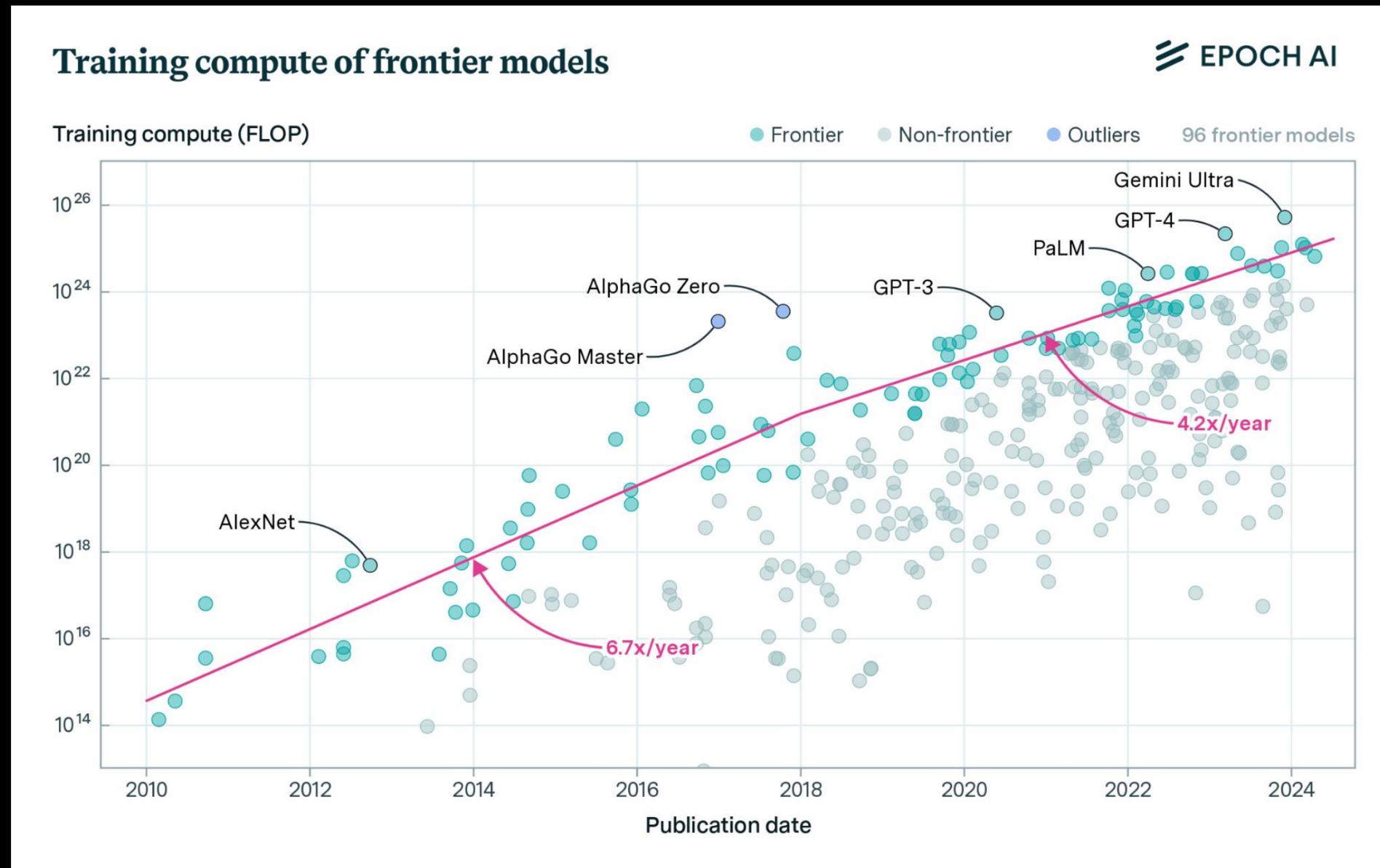
How predictable is language model benchmark performance?, Owen (2024)

Recap: Scaling Laws for AI Models

1. Next-word prediction is meaningful.
2. There are returns to scale
3. ... which are predictable and extrapolative (!!!)

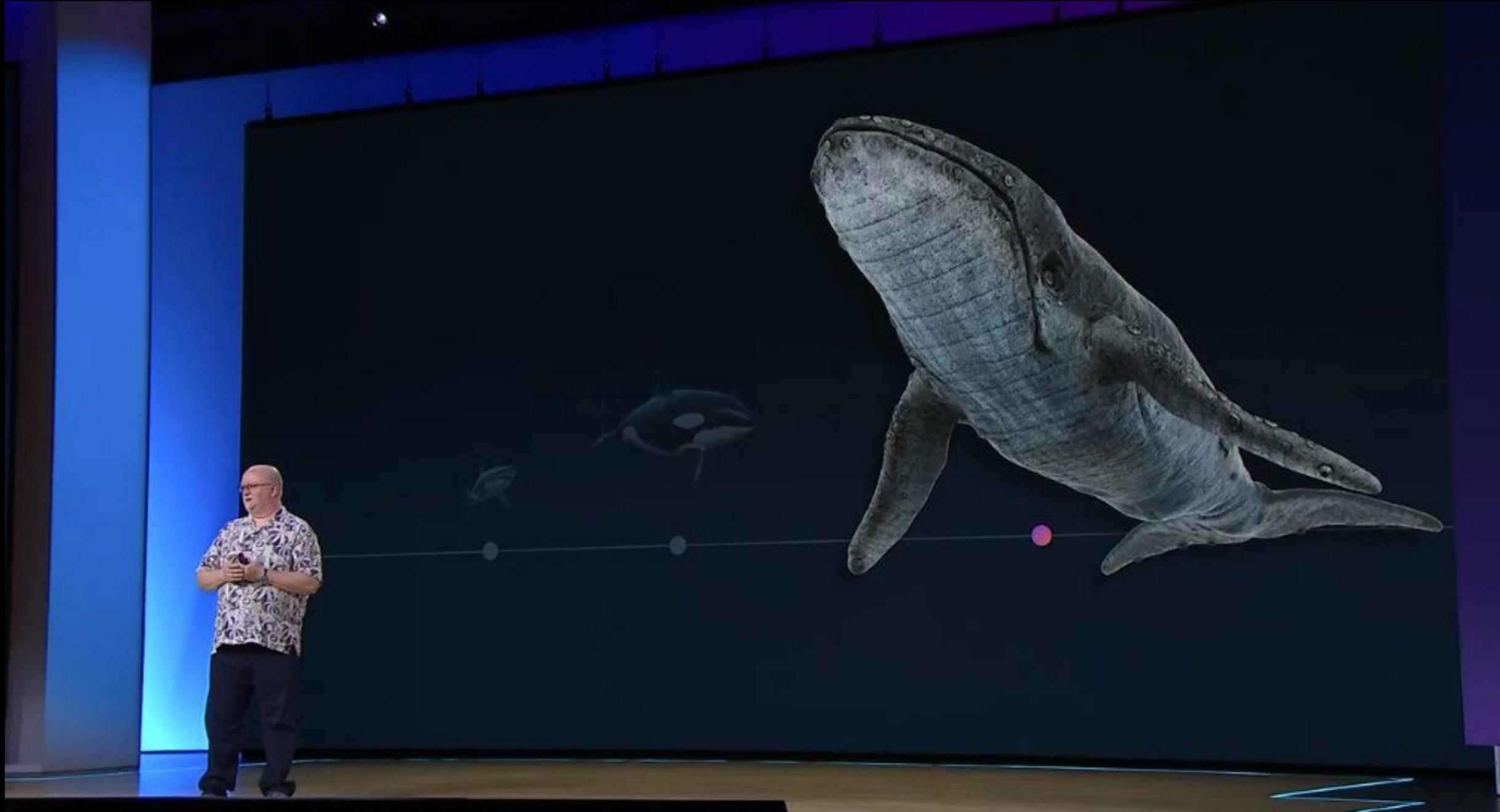
Implications for Infrastructure

Industry Compute Trends

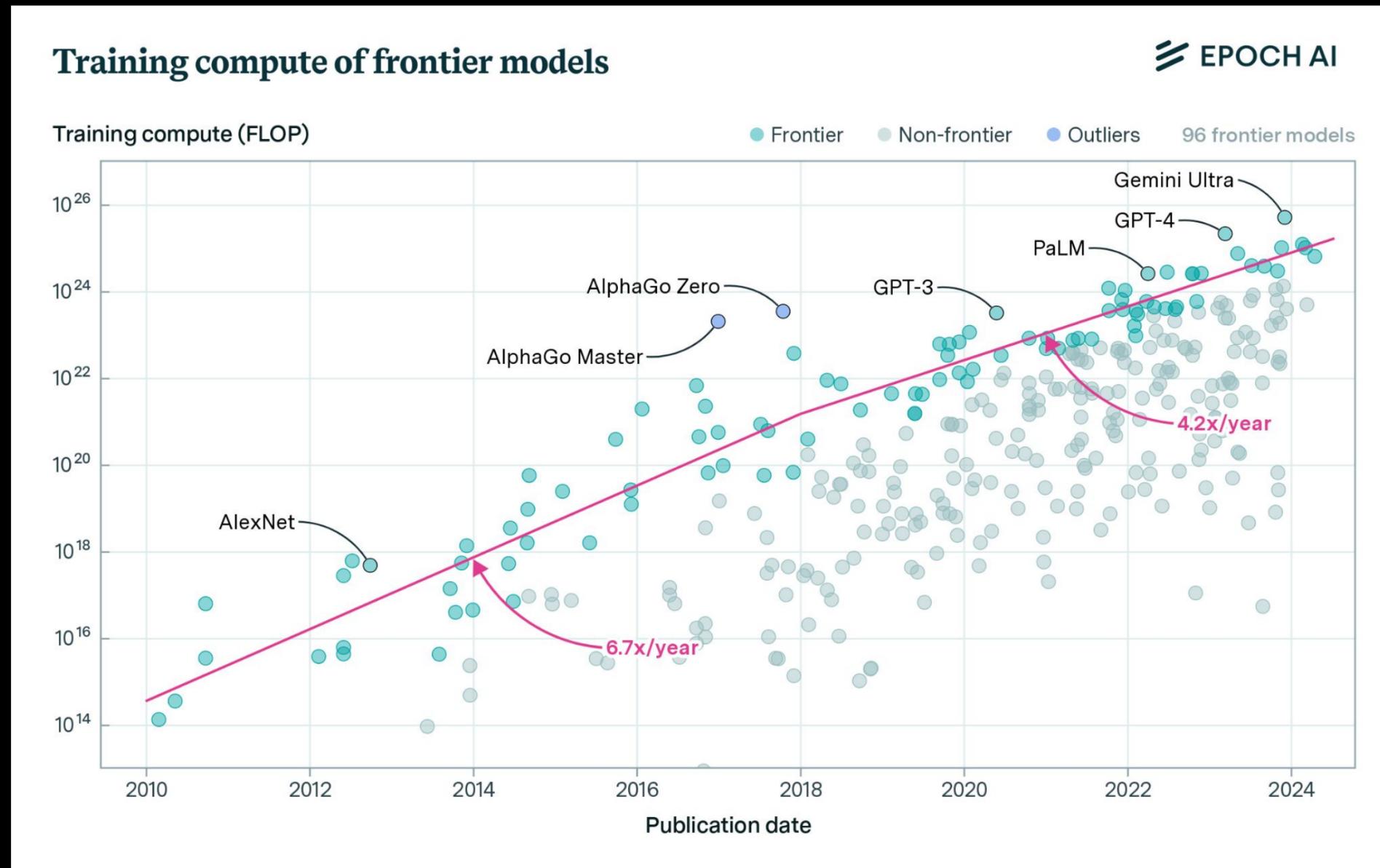


Training Compute of Frontier AI Models Grows by 4-5x per Year, Sevilla and Roldán (2024)

OpenAI Compute Trends

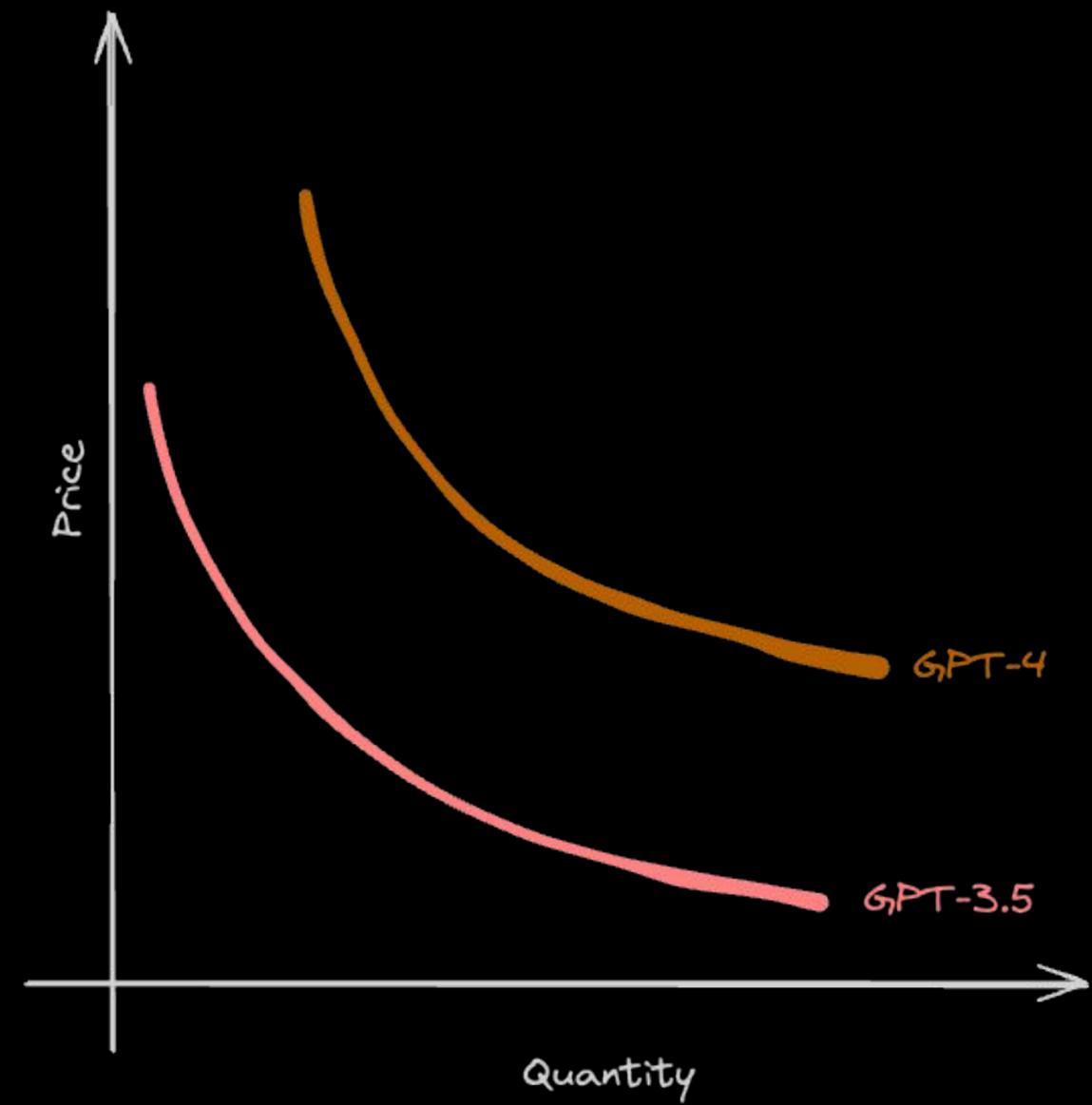
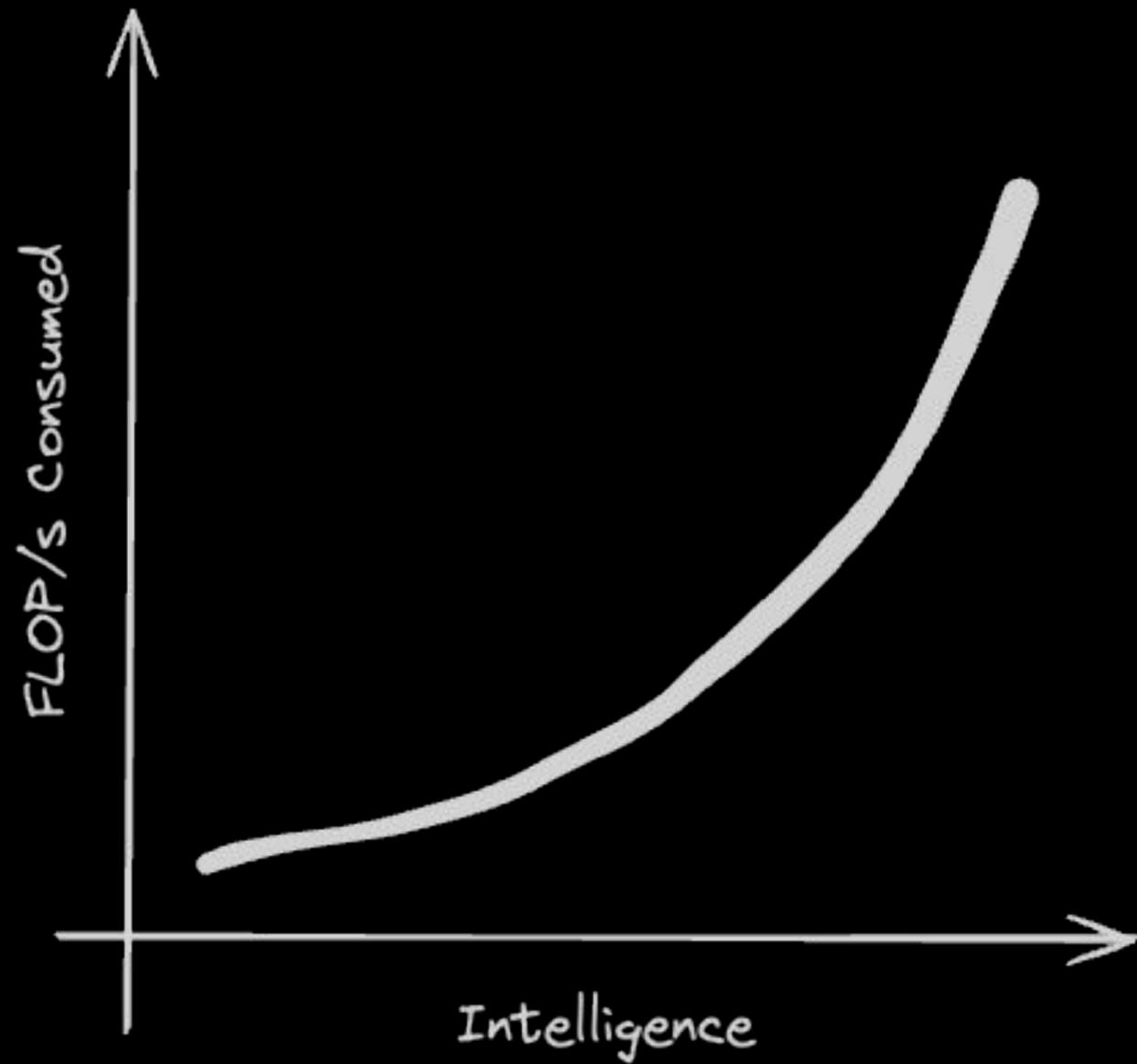


Industry Compute Trends



Training Compute of Frontier AI Models Grows by 4-5x per Year, Sevilla and Roldán (2024)

Inference Demand is Driven By Intelligence



The Bull Case for AI Compute

1. Compute scaling has been predictable and looks to continue.
2. Intelligence drives inference demand.
3. Technology and economics are ripe for scale this decade.



Sam Altman 

@sama



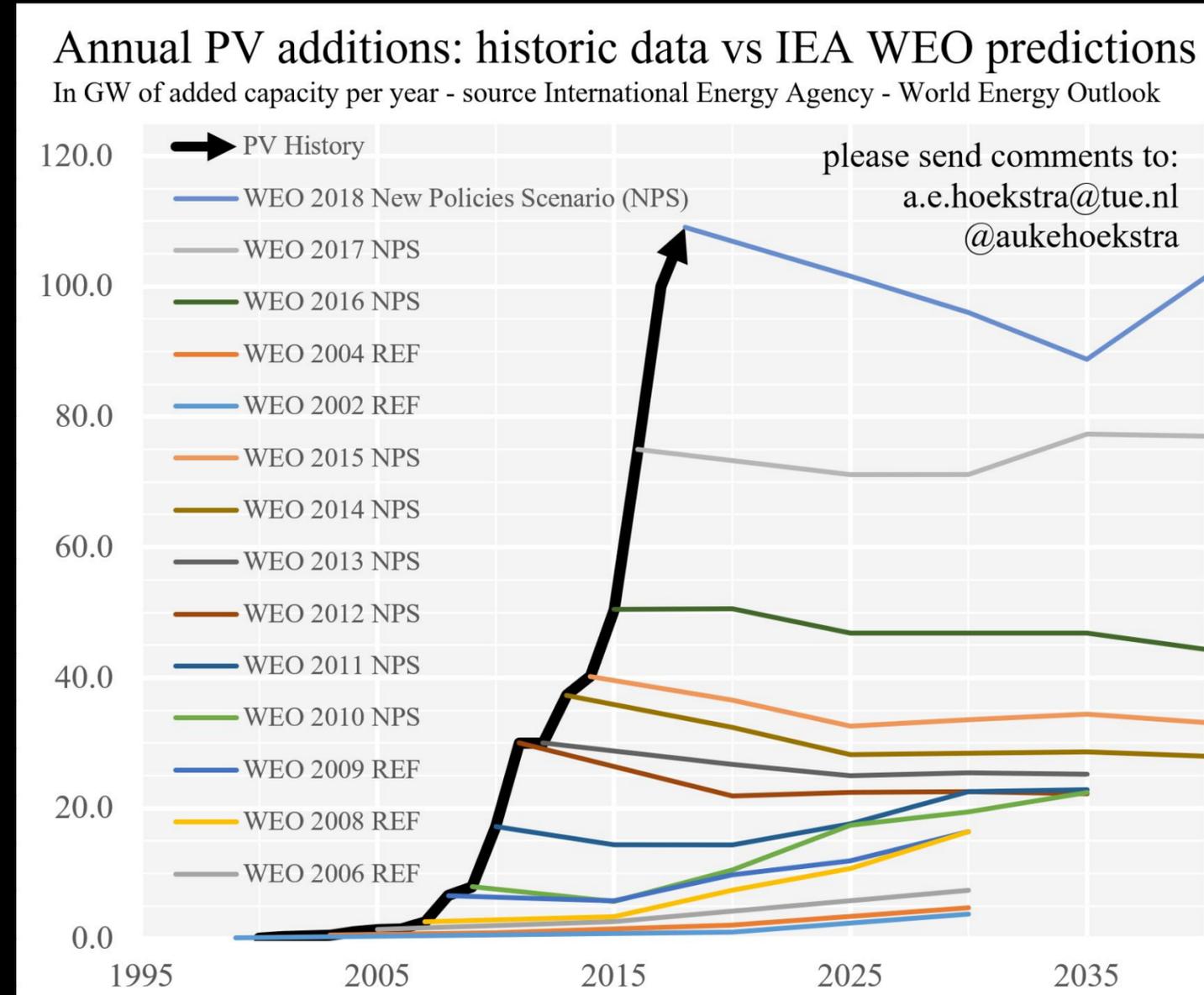
we believe the world needs more ai infrastructure--fab capacity, energy, datacenters, etc--than people are currently planning to build.

building massive-scale ai infrastructure, and a resilient supply chain, is crucial to economic competitiveness.

openai will try to help!

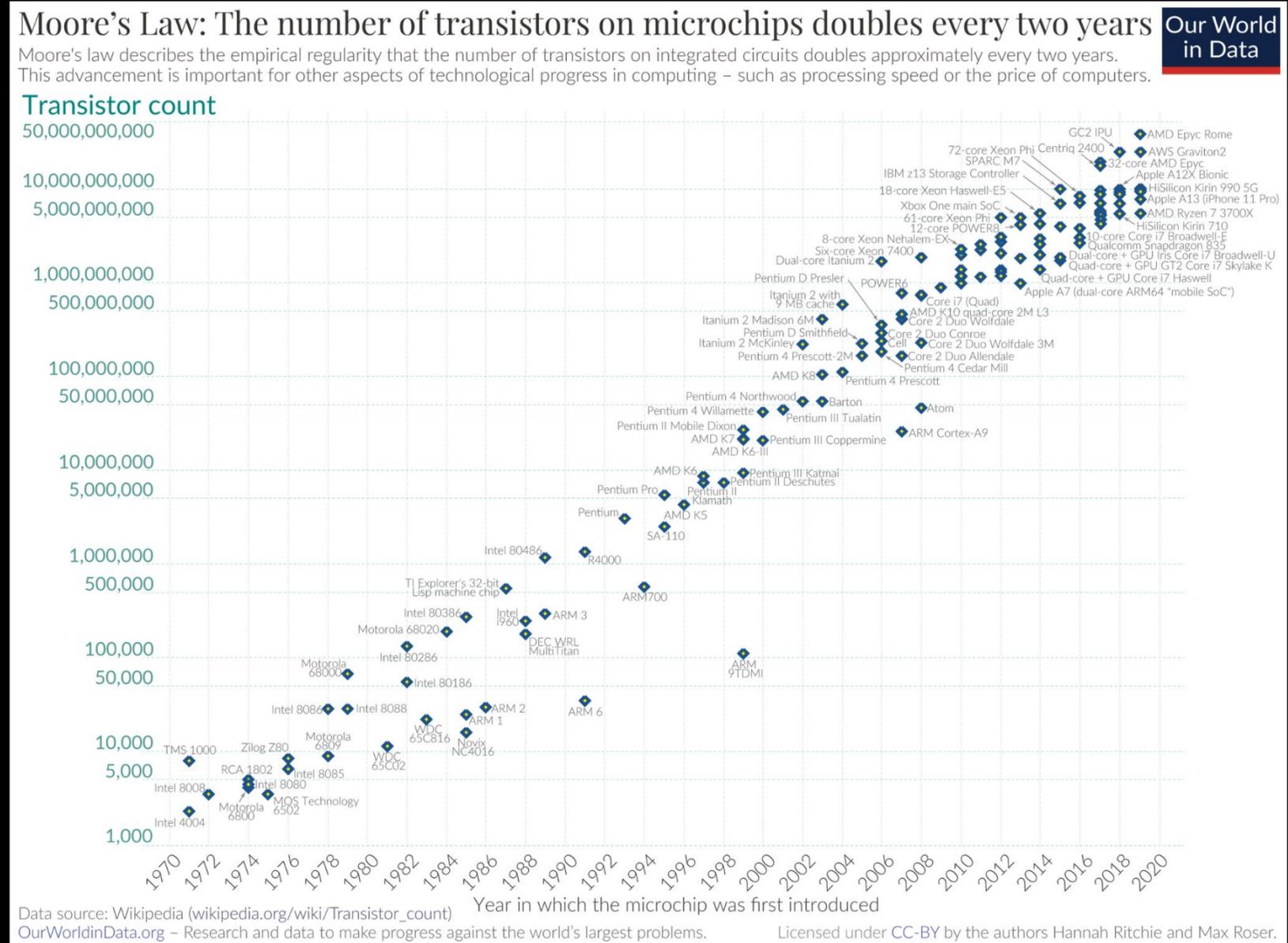
10:17 AM · Feb 7, 2024 · **360K** Views

“Sometimes lines really do go up”



Photovoltaic growth: reality versus projections of the International Energy Agency, Hoekstra (2018)

“Sometimes lines really do go up”



What is Moore's Law?, Roser and Ritchie (2020)

Design for Mass Deployment

Cluster-Level RAS

- Optics MTBF alone measured in minutes.
 - Not to mention HBM DUE, board failures, etc.
- SDCs: disturbingly common, and sometimes unreproducible.
- Failures have a very wide blast radius.

Cluster-Level RAS

- Minimize cost of repair.
 - Exception > Process Restart > GPU Reset > Node Reboot > RMA.
 - Example: Ideally, failed write over scale-up is a catchable exception.
- Minimize blast radius.
 - Example: Link flaps on one port should not affect neighboring ports.
 - Example: Ideally, uncorrectable memory error only affects own GPU.
 - Even in presence of coherent memory fabric.

Cluster-Level RAS

- Consider graceful degradation.
 - Some failures are more worth technician time than others.
 - Example: Disable faulty banks of second-tier memory instead of requiring RMA.
- Validation must be automated, fast, extensive, and performable in-field.
 - Example: In-depth correctness checks after SDC accusation.

Power Management

- Power bottlenecks mean we need to maximize power we have.
- Synchronized training steps results in power draw jitter.
- Need: Low-latency power telemetry and OOB power management.
- Want: Dynamic power sloshing.

Takeaways

1. Predictable scaling motivates rapidly expand AI training compute usage.
2. Delivering AI to the world will demand massive infrastructure buildout.
3. Design for mass deployment.
4. Performance is only one of many requirements.

Thank You!

