



Meek Models Shall Inherit the Earth

Neil Thompson

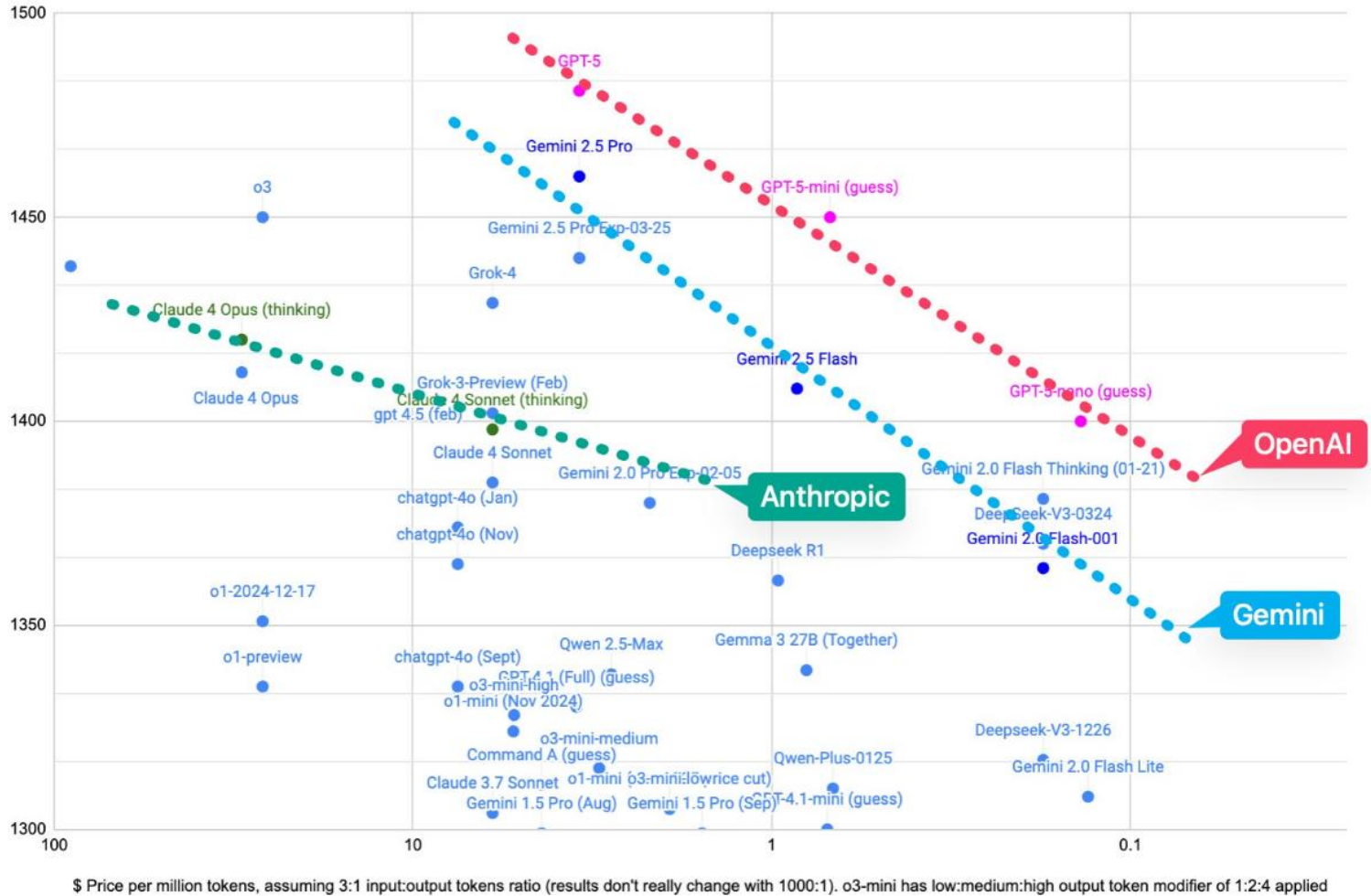
Director, MIT FutureTech

Group Lead, MIT Initiative on the Digital Economy



Plot of model pricing vs LMSys Elo (Aug 2025) - full analysis on <https://latent.space>

Frontier of model costs and performance



Source:
Latent Space

\$ Price per million tokens, assuming 3:1 input:output tokens ratio (results don't really change with 1000:1). o3-mini has low:medium:high output token modifier of 1:2:4 applied

Big vs Small Models

If only big models have advanced capabilities

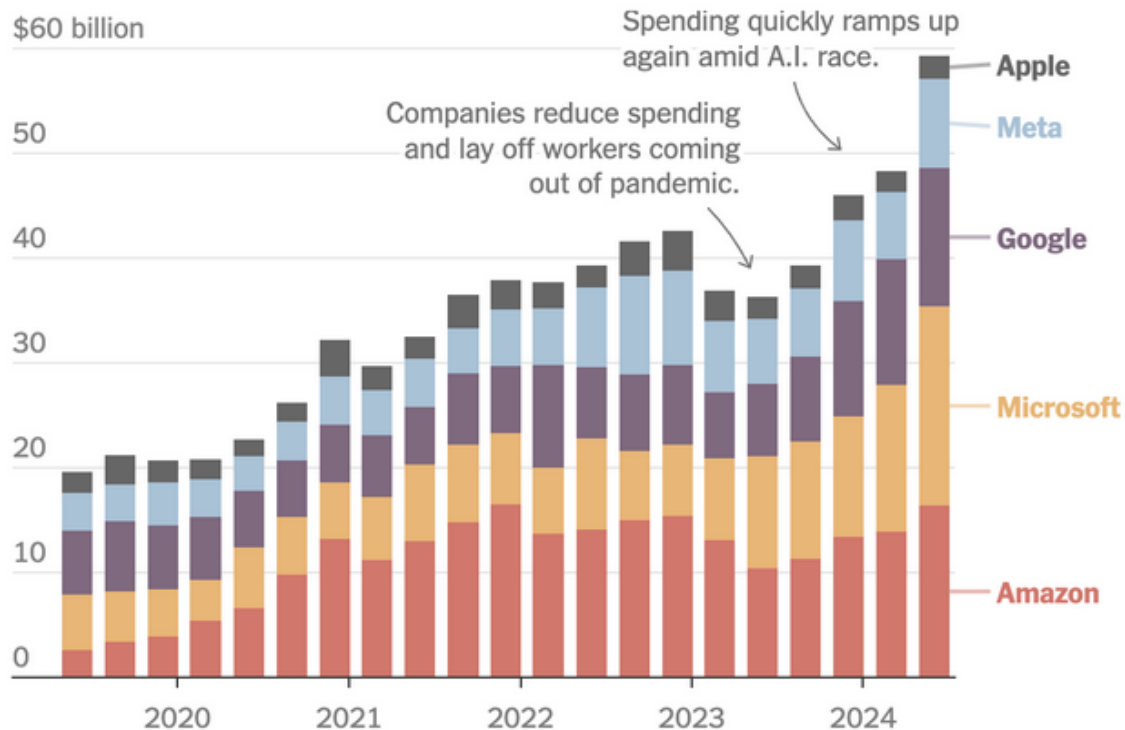
- Natural monopoly
 - A few big companies (countries) dominate
 - High financial and environmental costs
- Cybersecurity is crucial to protect models
 - Running a stolen model is much than training it
- Large developed-developing world AI gap

If small models also have advanced capabilities

- Advanced technologies diffuse widely
 - Developing world has better tools
 - So do cybercriminals
- Environmental footprint drops considerably

Big Tech Spending Has Shot Up

Capital expenditures in the latest quarter were up 63 percent from a year earlier.

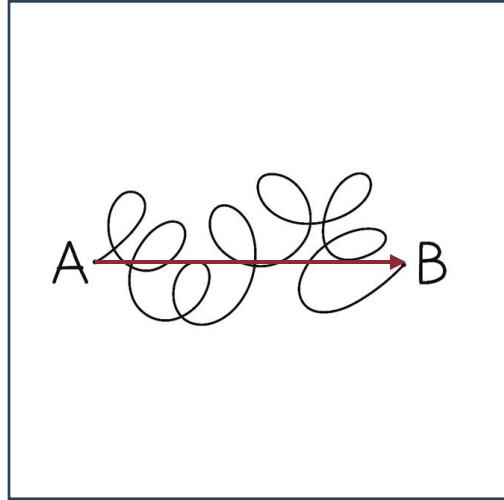


Source: FactSet and company filings • Karen Weise

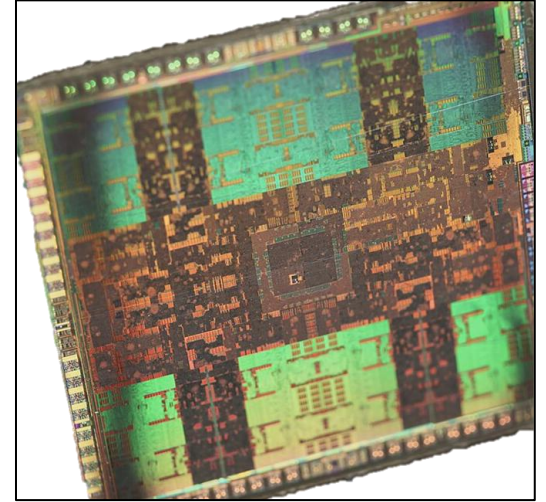
Where does AI improvement come from?



More chips, running longer

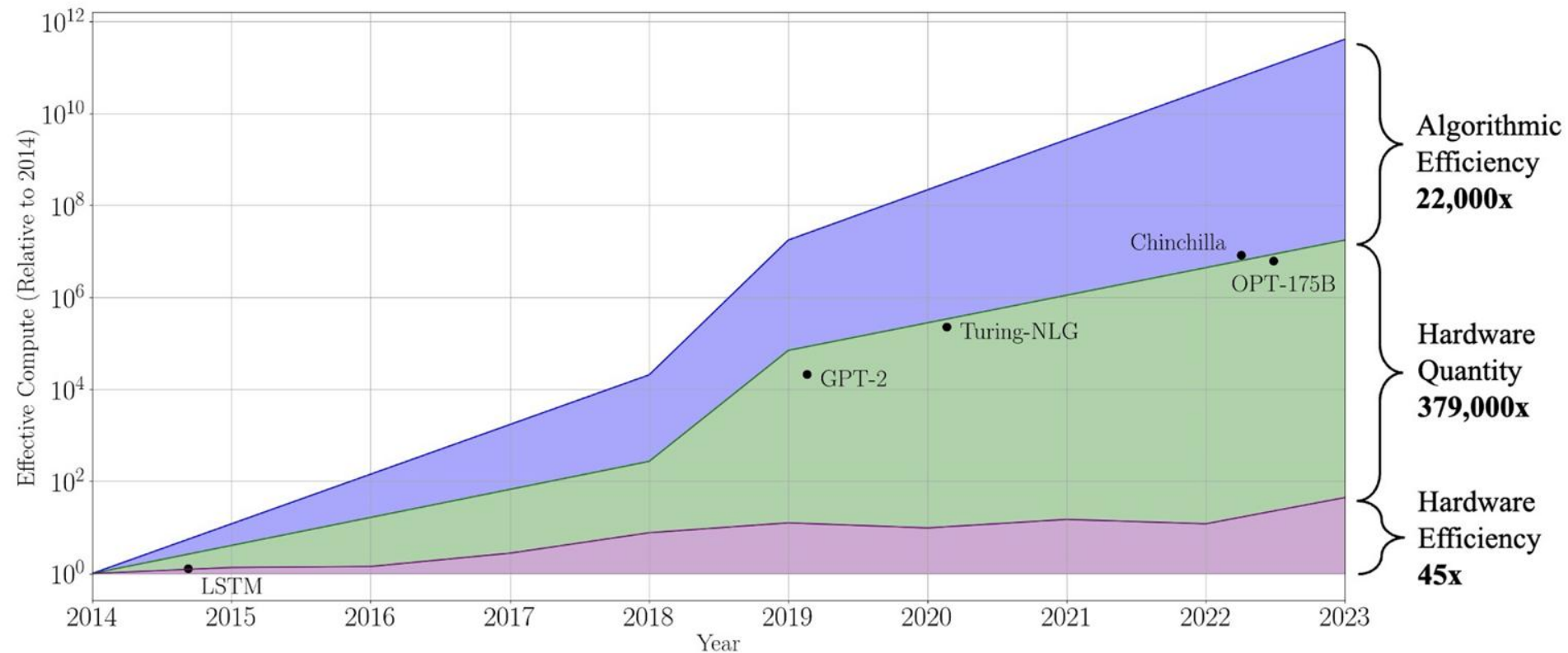


More efficient AI algorithms



More efficient GPUs

Where has past LLM progress come from?



Source: Thompson, based on Ho et al. (2024) and Del Sozzo et al (2025)

Consider building two models

Big Model

- Gets to use ever-newer GPUs (e.g. via cloud) & algorithm improvements (e.g. transformer)
- Starts with cost of \$1,000 and then grows each year at 3.6x (historical rate)

Meek Model

- Gets to use ever-newer GPUs (e.g. via cloud) & algorithm improvements (e.g. transformer)
- Cost of building model capped at \$1,000 per year

How does competition evolve?

$$L_{opt}(C) = 1070 C^{-0.155} + 1.7$$

$$L_{opt}(C) = A C^{-\alpha} + L_0$$

Training Loss Difference = Loss Meek – Loss SOTA

$$= A((g_{alg}g_h)^t C_0)^{-\alpha} - A((g_{alg}g_hg_i)^t C_0)^{-\alpha}$$

Algorithm progress:

$$g_{alg} = 2.8x$$

Hardware progress:

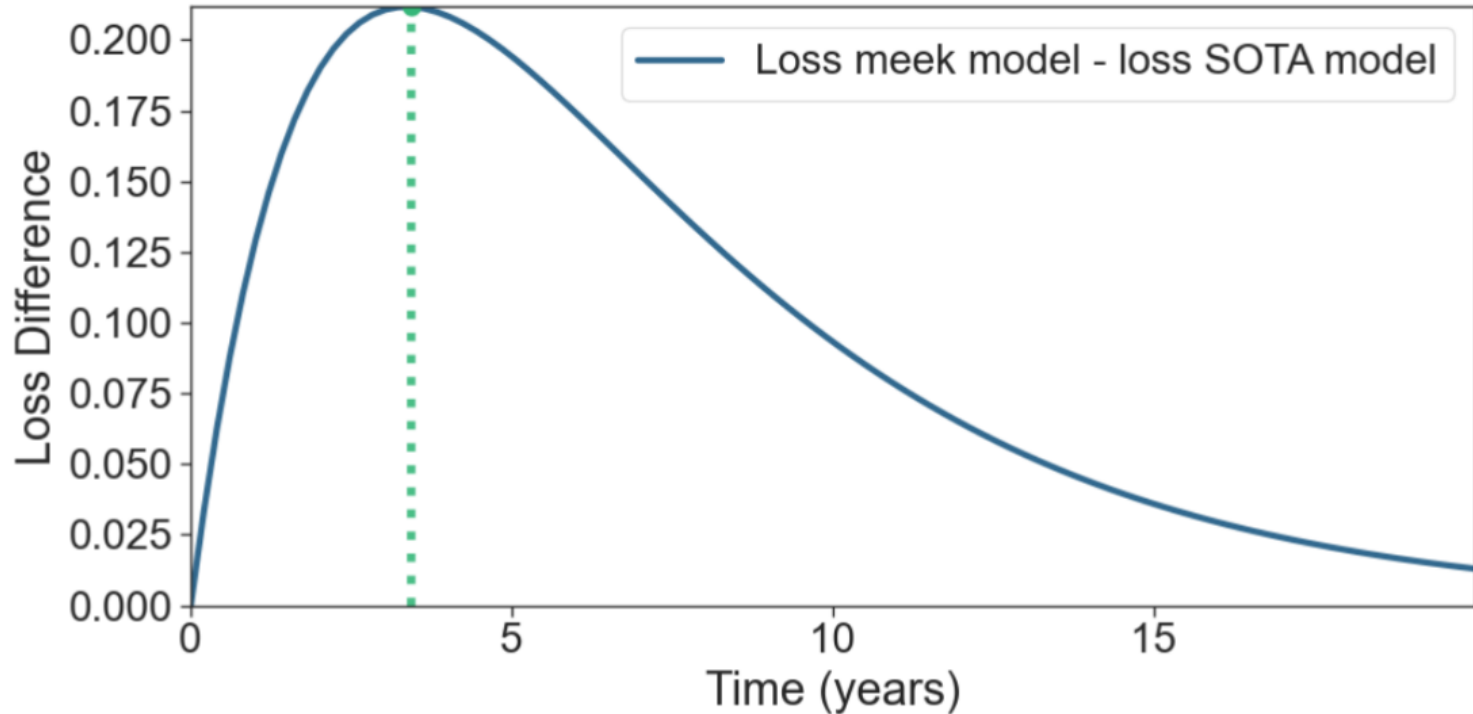
$$g_{alg} = 1.4x$$

Investment scale-up:

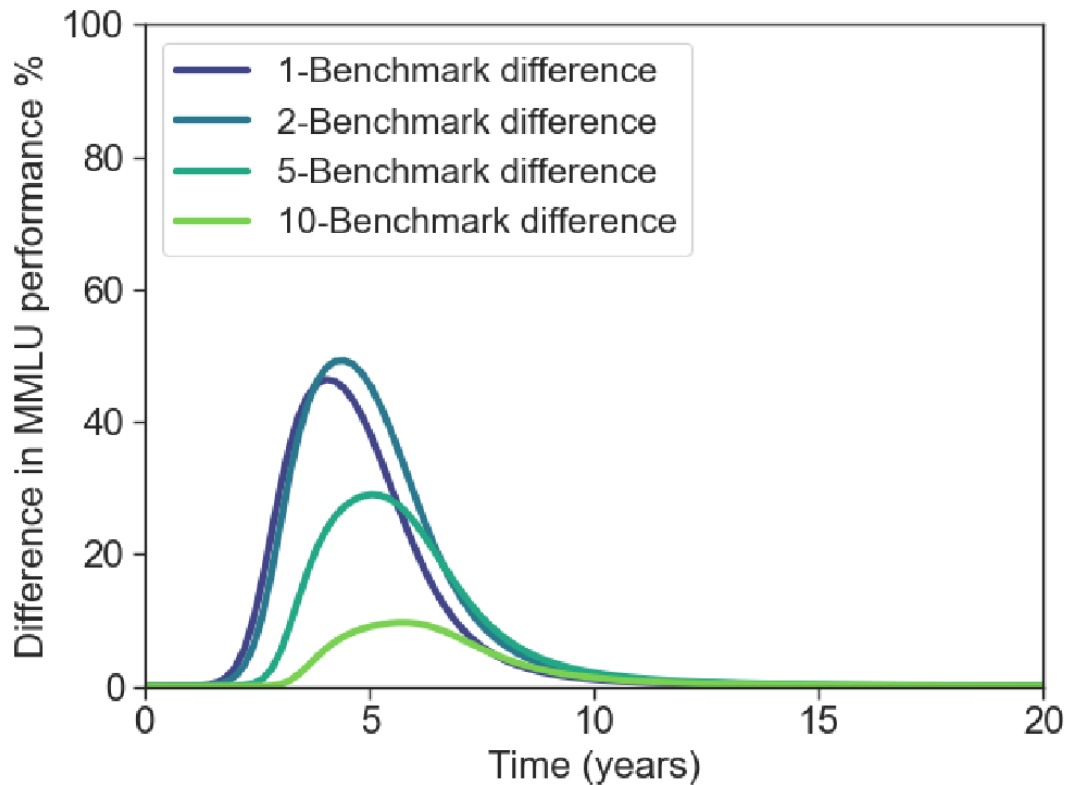
$$g_{alg} = 3.6x$$

Gap in performance rises, then falls

(Measured in loss)



Also true if converted to benchmark performance



Is this really happening?

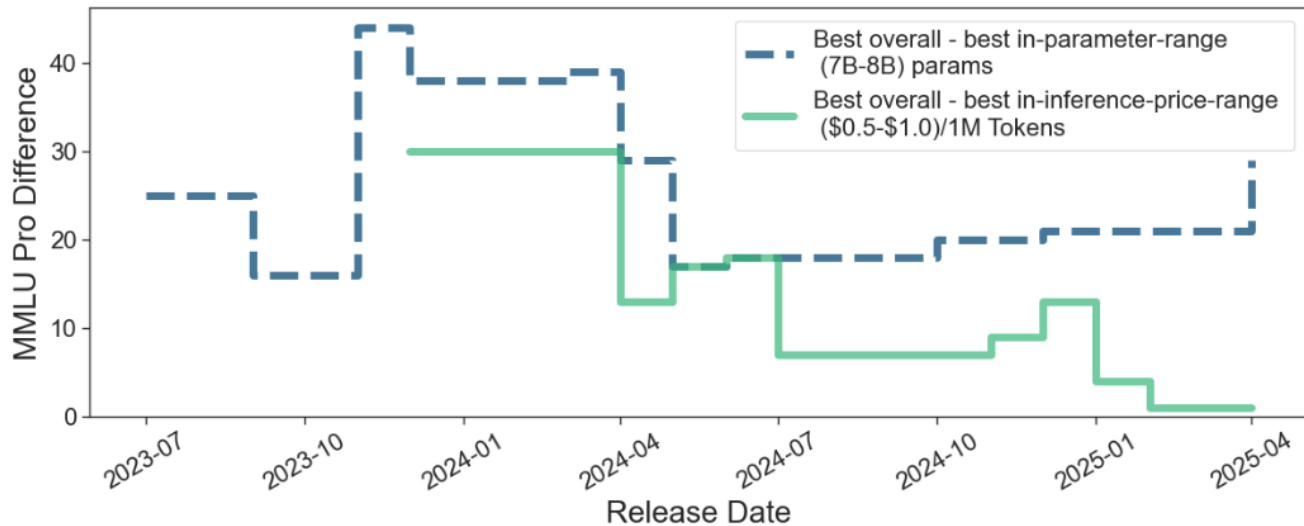


Figure 5: Graph depicting the difference in MMLU-Pro score between the model with the maximum score overall and the best score among models within a fixed inference price range 0.5\$ – 1\$ per 1M tokens. Models sourced from the Artificial Analysis LLM Leaderboard [Artificial Analysis \[2025\]](#).

Summary

- Cutting-edge performance driven by:
 - Scaling up computation (AI scaling laws)
 - Hardware progress (better GPUs)
 - Algorithm progress (more efficient code)
- Initially, scaling up computation has big returns to better performance
- But within a few years the yearly progress in a “meek” model (\$1,000 per year) quickly improves as fast (and then faster) than cutting-edge models
 - Arises because the returns to scaling-up computation decrease so quickly
- Implies much greater democratization of AI capabilities in the coming years



Meek Models Shall Inherit the Earth

Neil Thompson

Director, MIT FutureTech

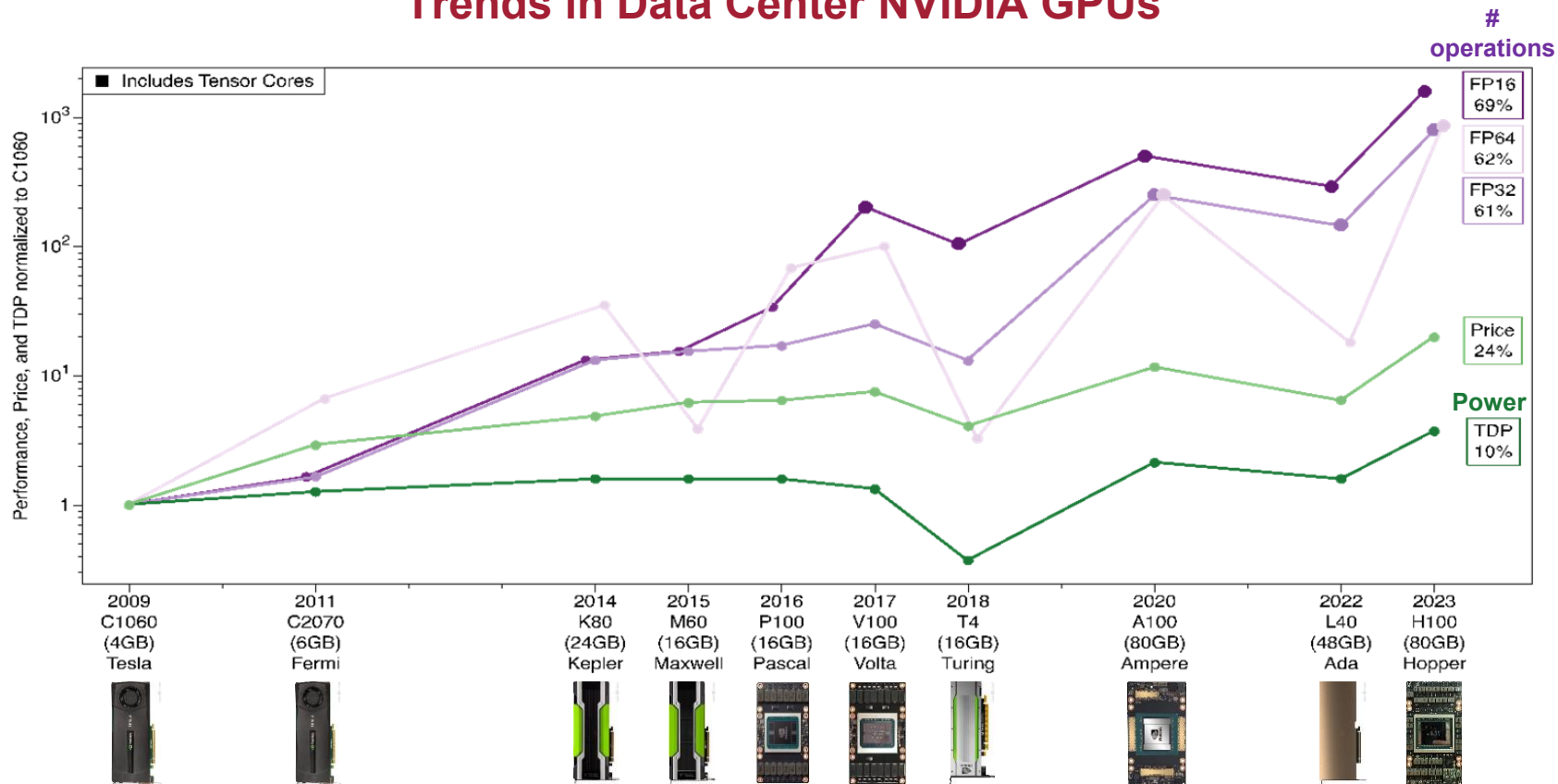
Group Lead, MIT Initiative on the Digital Economy



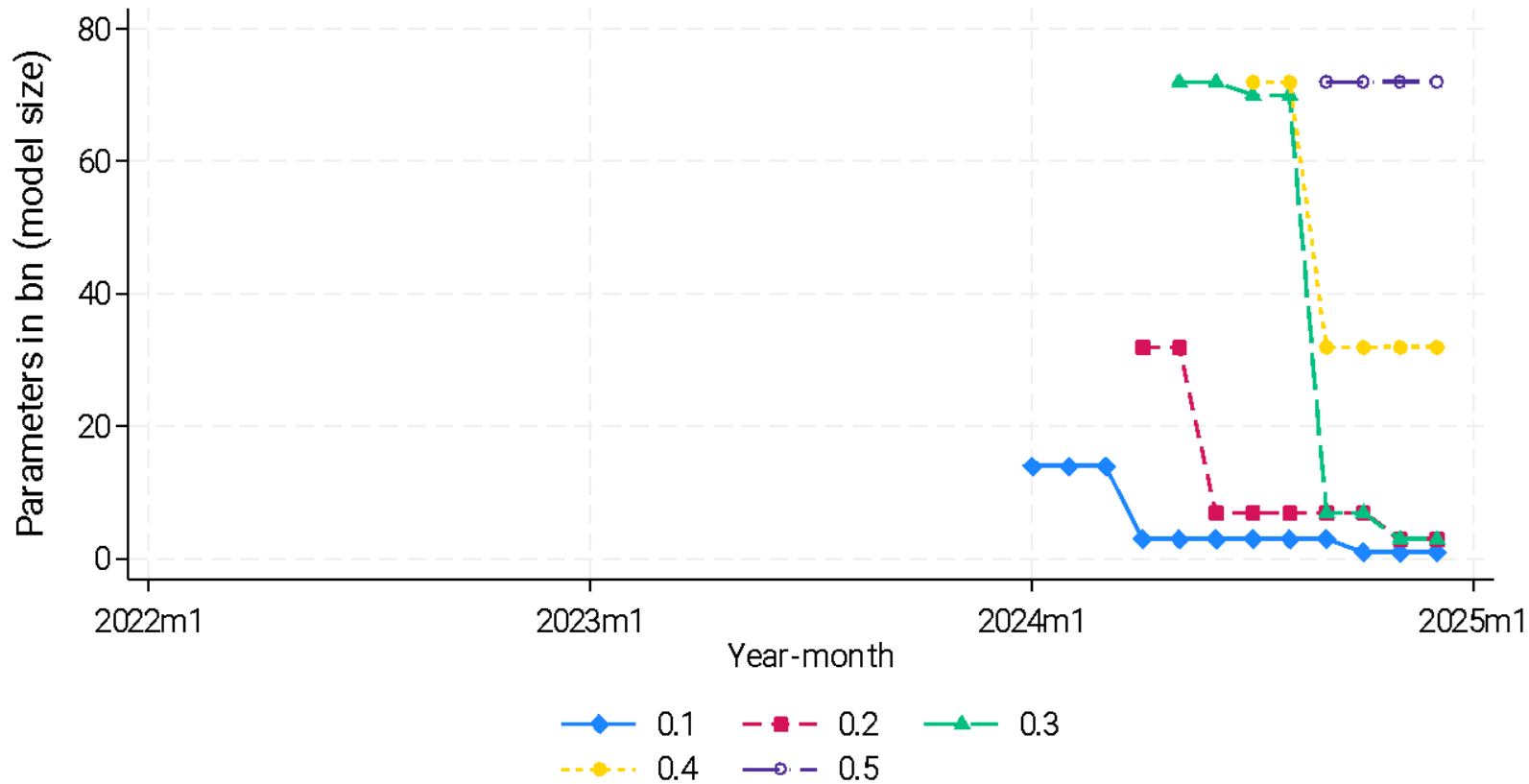
MIT FutureTech
Innovations that Shape the World



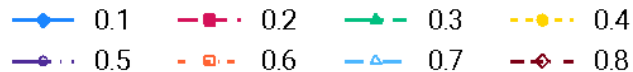
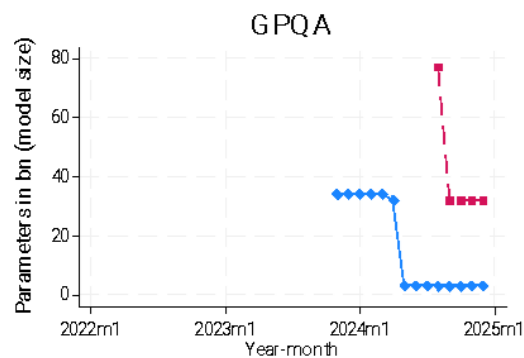
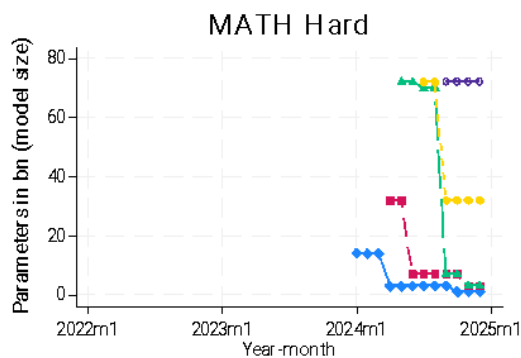
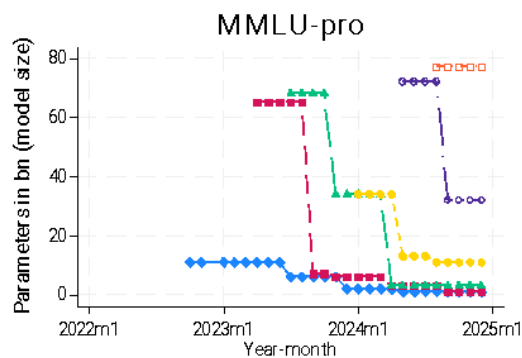
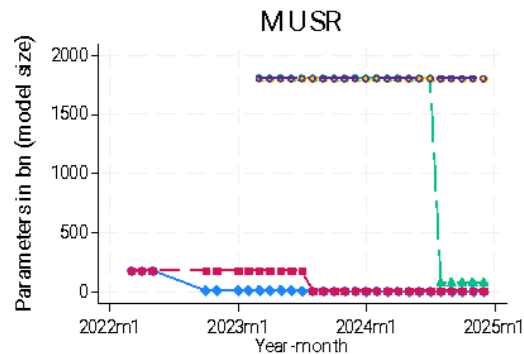
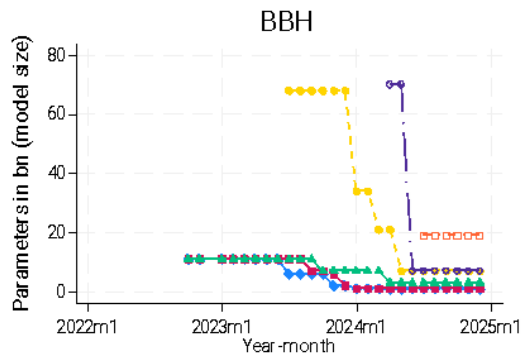
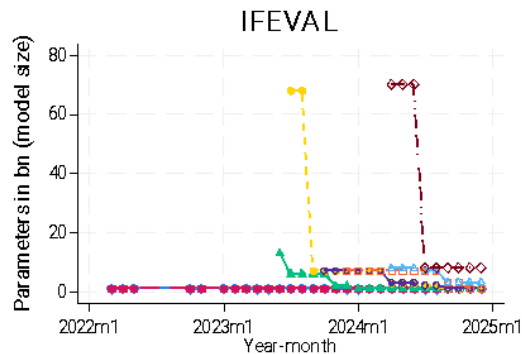
Over time compute cost falls: Trends in Data Center NVIDIA GPUs



MATH Hard



Required model size to achieve given benchmark performance



Notable AI Models

Training compute (FLOP)

483 Results

