What Will GPT-2030 Look Like?

Jacob Steinhardt

Simons Institute
February 20, 2024

Forecasting Model Capabilities

MATH: contest problems with non-obvious multistep reasoning

- Suppose that x, y, and z satisfy the equations $xyz = 4$, $x^3 + y^3 + z^3 = 4$, $xy^2 + x^2y + xz^2 + x^2z + yz^2 + y^2z = 12$. Calculate the value of $xy + yz + zx$.
- In the coordinate plane, the graph of $\|x + y - 1\| + \|\|x\| - x\| + \|\|x - 1\| + x - 1\| = 0$ is a certain curve. Find the length of this curve.

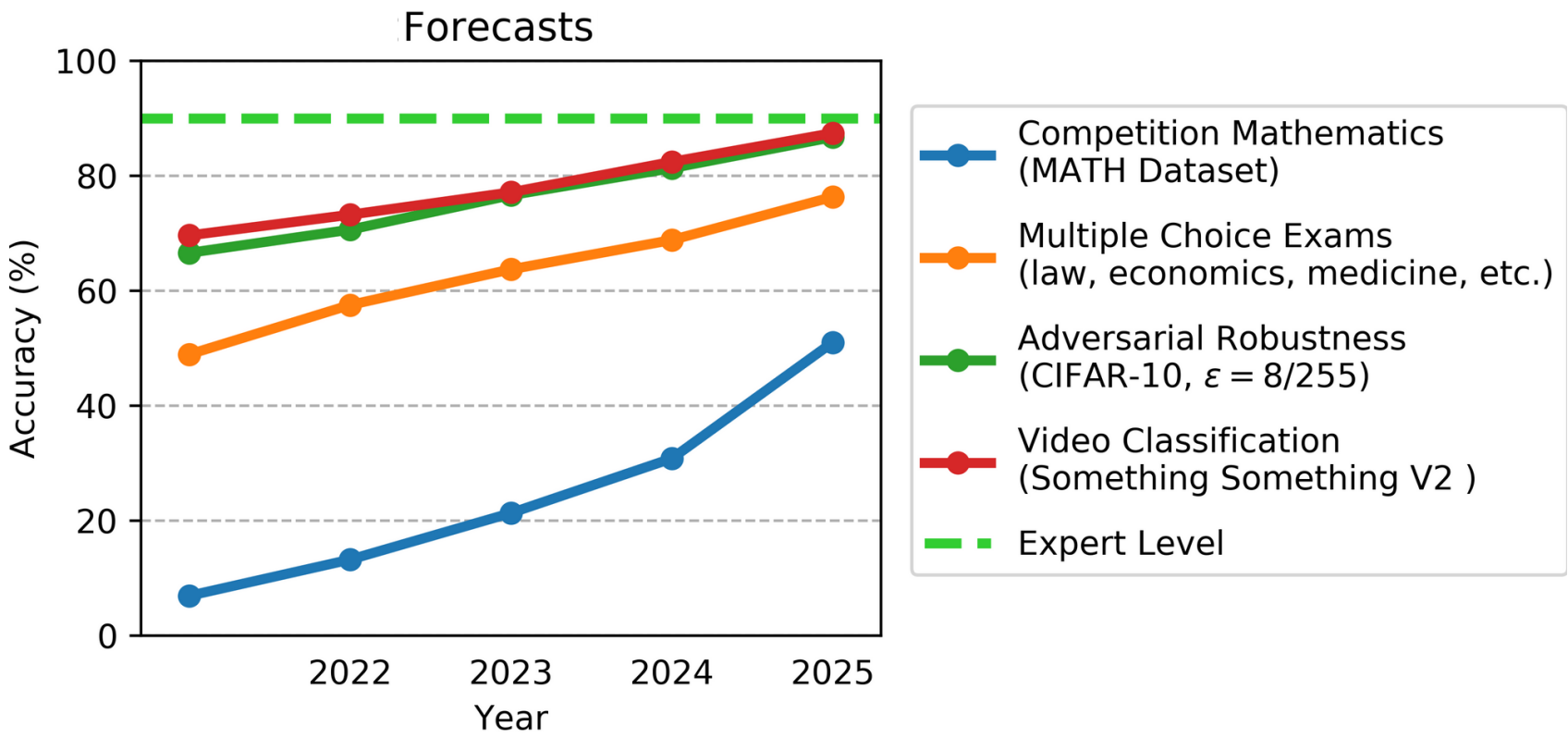
MMLU: multiple-choice high-school, college, and professional exams

- (Philosophy) According to Moore's "ideal utilitarianism," the right action is the one that brings about the greatest amount of:
  - (A) pleasure. (B) happiness. (C) good. (D) virtue.
- (College Medicine) In a genetic test of a newborn, a rare genetic disorder is found that has X-linked recessive transmission. Which of the following statements is likely true regarding the pedigree of this disorder?

Hendrycks et al. (2021), Hendrycks et al. (2020)

"What will SOTA be on June 30, 2022, 2023, etc.?" (Asked in 2021)

Summary of Benchmark Forecasts (from 2021)

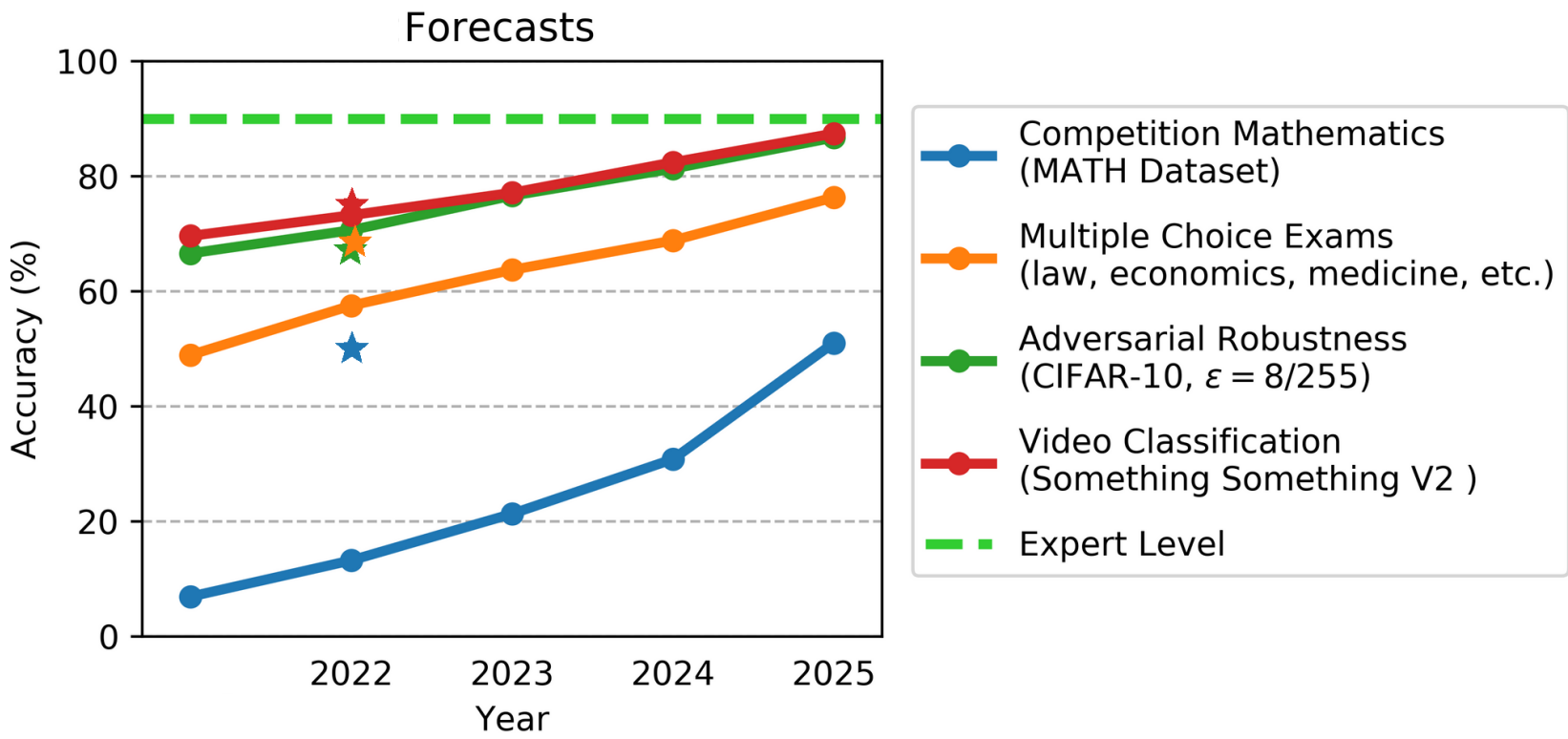


Forecasters predicted faster progress than most ML researchers

- Based on personal experience / anecdotes

"Lessons and Updates from AI Forecasting", Steinhardt (2021)

Results (2022)



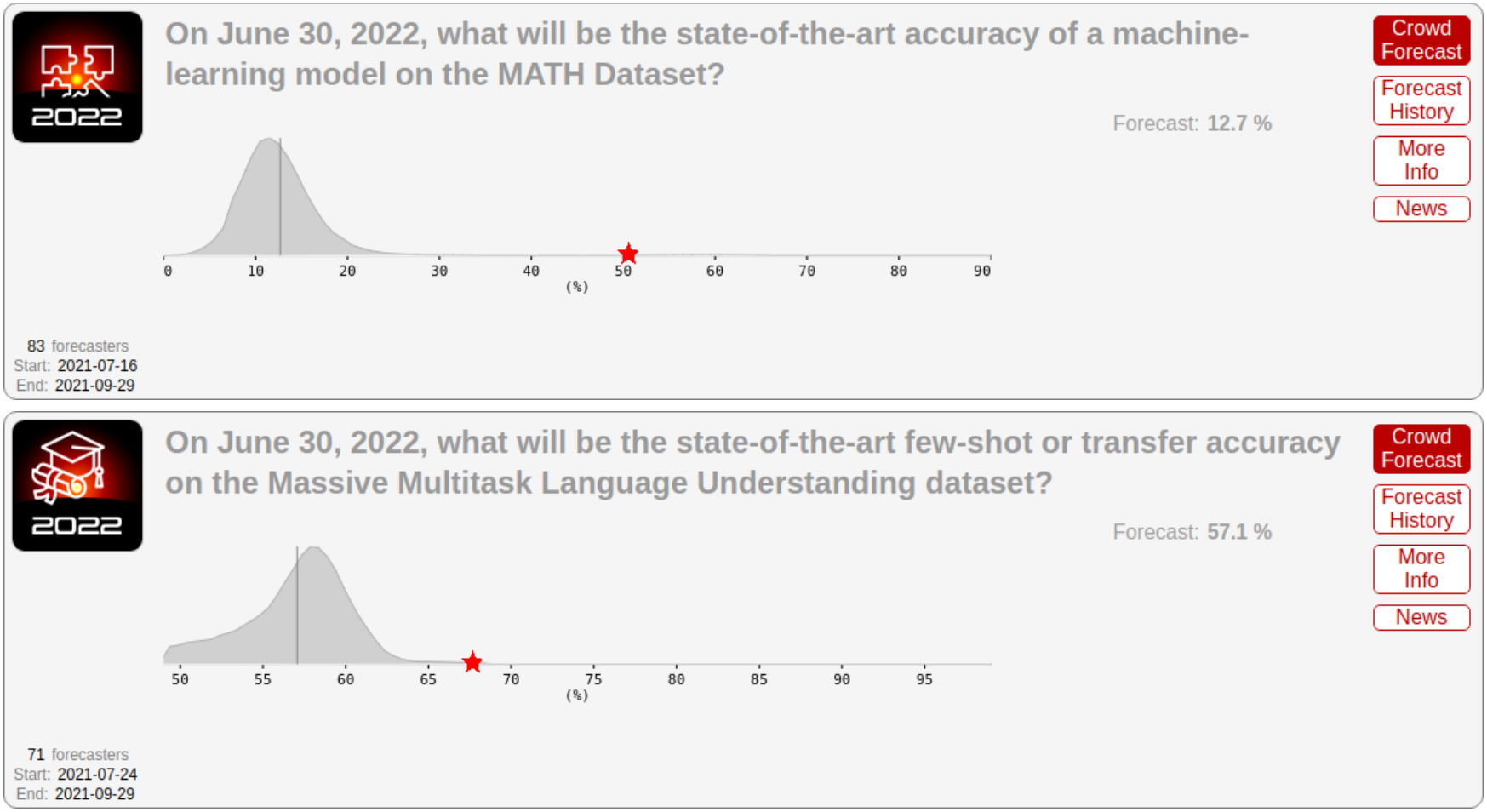
MATH, MMLU were far above the forecast

Video was somewhat above the forecast

Robustness was below the forecast

"AI Forecasting: One Year In", Steinhardt (2022)

Forecaster Calibration

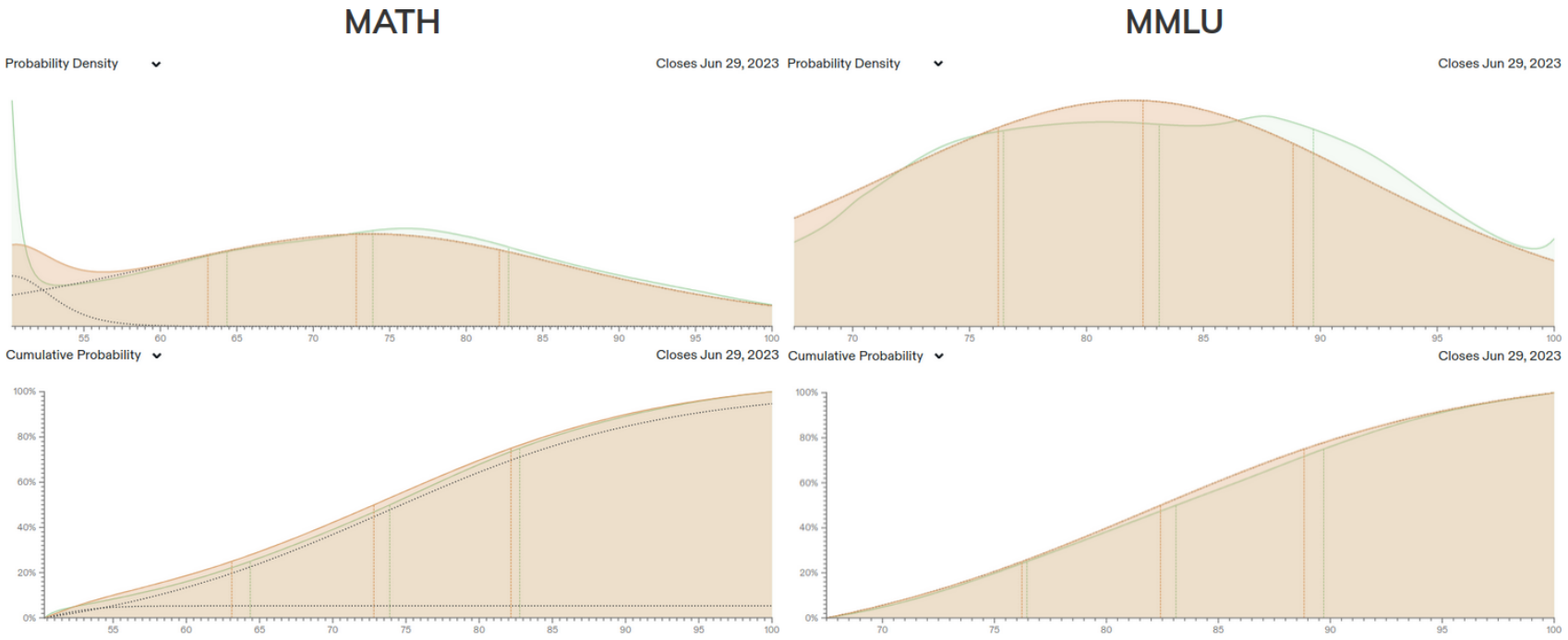


Actual SOTA far outside predicted range!

"AI Forecasting: One Year In", Steinhardt (2022)

Trying again for 2023

My updated MATH, MMLU forecasts:

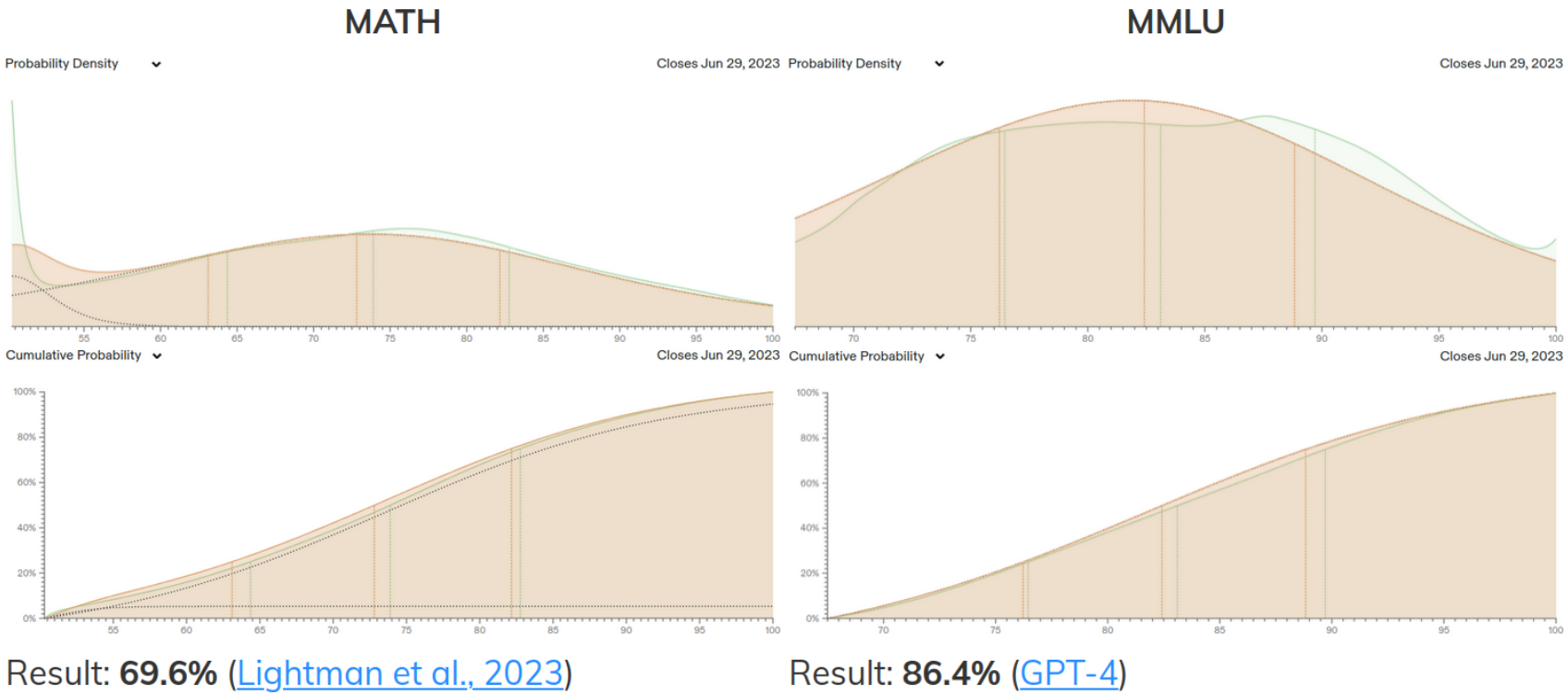


Result: **69.6%** (Lightman et al., 2023)

Result: **86.4%** (GPT-4)

Trying again for 2023

My updated MATH, MMLU forecasts:



Result: **69.6%** (Lightman et al., 2023)

Result: **86.4%** (GPT-4)

Reality vs. predictions: 41st percentile (MATH), 66th (MMLU)

Steinhardt (2022), "Forecasting ML Benchmarks in 2023"

Most Important Forecasting Skills

Be calibrated

Consider the full outcome space (MECE + "other" option)

Zeroth order forecast (know current / historical state of world)

First order forecast

Most Important Forecasting Skills

Be calibrated

Consider the full outcome space (MECE + "other" option)

Zeroth order forecast (know current / historical state of world)

First order forecast ← our focus

GPT-2030

Foundation model trained in 2030 on language + other modalities

Predict following characteristics:

- "Training overhang" (# of parallel copies)
- Speed of inference
- Cost of inference
- Capabilities and modalities
- Parallel learning
- Degree of autonomy

"What Will GPT-2030 Look Like?", Steinhardt (2023)

Training Overhang

Train FLOPs $\gg$ forward pass FLOPs

Currently:

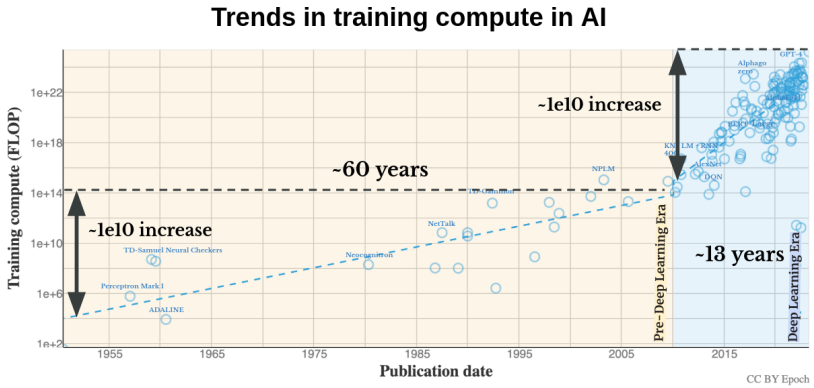
- Training FLOPs $\propto$ Data $\times$ Model size
- Data, model grow proportionally (Chinchilla)
- $\implies$ Overhang grows as sqrt(Training FLOPs)

Training Overhang

Train FLOPs $\gg$ forward pass FLOPs

Currently:

- Training FLOPs $\propto$ Data $\times$ Model size
- Data, model grow proportionally (Chinchilla)
- $\implies$ Overhang grows as sqrt(Training FLOPs)



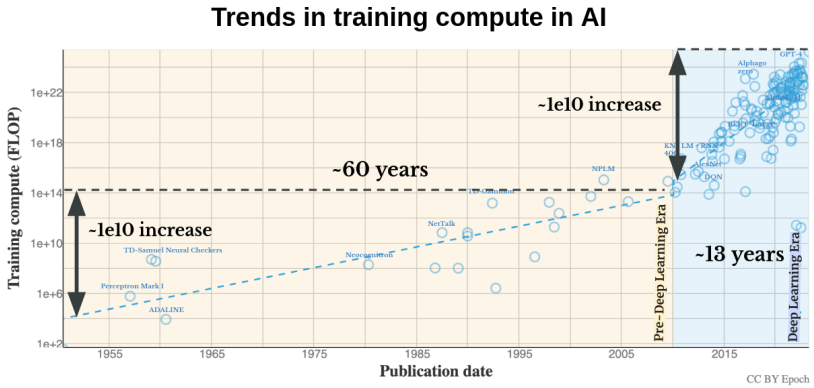
Sevilla et al. (2022)

Training Overhang

Train FLOPs $\gg$ forward pass FLOPs

Currently:

- Training FLOPs $\propto$ Data $\times$ Model size
- Data, model grow proportionally (Chinchilla)
- $\implies$ Overhang grows as sqrt(Training FLOPs)

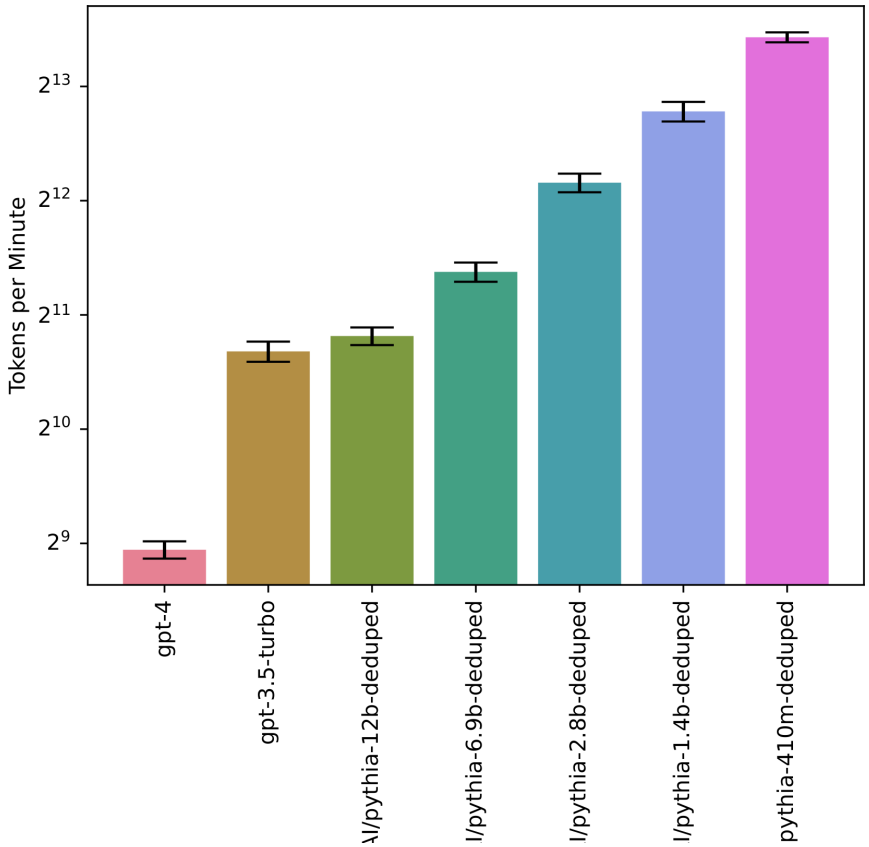


Sevilla et al. (2022)

Projection:

- Training FLOPs grows 0.62 OOM/year $\implies$ overhang grows 0.31 OOM/year
- Extrapolation: 4.2e14 in 2030 $\implies$ 1.6 million "human-years" of work [0.4M, 10M]

Inference Speed



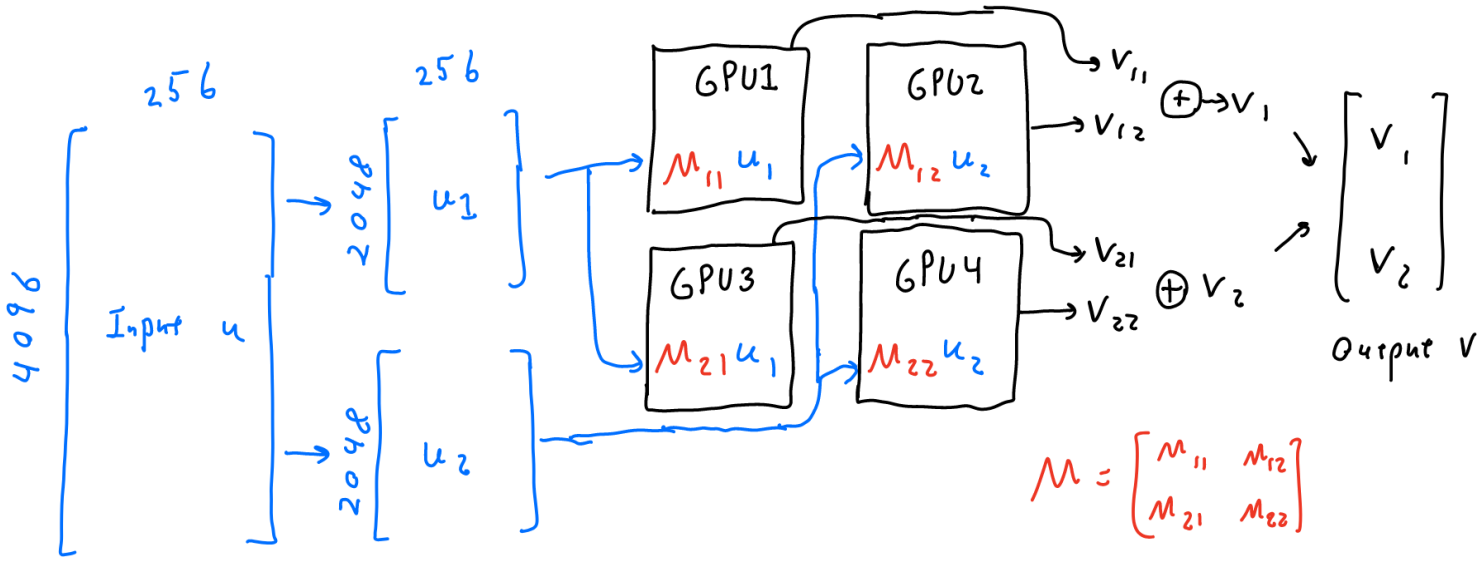
GPT-4 improving over time (at least 700/minute now)

Likely room to further improve

Prediction: 2500 toks/minute, or 5x human speed (range: [0.5x, 20x])

Inference Speed: Turbo Boost

Inference speed is not constant (trade throughput for latency)



Can increase speed by $\approx k^2$ if willing to accept $1/k$ utilization

Likely can take $k = 10$ or maybe higher $\implies$ >100x human speed

"How fast can we perform a forward pass?", Steinhardt (2022)

Inference Cost (Rough)

Grows as Model Size x ($/FLOP)

Model size: 0.31 OOM/year

$/FLOP: 0.12 OOM/year decrease

Increase by $10^{7 \cdot 0.19} = 21$x

$2.50 / 1k tokens ($75/"hour") at OpenAI pricing rates (c. 06/2023)

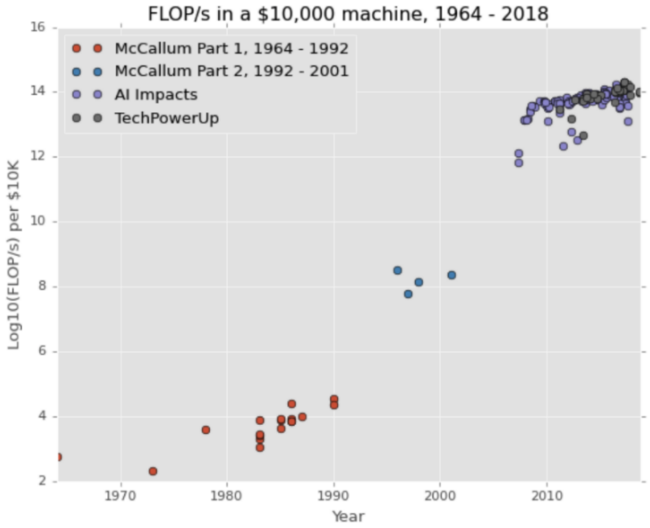
Inference Cost (Rough)

Grows as Model Size x ($/FLOP)

Model size: 0.31 OOM/year

$/FLOP: 0.12 OOM/year decrease

Increase by $10^{7 \cdot 0.19} = 21$x



From Cotra 2020; original chart from Kathleen Finlinson

$2.50 / 1k tokens ($75/"hour") at OpenAI pricing rates (c. 06/2023)

Capabilities

Likely superhuman at math, programming, hacking

- APPS dataset: 7.8% (AlphaCode) $\rightarrow$ 25.5% (Parsel)
- Metaculus median: 80% in 2027 (very strong human)
- MATH dataset: high velocity, 92% by 2025 (Metaculus)
- IMO gold medal by LLM: 2028 (Metaculus)

Possibly superhuman at manipulation, protein design

Capabilities

Likely superhuman at math, programming, hacking

- APPS dataset: 7.8% (AlphaCode) $\rightarrow$ 25.5% (Parsel)
- Metaculus median: 80% in 2027 (very strong human)
- MATH dataset: high velocity, 92% by 2025 (Metaculus)
- IMO gold medal by LLM: 2028 (Metaculus)

Possibly superhuman at manipulation, protein design

Likely modalities: language, code, images

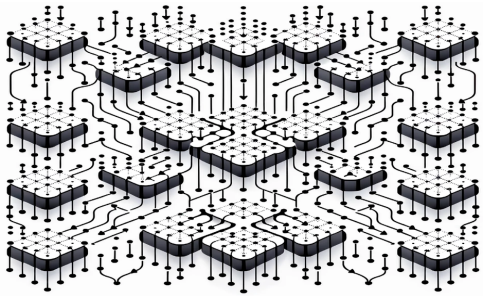
Possible: video, speech, proteins, astronomical data, network traffic(?)

Tool use, possible physical actuation

Continual Learning

All copies of model share weights

Therefore, can "pool experience" via batch gradient updates



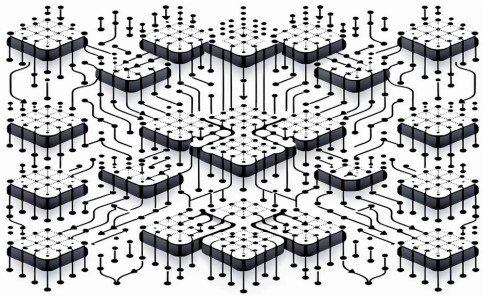
Doesn't slow down learning below "critical batch size"

McCandlish et al. (2018)

Continual Learning

All copies of model share weights

Therefore, can "pool experience" via batch gradient updates



Doesn't slow down learning below "critical batch size"

McCandlish et al. (2018)

Batch sizes in millions routine in practice

10^6 copies $\implies$ learn 10^6 times faster (online learning)

2500 "years" of experience each day

- Possibly a huge deal

Autonomy (Very Speculative)

GPT-2030 is blindingly fast, with many copies

But can we use this speed?

GPT-4 can write ≈ 2000 tokens before falling over

Perhaps 4x better than GPT-3 (?)

Assume GPT-2030 is 500x better than GPT-4

1M tokens $\approx$ 1 "day"

Implications

2M copies; $<0.1\%$ of human workforce, but can:

- assign arbitrary tasks
- distill (run cheaper)
- possibly coordinate much better

Run for an AI-day (few human-hours) before needing feedback

Likely high-quality reports on outputs

- Undergrad coding wizards with very good explanation skills

Implications

2M copies; $<0.1\%$ of human workforce, but can:

- assign arbitrary tasks
- distill (run cheaper)
- possibly coordinate much better

Run for an AI-day (few human-hours) before needing feedback

Likely high-quality reports on outputs

- Undergrad coding wizards with very good explanation skills
- Aside: I don't think this would 5x my productivity (maybe 2x). But it might 5x grad students, by turning them into "profs"

Implications

Math is done

AI: we'll be supervising teams of AI models

Companies could create master manipulators if they wanted

Sensitive tech (proteins, cyber) might be widely accessible

The Future is (Very) Uncertain

- 5, 15, 40 years?
- Will GDP growth $= 2\%$, 4%, or 30%?
- Cost to run: \$75/hour or \$7/hour?
- Misalignment, cyberattacks, inequality, surveillance, ...?

Overview of risks: Hendrycks, Mazeika, Woodside (2023)

The Future is (Very) Uncertain

- 5, 15, 40 years?
- Will GDP growth $= 2\%$, 4%, or 30%?
- Cost to run: \$75/hour or \$7/hour?
- Misalignment, cyberattacks, inequality, surveillance, ...?

Overview of risks: Hendrycks, Mazeika, Woodside (2023)

Interventions I favor:

- Make sure key functions grow with AI progress

Monitoring: Tong, Jones, S. (2023)

Explaining complex phenomena: Zhong et al. (2022, 2023)

Forecasting: Zou et al. (2022)

Cybersecurity, biosecurity, alignment

- Ensure AI has key properties (that remain past human-level)

Honesty: Burns & Ye et al. (2022), faithful explanations of internal state

- Basic science of AI
- Better forecasts and scaling projections