AI progress

National Academy Committee on Automation and the U.S. Workforce, July 6th 2022

Structure of this talk:

- Past ten years of AI progress
- Past year of AI progress (and why it has been surprising)
- Something strange is beginning to happen…

- Why measurement is getting increasingly hard
- Open questions and problems

What happened over the past ten years?

Scaling up of systems and convergence around an increasingly common, universal set of tools.

Shift from academia to industry in terms of research.

Shift from Western world to 'everyone' (who can afford a big computer) at the frontier.

A yellow school bus park[illegible] parking lot.

A red school bus parked in a parking lot.

The decadent chocolate desert is on the table.

A bowl of bananas is on the table.

Generating Images from Captions with Attention - 2015.

A family standing in front of a sign while wearing skis and holding ski poles.

A train being operated on a train track.

Three boys playing a soccer game on a green soccer field.

DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis - 2020

Teddy bears mixing sparkling chemicals as mad scientists in a steampunk style

DALL-E 2 - 2022

Playing Atari with Deep Reinforcement Learning - 2013

OpenAI Five - Dota 2 strategy game - 2019

Mastering the game of Go with deep neural networks and tree search - 2016

AlphaDogFight - 2020

GAN PROGRESS ON FACE GENERATION

Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021; Vahdat et al., 2021

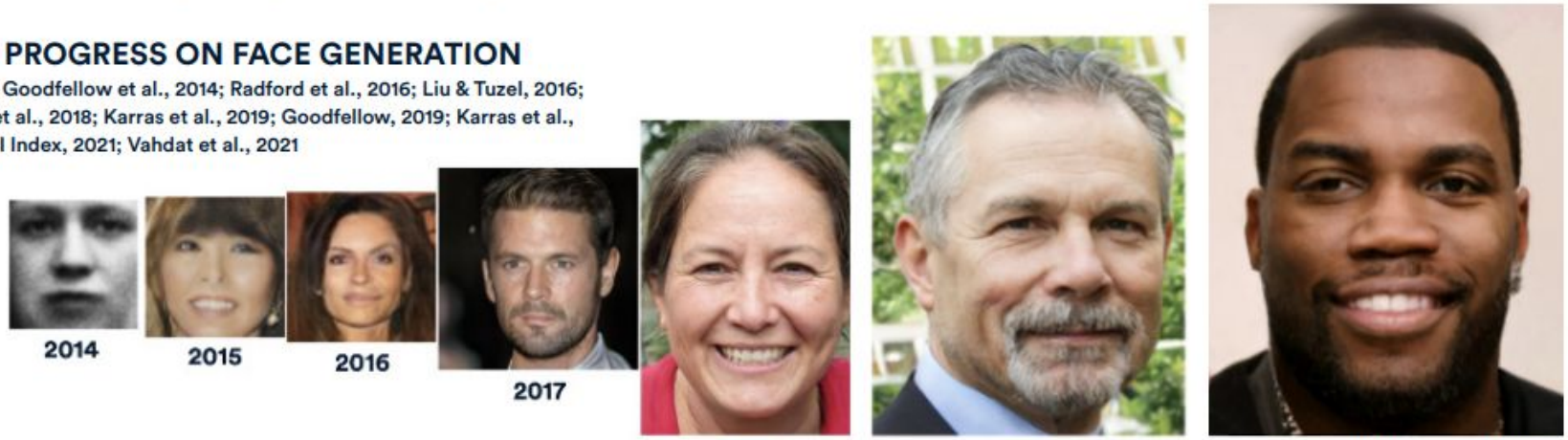


Figure 2.1.4

Why are we here?

arXiv: [Scaling Laws for Neural Language Models]

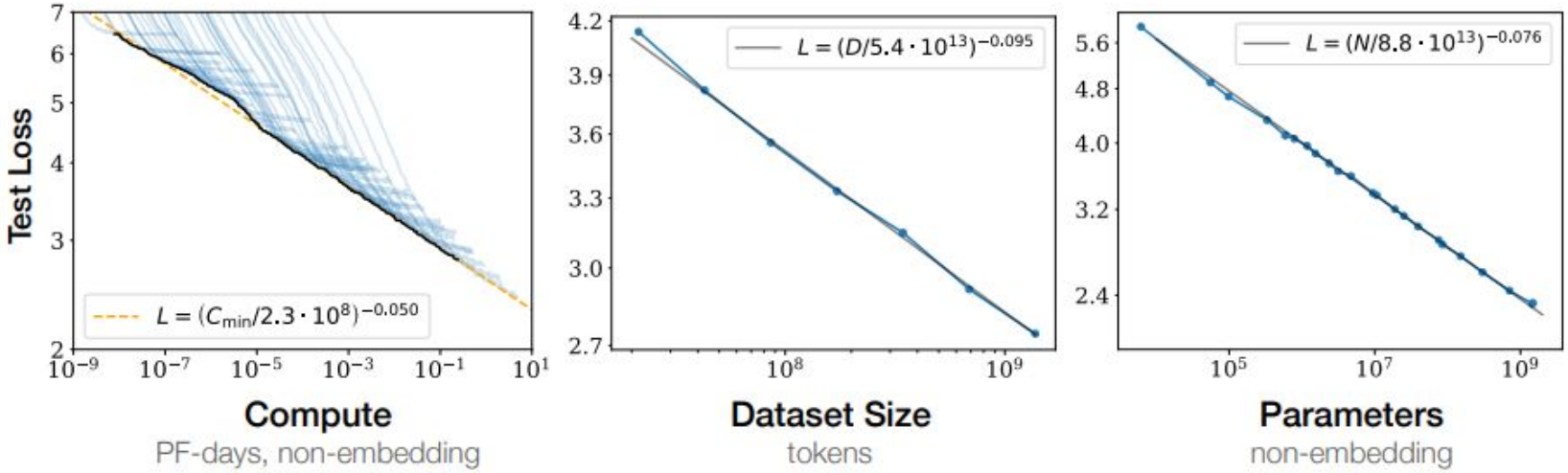


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Why are we here?

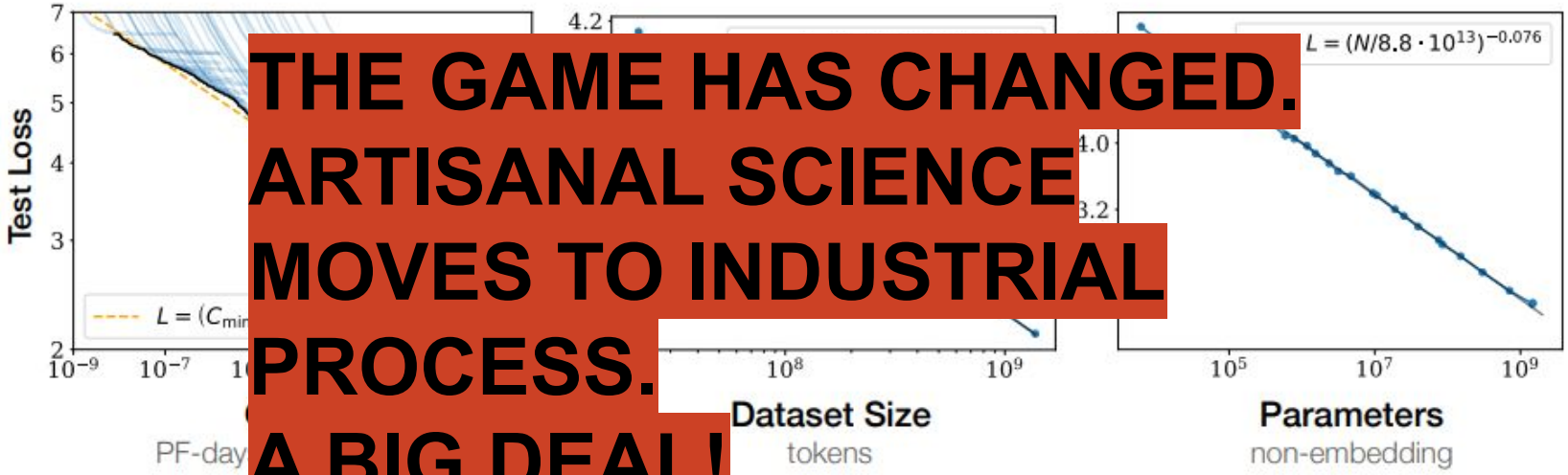


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

One Big Bug

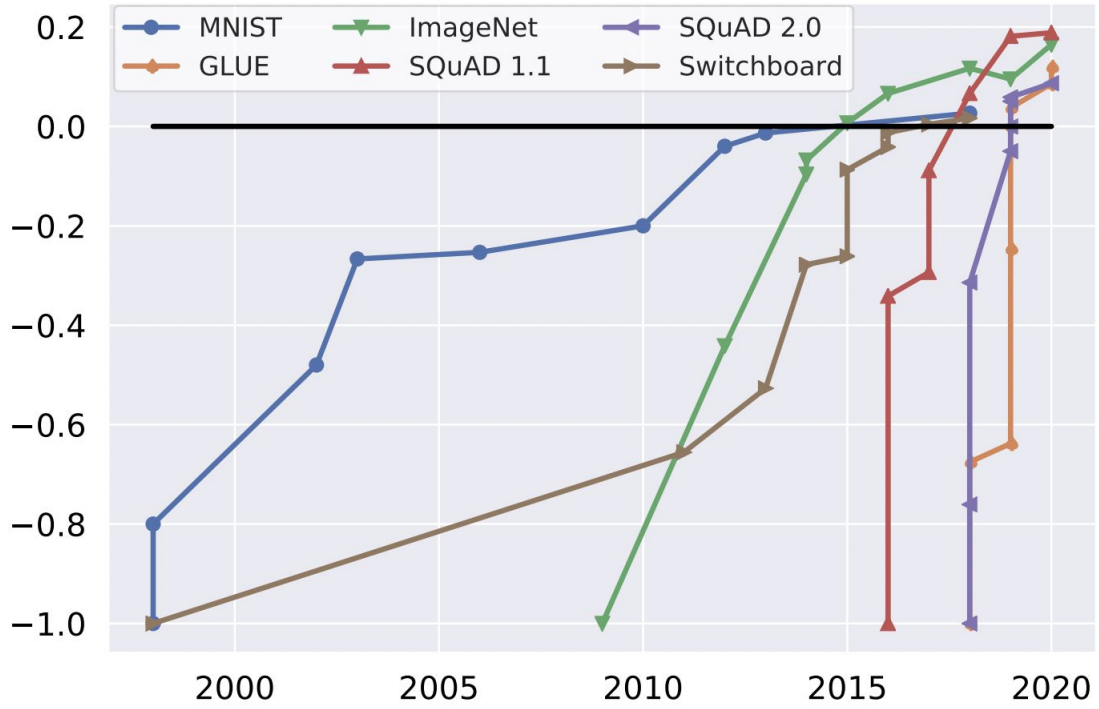


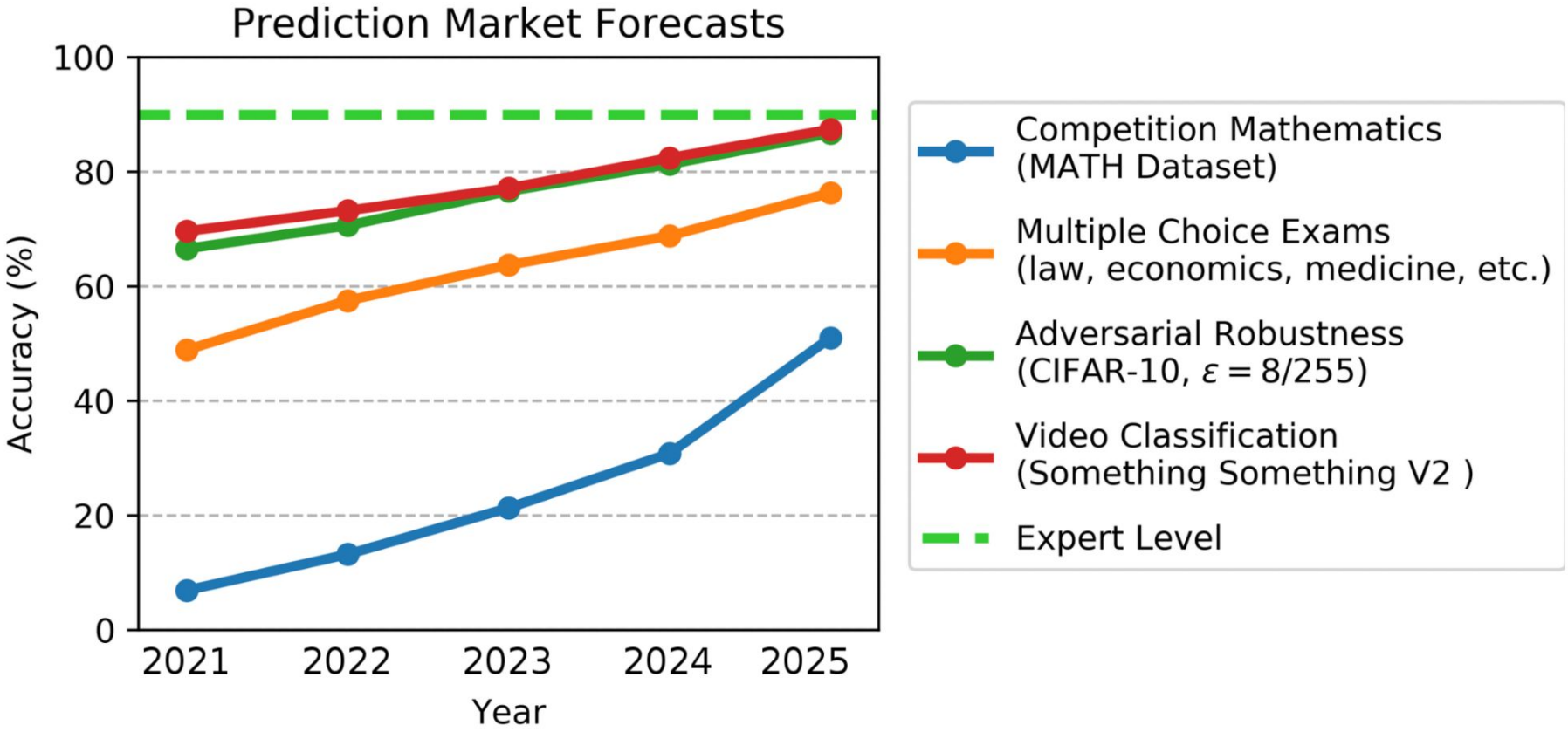
Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

What happened over the past year?

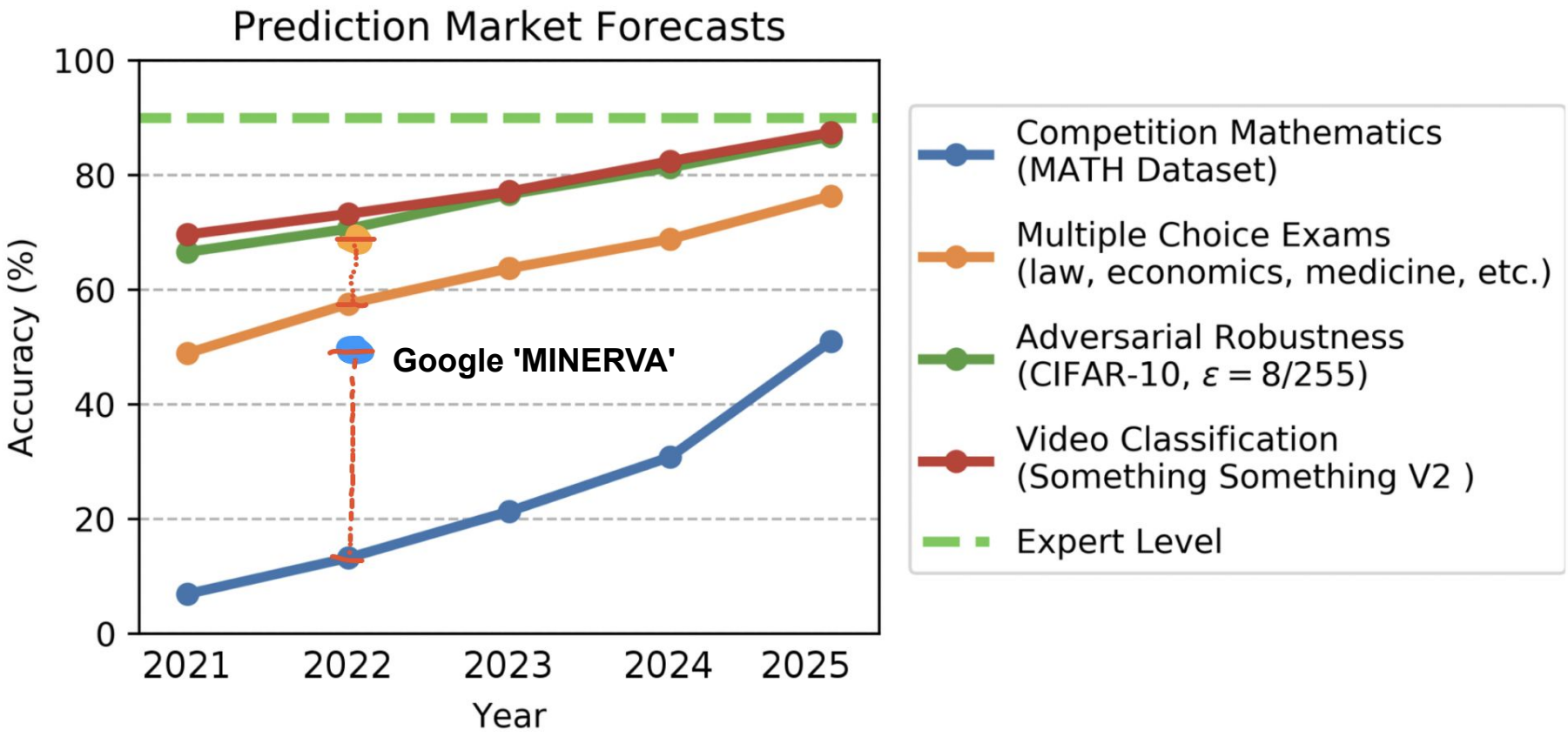
The big bug is getting worse.

Appearance of large-scale generative models, sometimes with RL, leading to increasingly capable and rich systems.

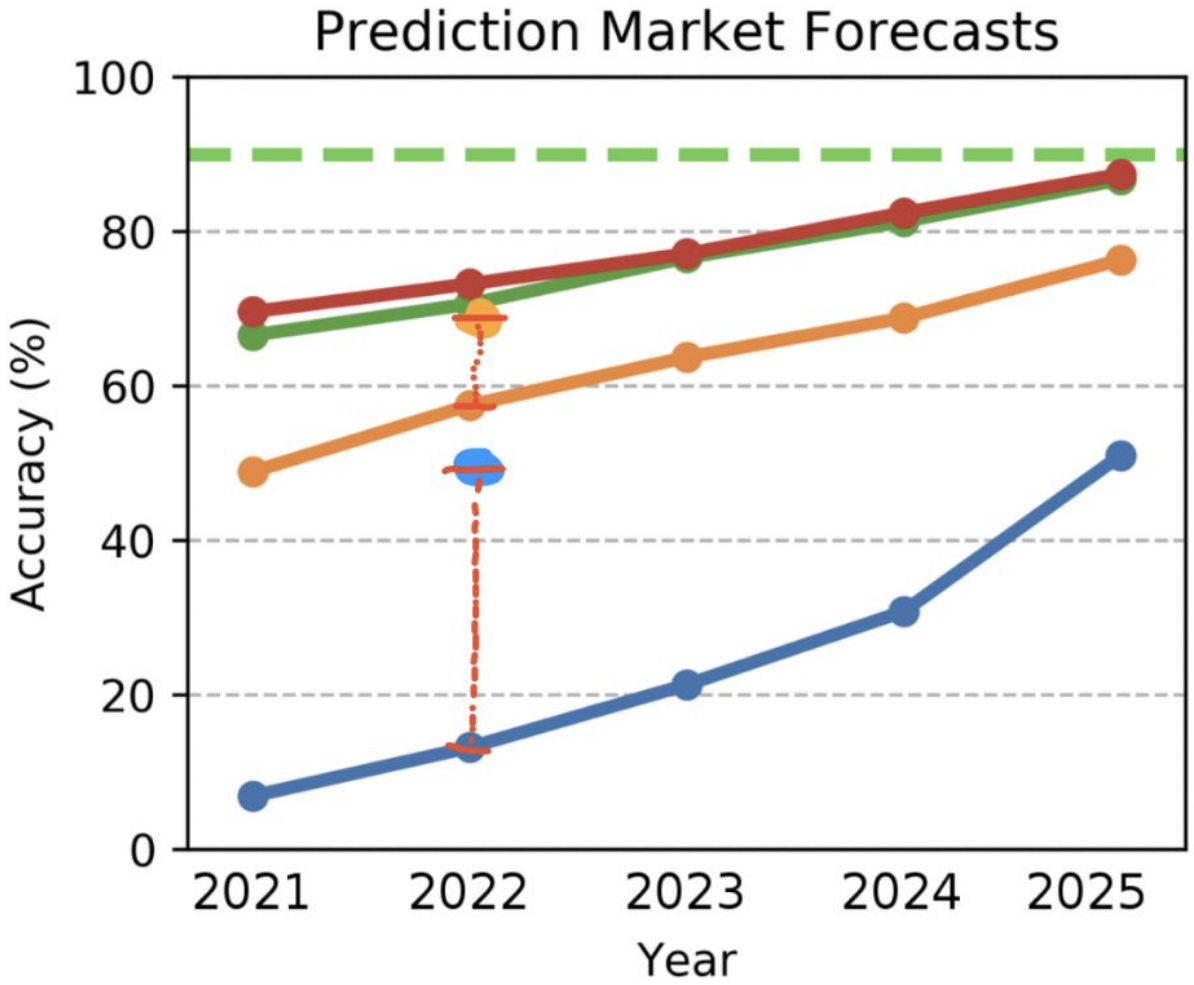
- GATO (flexible general agent)
- Dall-E-2 / CogView / Imagen (flexible image generation)
- Copilot (flexible code generation)
- LM++, e.g, Palm, Anthropic RLHF assistant (flexible text models)



From: https://bounded-regret.ghost.io/ai-forecasting-one-year-in/



From: https://bounded-regret.ghost.io/ai-forecasting-one-year-in/



What is happening here?

What forces are driving this?

What does combo of bad benchmarks and bad predictions mean for future?

In what other parts of science has progress beat expectations repeatedly?

Open problems

- How can we better measure progress, given rate of benchmark invalidation?
- We KNOW these models can augment people (e.g., code models). What's the best way to measure improved productivity?
- How can we create better ways to forecast this progress?
- What's the correct way to measure a multi-tool like a language model?