A few technical points to keep in mind when discussing technologies like ChatGPT

and a few hypotheses

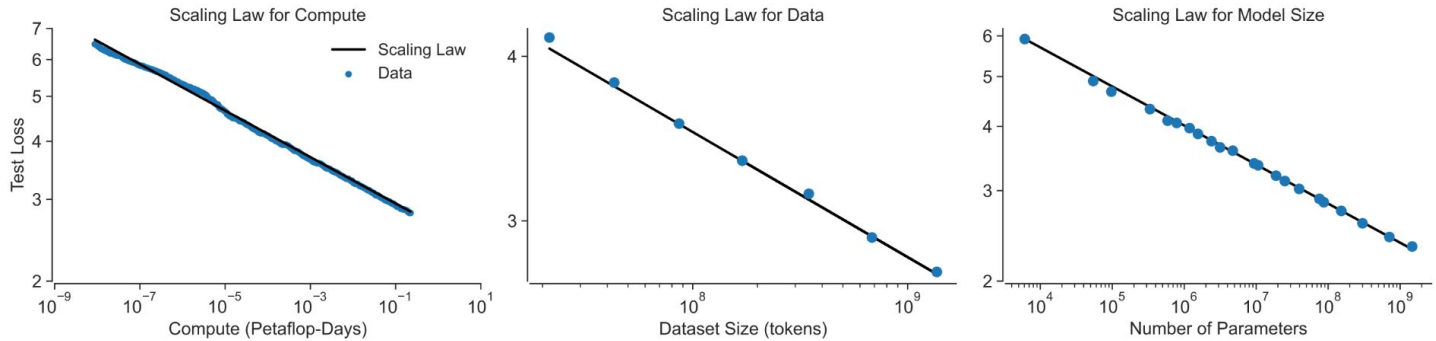
NYU & Anthropic

Sam Bowman
@sleepinyourhat

Large language models follow some shockingly predictable trends

Huge shift in language technology ca. 2019–2020:

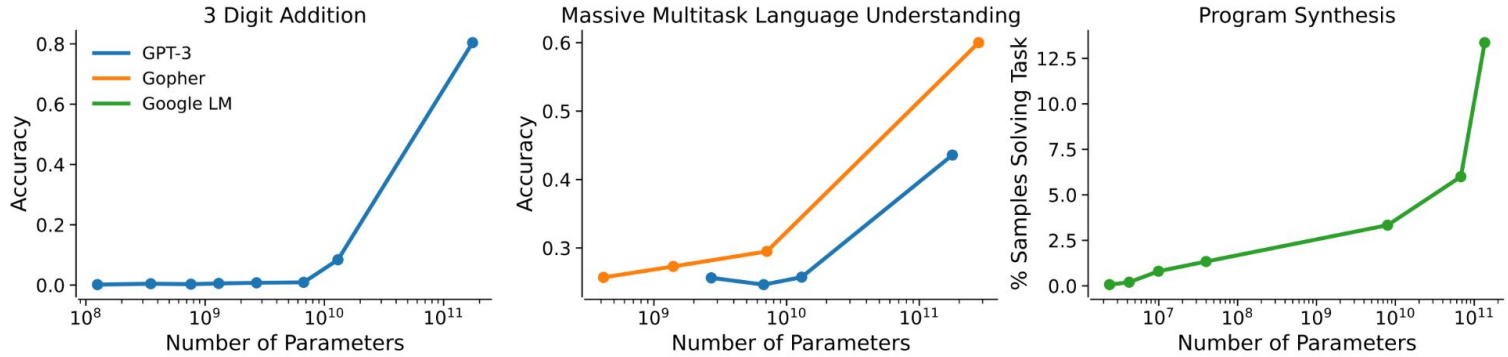
> There is now a recipe that predictably turns computer time (i.e., money) and public data into commercially valuable technical advances.



Kaplan et al. ‘20, Ganguli et al. ‘22

Large language models follow some shockingly predictable trends, **but many important developments are deeply *un*predictable**

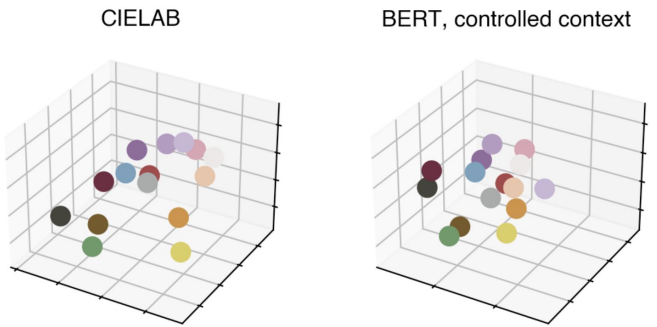
These trends correlate closely enough with market value that they drive investment, but they don't let you predict what systems will be able to do.



Kaplan et al. '20, Ganguli et al. '22

Large language models follow some shockingly predictable trends, but many important developments are deeply unpredictable, **and there aren't clear limits to the behaviors we should expect to see.**

There's evidence that these systems are starting to develop internal models of the world, including some theory of mind, notions of whether statements are true, and representations of color and geometry.



Abdou et al. '21, Andreas et al. '22, Wei et al. '22, Burns et al. '22, Kosinski '23, Patel & Pavlick '23

Large language models follow some shockingly predictable trends, but many important developments are deeply unpredictable, and there aren't clear limits to the behaviors we should expect to see. **Making these models use their capabilities to *do what we want* is a difficult open problem.**

Models' increasing ability to *reason about the environments they're operating in* makes this problem harder in other ways.

Zhao et al. '21, Ngo et al. '22, Perez et al. '22

Large language models follow some shockingly predictable trends, but many important developments are deeply unpredictable, and there aren't clear limits to the behaviors we should expect to see. Making these models use their capabilities to *do what we want* is a difficult open problem.

A Few Hypotheses

- Expect to see power and capital within tech consolidate further as the most capable models grow exponentially in price.
- Expect a few important harms from current systems to go away on their own. Many more will get much worse without difficult interventions.
- Many of the biggest impacts from this technology in five years will look very unlike what we see with today's ChatGPT. Some will have geopolitical consequences. This makes regulation especially thorny.

Thanks!

Large language models follow some shockingly predictable trends, but many important developments are deeply unpredictable, and there aren't clear limits to the behaviors we should expect to see. Making these models use their capabilities to do what we want is a difficult open problem.