ANTHROP\C

Predictability and Surprise in Large Generative Models

This memo is a summary of research conducted at Anthropic. Ganguli, D., et al. (2022). Predictability and Surprise in Large Generative Models. Association for Computing Machinery. https://dl.acm.org/doi/abs/10.1145/3531146.3533229

In recent years, AI researchers have built far more general and useful AI systems. The best current examples are models that can process and generate text (e.g. OpenAI's *GPT-3*, Microsoft & NVIDIA's *Megatron-Turing NLG*, and DeepMind's *Chinchilla*). Unlike previous AI systems designed to perform a single task, these models can be used in a variety of applications, without being explicitly trained for those purposes.

These models share an unusual combination of interrelated characteristics. In some respects they are reliably predictable, while in other ways, they are quite unpredictable. Aggregate model performance follows a *predictable* trend in relation to the resources expended on training. **By scaling up the size of models, the computing power (compute) used to train them, and the amount of data they're trained on (in the correct proportions), models demonstrate improved general performance in a predictable manner (a trend referred to as a "scaling law," Fig 1.1).**

On the other hand, individual capabilities and outputs can't be predicted ahead of time — developers can't tell you precisely what new behaviors will emerge as they scale up models. For example, the ability to complete a specific task can sometimes emerge abruptly as developers increase the size of a model (Fig 1.2). The unpredictable nature of these models makes it difficult to fully account for the consequences of their development and deployment, demonstrating the importance of empirical safety research.

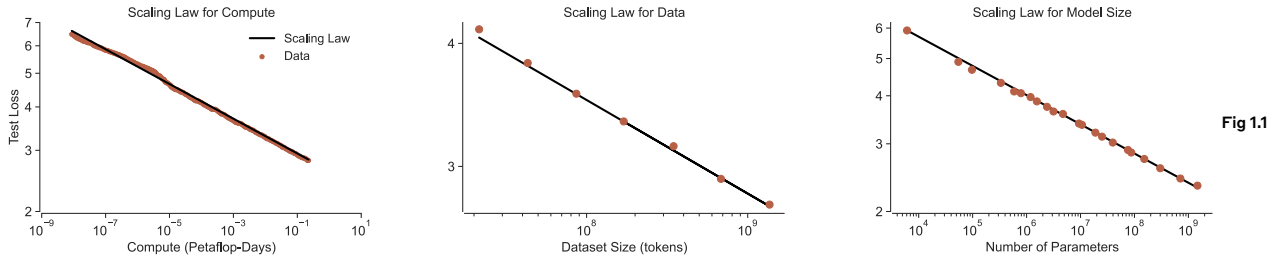


Fig 1.1

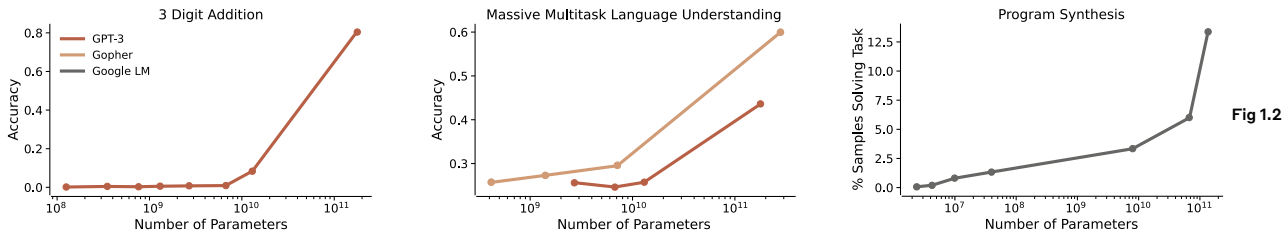


Fig 1.2

Despite these risks, more organizations from around the world are developing these systems, as a significant amount of the uncertainty inherent to AI development has been reduced (Fig 1.3).

Organizations (or countries) may also be motivated by the scientific potential these models create for novel AI research or the prestige associated with being on the technological frontier.

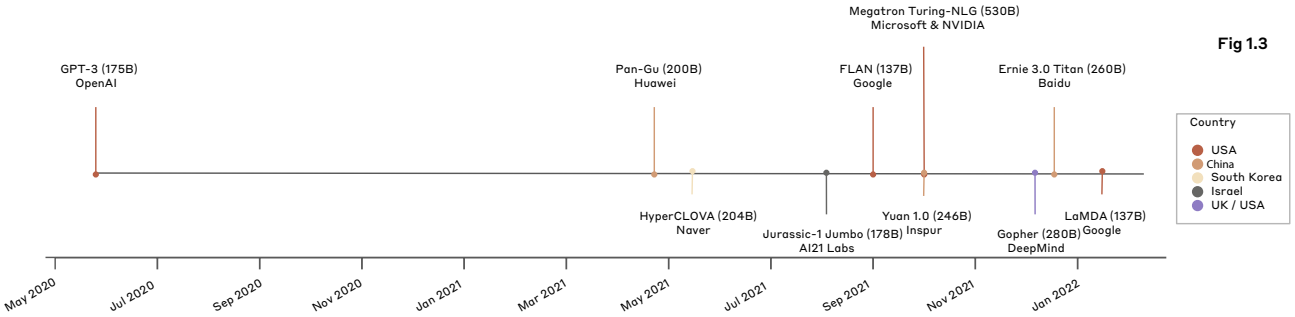
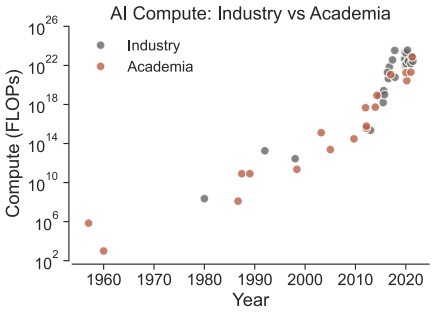


Fig 1.3

Due to the expense involved in building these models and the technical talent required to engineer them, **private sector organizations from a number of countries are the ones building these models – not academia, civil society, or public sector organizations.** Over the past decade, we've seen a dramatic shift in AI R&D as more computationally-intensive research is conducted by private sector actors, while academia increasingly lags in its ability to build or investigate models at the frontier (Fig 1.4).



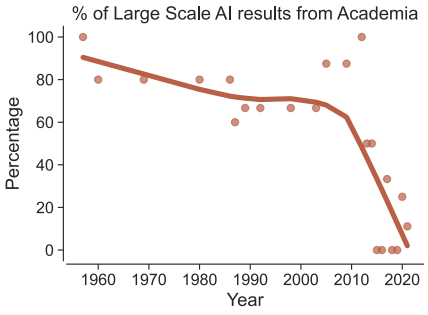


Fig 1.4

Based on the distinguishing features of these models, and the economic motivations for their development and deployment, we predict large generative models will increasingly be developed and deployed despite their potential for harm. We think there are some policy interventions available that can increase the chance of these models being developed and deployed in positive ways:

- Reduce compute asymmetries between the private sector and academia
- Improve the technical tools publicly available for model evaluation
- Increase our understanding of abrupt jumps in capabilities
- Improve shared knowledge of how to "red team" models
- Explore and prototype novel governance structures and government interventions