BUZZ CHRONICLES > ALL
Saved by @ankitsrihbti
See On Twitter

Twitter Thread by Owain Evans

Owain Evans
@OwainEvans_UK

Thread on @AnthropicAI's cool new paper on how large models are both predictable (scaling laws) and surprising (capability jumps).
1. That there’s a capability jump in 3-digit addition for GPT3 (left) is unsurprising. Good challenge to better predict when such jump will occur.

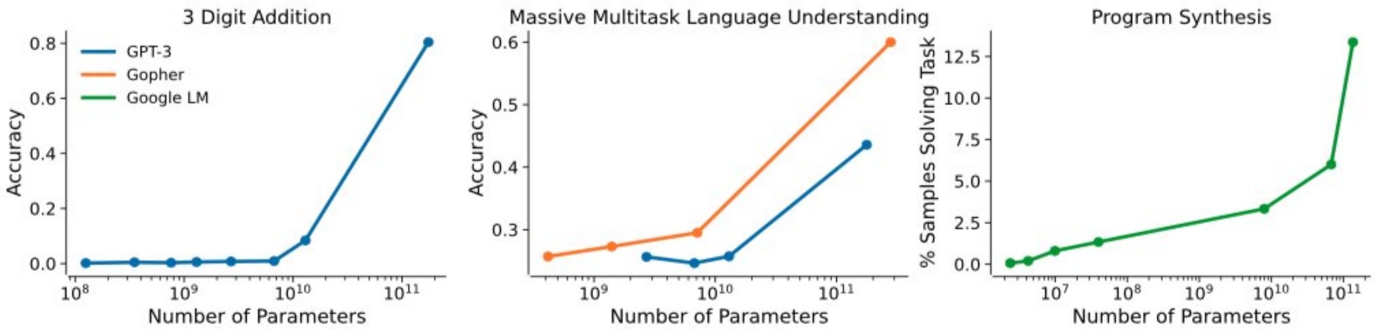

Figure 2 Three examples of abrupt specific capability scaling described in Section 2.2, based on three different models: GPT-3 (blue), Gopher (orange), and a Google language model (green). **(Left)** 3-Digit addition with GPT-3 [11]. **(Middle)** Language understanding with GPT-3 and Gopher [56]. **(Right)** Program synthesis with Google language models [4].

2. The MMLU capability jump (center) is very different b/c it’s many diverse knowledge questions with no simple algorithm like addition.
This jump is surprising and I’d like to understand better why it happens at all.

3. Program Synthesis jump (right) feels like it should be in between 1 and 2. Less diversity than 2 and we can also imagine models grokking certain concepts in programming leading to a jump.

I’d love to see more work on this topic of predictability and surprise and how they relate to forecasting alignment/risk.
Related work:
1. @gwern's list of capability jumps and classic article on scaling
https://t.co/C8qljJa13A
https://t.co/er791mhbjP

2. @JacobSteinhardt's insightful blog series. https://t.co/0ckLgOgBiV
3. Lukas Finnveden's post on GPT-n extrapolation /scaling on different task

shttps://www.lesswrong.com/posts/k2SNji3jXaLGhBeYP/extrapolating-gpt-n-performance