Review of Epoch's *Scaling transformative autoregressive models*

Nuño Sempere

2023-04-11

We want to forecast the arrival of human-level AI systems. This is a complicated task, and previous attempts have been kind of mediocre. So this paper proposes a new approach.

The approach has some key assumptions. And then it needs some auxiliary hypotheses and concrete estimates flesh out those key assumptions. Its key assumptions are:

- That a sufficient condition for reaching human-level performance might be indistinguishability: if you can't determine whether a git repository was produced by an expert human programmer or by an AI, this should be a sufficient (though not necessary) demonstration for the AI to have acquired the capability of programming.
- That models' performance will continue growing as predicted by current scaling laws.

Then, note that today's language models are in fact trained to mimic human text. And note that the error that their training method aims to minimize can be decomposed into some irreducible entropy part, and some part in which is due to the model not yet being a good enough mimic. So then, we can ask, when will a language models' loss be low enough that it is approximating human text enough that we can be very sure that it has acquired enough human capabilities that the model will have transformative effects?

For example, if a model is able to produce scientific papers such that it takes multiple papers to distinguish that model from a talented human scientist, then this would be enough to conclude that a model has acquired the capability of doing scientific research as good as that of the talented human scientist.

My high level thought is that this is an elegant approach. It warms my blackened heart a bit. I'm curious about why previous attempts, e.g., various reports commissioned by Open Philanthropy, didn't think of it in some shape.

One thought though, is that the scaling laws that they are referring to do have a large number of degrees of freedom. Their shape is:

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

and we notice that there are four degrees of freedom (A,B, alpha and beta), as well as the choice of overall shape of the formula. The E parameter represents irreducible uncertainty, but also seems to be only empirically estimatable by estimating the other parameters, and thus also seems to take the role of a dangling parameter. I am reminded of the quote "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk". I don't really have strong opinions about whether scaling laws are overfitting on past data. But I do think that most of the uncertainty is going to come from uncertainty about whether scaling laws will continue as they have. Still, having a conditional forecast based on scaling laws continuing is fine.

Anyways, then to flesh out their approach, the authors need some auxiliary assumptions. Some which I could notice are:

1. That at the moment when a rater exceeds some threshold of certainty when trying to catch a model vs a human, their certainty will be very close to their threshold of certainty.
2. That human language is ergodic.
3. That a human trying to distinguish between a mimic model's output and that of a human would exhibit a particular mathematical shape, where a human moves in the same direction as an ideal bayesian reasoner each step, just more slowly.

Honestly, these seem like reasonable analytical approximations. But I wish there had been some discussion of when language cannot be ergodic. What I'm thinking is, that it does sound like a reasonable approximation, but that I haven't thought much about it, and that maybe there is a catch, like texts containing novel scientific discoveries maybe not being ergodic? Also, ergodic over what domain? The domain of texts which could be written, or will ever be written? But then language models don't have future or counterfactual texts in their training data? I don't know, and I haven't thought much about ergodicity either, so I'm uncertain here.

I also found assumption number 3. to be irritating, because the model assumes that human judges have a "slowdown" factor in their updates, but then humans could notice that, estimate their slowdown factor, and then update faster. My guess is that instead of, or in addition to a slowdown factor, humans have an error rate, where they process evidence as pointing to one side that an ideal reasoner would process that same evidence as pointing to another side.

I also didn't really like assumption number 1, about the endline uncertainty when waiting for enough evidence to exceed a threshold of uncertainty being close to that threshold. That assumption wouldn't be the case if, for example, most of the outputs of a model where nearly indistinguishable from a human,

but it infrequently had some clear tells or artifacts, like inexact human hands in some arts models or unnatural moves in chess or in go. I also

But overall, despite being irritated by these auxiliary assumptions, I think they are generally a reasonable move. I don't really expect assumptions 1 and 2 to introduce much error. I'm a bit more wary about assumption number two, because as I said I haven't thought much about ergodicity, but it seems respectable to add more analytical clarity at the cost of wrapping the whole result in another conditional (… if language is such that it can be reasonably modelled as ergodic, then…). Also, the alternative might be to explore each tendril down the forking paths road, but this would be too effortful.

To conclude the analysis, the authors need some forecasts about how indistinguishable AI-produce texts have to be before they are transformative, operationalized as how many tokens one would have to see until one can be confident that they are in fact produced by an AI. Their given forecasts don't sem particularly objectionable, thought that I haven't had enough time to sit down and examine them carefully. More generally though, once the authors have outlined an analytical approach, it doesn't seem that difficult to commission forecasts to feed into their model.

As for suggestions, my main one would be to add more analytical clarity, and to be a bit obsessive about what analytical assumptions, key assumptions or otherwise, are going into the model. Maybe have a chart. If one does this, then the shape of what the paper produces is, I think:

- A conditional forecast
- of an upper bound of compute needed
- to reach transformative AI

The forecast is conditional on scaling laws continuing as they have, and on the various analytical assumptions not introducing too much error. And it's not a forecast of when models will be transformative, but of an upper bound, because as we mentioned at the beginning, indistinguishability is a sufficient but not a necessary condition for transformative AI. The authors point this out at the beginning, but I think this could be pointed out more obviously.

The text also generally needs an editor (e.g., use the first person plural, as there are two authors). As I was reading it, I felt the compulsion to rewrite it in better prose. But I didn't think that it was worth it for me to do that, or tpoint out style mistakes—besides my wish for greater clarity—because you can just hire an editor for this. And also, an alert reader should be able to extract the core of what you are saying even though prose could be improved. I did write down some impressions as I was reading in a different document, though.

Overall I liked it, and would recommend that it be published. It's the kind of thing that, even if one thinks that it is not enough for a forecast on its own, seems like it would be a valuable input into other forecasts.