Is AI progress slowing down?

Making sense of recent technology trends and claims

ARVIND NARAYANAN AND SAYASH KAPOOR
DEC 18, 2024

By Arvind Narayanan, Benedikt Ströbl, and Sayash Kapoor.

After the release of GPT-4 in March 2023, the dominant narrative in the tech world was that continued scaling of models would lead to artificial general intelligence and then superintelligence. Those extreme predictions gradually receded, but up until a month ago, the prevailing belief in the AI industry was that model scaling would continue for the foreseeable future.

Then came three back-to-back news reports from The Information, Reuters, and Bloomberg revealing that three leading AI developers — OpenAI, Anthropic, and Google Gemini — had all run into problems with their next-gen models. Many industry insiders, including Ilya Sutskever, probably the most notable proponent of scaling, are now singing a very different tune:

> "The 2010s were the age of scaling, now we're back in the age of wonder and discovery once again. Everyone is looking for the next thing," Sutskever said. "Scaling the right thing matters more now than ever." (Reuters)

The new dominant narrative seems to be that model scaling is dead, and "inference scaling", also known as "test-time compute scaling" is the way forward for improving AI capabilities. The idea is to spend more and more computation when using models to perform a task, such as by having them "think" before responding.

This has left AI observers confused about whether or not progress in AI capabilities is slowing down. In this essay, we look at the evidence on this question, and make four main points:

1. Declaring the death of model scaling is premature.
2. Regardless of whether model scaling will continue, industry leaders' flip flopping on this issue shows the folly of trusting their forecasts. They are not significantly better informed than the rest of us, and their narratives are heavily influenced by their vested interests.
3. Inference scaling is real, and there is a lot of low-hanging fruit, which could lead to rapid capability increases in the short term. But in general, capability improvements from inference scaling will likely be both unpredictable and unevenly distributed among domains.

4. The connection between capability improvements and AI's social or economic impacts is extremely weak. The bottlenecks for impact are the pace of product development and the rate of adoption, not AI capabilities.

Is model scaling dead?

There is very little new information that has led to the sudden
We've long been saying on this newsletter that there are impor
to model scaling. Just as we cautioned back then about scaling
now caution against excessive pessimism about model scaling.

"Scaling as usual" ended with GPT-4 class models, because the
trained on most of the readily available data sources. We alread
new ideas would be needed to keep model scaling going. So un
evidence that many such ideas have been tried and failed, we c
that there isn't more mileage to model scaling.

As just one example, it is possible that including YouTube videc
videos, not transcribed text — in the training mix for multimodal models will unlock new capabilities. Or it might not help; we just won't know until someone tries it, and we don't know if it has been tried or not. Note that it would probably have to be Google, because the company is unlikely to license YouTube training data to competitors.[1]

If things are still so uncertain regarding model scaling, why did the narrative flip? Well, it's been over two years since GPT-4 finished training, so the idea that next-gen models are simply taking a bit longer than expected was becoming less and less credible. And once one company admits that there are problems, it becomes a lot easier for others to do so. Once there is a leak in the dam, it quickly bursts. Finally, now that OpenAI's reasoning model o1 is out, it has given companies an out when admitting that they have run into problems with model scaling, because they can save face by claiming that they will simply switch to inference scaling.

To be clear, there is no reason to doubt the reports saying that many AI labs have conducted larger training runs and yet not released the resulting models. But it is less clear what to conclude from it. Some possible reasons why bigger models haven't been released include:

- Technical difficulties, such as convergence failures or complications in achieving fault tolerance in multi-datacenter training runs.
- The model was not much better than GPT-4 class models, and so would be too underwhelming to release.
- The model was not much better than GPT-4 class models, and so the developer has been spending a long time trying to eke out better

performance through fine tuning.

To summarize, it's possible that model scaling has indeed reached its limit, but it's also possible that these hiccups are temporary and eventually one of the companies will find ways to overcome them, such as by fixing any technical difficulties and/or finding new data sources.

Let's stop deferring to insiders

Not only is it strange that the new narrative emerged so quickly, it's also interesting that the old one persisted for so long, despite the potential limitations of model scaling being obvious. The main reason for its persistence is the assurances of industry leaders that scaling would continue for a few more years. [2] In general, journalists (and most others) tend to defer to industry insiders over outsiders. But is this deference justified?

Industry leaders don't have a good track record of predicting AI developments. A good example is the overoptimism about self-driving cars for most of the last decade. (Autonomous driving is finally real, though Level 5 — full automation — doesn't exist yet.) As an aside, in order to better understand the track record of insider predictions, it would be interesting to conduct a systematic analysis of all predictions about AI made in the last 10 years by prominent industry insiders.

There are some reasons why we might want to give more weight to insiders' claims, but also important reasons to give *less* weight to them. Let's analyze these one by one. It is true that industry insiders have proprietary information (such as the performance of as-yet-unreleased models) that might make their claims about the future more accurate. But given how many AI companies are close to the state of the art, including some that openly release model weights and share scientific insights, datasets, and other artifacts, we're talking about an advantage of at most a few months, which is minor in the context of, say, 3-year forecasts.

Besides, we tend to overestimate how much additional information companies have on the inside — whether in terms of capability or (especially) in terms of safety. Insiders warned for a long time that "if only you know what we know..." but when whistleblowers finally came forward, it turns out that they were mostly relying on the same kind of speculation that everyone else does. [3]

Another potential reason to give more weight to insiders is their technical expertise. We don't think this is a strong reason: there is just as much AI expertise in academia as in industry. More importantly, deep technical expertise isn't that important to support the kind of crude trend extrapolation that goes into AI forecasts. Nor is technical expertise enough — business and

social factors play at least as big a role in determining the course of AI. In the case of self-driving cars, one such factor is the extent to which societies tolerate public roads being used for experimentation. In the case of large AI models, we've argued before that the most important factor is whether scaling will make business sense, not whether it is technically feasible. So not only do techies not have much of an advantage, their tendency to overemphasize the technical dimensions tends to result in overconfident predictions.

In short, the reasons why one might give more weight to insiders' views aren't very important. On the other hand, there's a huge and obvious reason why we should probably give less weight to their views, which is that they have an incentive to say things that are in their commercial interests, and have a track record of doing so.

As an example, Sutskever had an incentive to talk up scaling when he was at OpenAI and the company needed to raise money. But now that he heads the startup Safe Superintelligence, he needs to convince investors that it can compete with OpenAI, Anthropic, Google, and others, despite having access to much less capital. Perhaps that is why he is now talking about running out of data for pre-training, as if it were some epiphany and not an endlessly repeated point.

To reiterate, we don't know if model scaling has ended or not. But the industry's sudden about-face has been so brazen that it should leave no doubt that insiders don't have any kind of crystal ball and are making similar guesses as everyone else, and are further biased by being in a bubble and readily consuming the hype they sell to the world.

In light of this, our suggestion — to everyone, but especially journalists, policymakers, and the AI community — is to end the deference to insiders' views when they predict the future of technology, especially its societal impacts. This will take effort, as there is a pervasive unconscious bias in the U.S., in the form of a "distinctly American disease that seems to equate extreme wealth, and the power that comes with it, with virtue and intelligence." (from Bryan Gardiner's review of Marietje Schake's The Tech Coup.)

Will progress in capabilities continue through inference scaling?

Of course, model scaling is not the only way to improve AI capabilities. Inference scaling is an area with a lot of recent progress. For example, OpenAI's o1 and the open-weights competitor DeepSeek R1 are reasoning models: they have been fine tuned to "reason" before providing an answer. Other methods leave the model itself unchanged but employ tricks like generating many solutions and ranking them by quality.

There are two main open questions about inference scaling that will determine how significant of a trend it will be.

1. What class of problems does it work well on?
2. For problems where it does work well, how much of an improvement is possible by doing more computation during inference?

The per-token output cost of language models has been rapidly decreasing due to both hardware and algorithmic improvements, so if inference scaling yields improvements over many orders of magnitude — for example, generating a million tokens on a given task yields significantly better performance than generating a hundred thousand tokens — that would be a big deal. [4]

The straightforward, intuitive answer to the first question is that inference scaling is useful for problems that have clear correct answers, such as coding or mathematical problem solving. In such tasks, at least one of two related things tend to be true. First, symbolic reasoning can improve accuracy. This is something LLMs are bad at due to their statistical nature, but can overcome by using output tokens for reasoning, much like a person using pen and paper to work through a math problem. Second, it is easier to verify correct solutions than to generate them (sometimes aided by external verifiers, such as unit tests for coding or proof checkers for mathematical theorem proving).

In contrast, for tasks such as writing or language translation, it is hard to see how inference scaling can make a big difference, especially if the limitations are due to the training data. For example, if a model works poorly in translating to a low-resource language because it isn't aware of idiomatic phrases in that language, the model can't reason its way out of this.

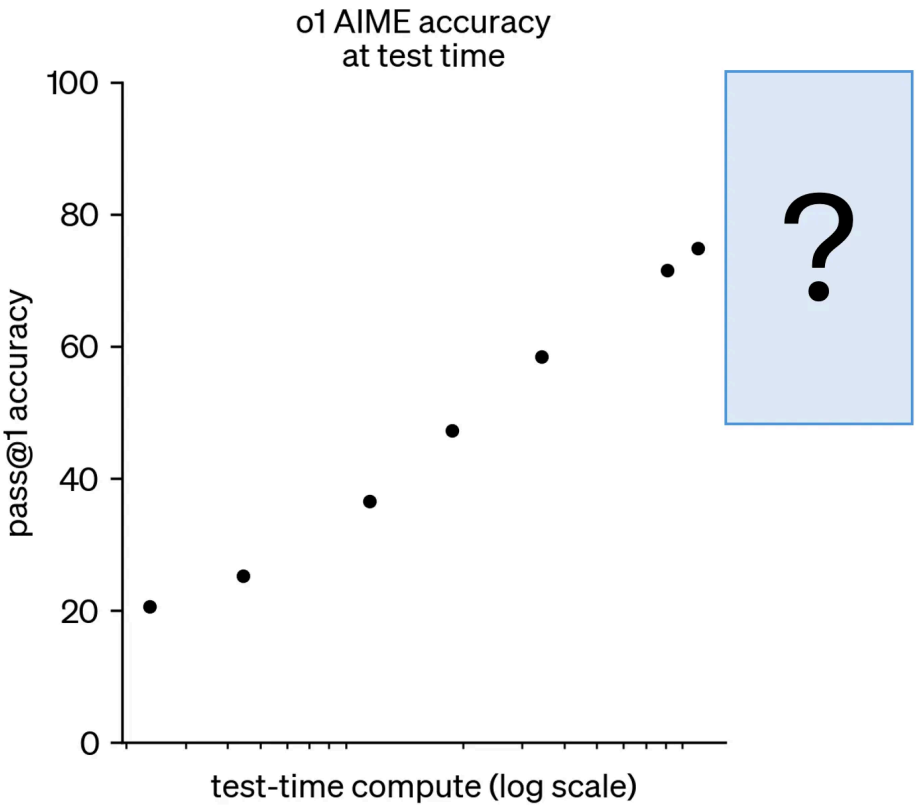
The early evidence we have so far, while spotty, is consistent with this intuition. Focusing on OpenAI o1, it improves compared to state-of-the-art language models such as GPT-4o on coding, math, cybersecurity, planning in toy worlds, and various exams. Improvements in exam performance seem to strongly correlate with the importance of reasoning for answering questions, as opposed to knowledge or creativity: big improvements for math, physics and LSATs, smaller improvements for subjects like biology and econometrics, and negligible improvement for English.

Tasks where o1 doesn't seem to lead to an improvement include writing, certain cybersecurity tasks (which we explain below), avoiding toxicity, and an interesting set of tasks at which thinking is known to make humans *worse*.

We have created a webpage compiling the available evidence on how reasoning models compare against language models. We plan to keep it updated for the time being, though we expect that the torrent of findings will soon become difficult to keep up with.

Now let's consider the second question: how large of an improvement can we get through inference scaling, assuming we had an infinite inference compute budget.

OpenAI's flagship example to show off o1's capabilities was AIME, a math benchmark. Their graph leaves this question tantalizingly open. Is the performance about to saturate, or can it be pushed close to 100%? Also note that the graph conveniently leaves out x-axis labels.



Source: OpenAI

An attempt by external researchers to reconstruct this graph shows that (1) the cutoff for the x-axis is probably around 2,000 tokens, and (2) when o1 is asked to think longer than this, it doesn't do so. So the question remains unanswered, and we need to wait for experiments using open-source models

to get more clarity. It is great to see that there are vigorous efforts to publicly reproduce the techniques behind o1.

In a recent paper called *Inference Scaling fLaws* (the title is a pun on inference scaling laws), we look at a different approach to inference scaling — repeatedly generating solutions until one of them is judged as correct by an external verifier. While this approach has been associated with hopes of usefully increasing scaling by many orders of magnitude (including by us in our own past work), we find that this is extremely sensitive to the quality of the verifier. If the verifier is slightly imperfect, in many realistic settings of a coding task, performance maxes out and actually starts to *decrease* after about 10 attempts.

Generally speaking, the evidence for inference scaling "laws" is not convincing, and it remains to be seen if there are real-world problems where generating (say) millions of tokens at inference time will actually help.

Is inference scaling the next frontier?

There is a lot of low-hanging fruit for inference scaling, and progress in the short term is likely to be rapid. Notably, one current limitation of reasoning models is that they don't work well in agentic systems. We have observed this in our own benchmark CORE-Bench that asks agents to reproduce the code provided with research papers — the best performing agent scores 38% with Claude 3.5 Sonnet compared to only 24% with o1-mini. [5] This also explains why reasoning models led to an improvement in one cybersecurity eval but not another — one of them involved agents.

We think there are two reasons why agents don't seem to benefit from reasoning models. Such models require different prompting styles than regular models, and current agentic systems are optimized for prompting regular models. Second, as far as we know, reasoning models so far have *not* been trained using reinforcement learning in a setting where they receive feedback from the environment — be it code execution, shell interaction, or web search. In other words, their tool use ability is no better than the underlying model before learning to reason. [6]

These seem like relatively straightforward problems. Solving them might enable significant new AI agent capabilities — for example, generating complex, fully functional apps from a prompt. (There are already tools that try to do this, but they don't work well.)

But what about the long run? Will inference scaling lead to the same kind of progress we've seen with model scaling over the last 7 years? Model scaling

was so exciting because you "merely" needed to make data, model size, and compute bigger; no algorithmic breakthroughs were needed.

That's not true (so far) with inference scaling — there's a long list of inference scaling techniques, and what works or doesn't work is problem-dependent, and even collectively, they only work in a circumscribed set of domains. AI developers are trying to overcome this limitation. For example, OpenAI's reinforcement finetuning service is thought to be a way for the company to collect customer data from many different domains for fine-tuning a future model.

About a decade ago, reinforcement learning (RL) led to breakthroughs in many games like Atari. There was a lot of hype, and many AI researchers hoped we could RL our way to AGI. In fact, it was the high expectations around RL that led to the birth of explicitly AGI-focused labs, notably OpenAI. But those techniques didn't generalize beyond narrow domains like games. Now there is similar hype about RL again. It is obviously a very powerful technique, but so far we're seeing limitations similar to the ones that led to the dissipation of the previous wave of hype.

It is impossible to predict whether progress in AI capabilities will slow down. In fact, forget prediction — reasonable people can have very different opinions on whether AI progress has already slowed down, because they can interpret the evidence very differently. That's because "capability" is a construct that's highly sensitive to how you measure it.

What we can say with more confidence is that the *nature* of progress in capabilities will be different with inference scaling than with model scaling. In the last few years, newer models predictably brought capability improvements each year across a vast swath of domains. There was a feeling of pessimism among many AI researchers outside the big labs that there was little to do except sit around and wait for the next state-of-the-art LLM to be released.

With inference scaling, capability improvements will likely be uneven and less predictable, driven more by algorithmic advances than investment in hardware infrastructure. Many ideas that were discarded during the reign of LLMs, such as those from the old planning literature, are now back in the mix, and the scene seems intellectually more vibrant than in the last few years.

Product development lags capability increase

The furious debate about whether there is a capability slowdown is ironic, because the link between capability increases and the real-world usefulness of AI is extremely weak. The development of AI-based applications lags far behind the increase of AI capabilities, so even existing AI capabilities remain

greatly underutilized. One reason is the capability-reliability gap — even when a certain capability exists, it may not work reliably enough that you can take the human out of the loop and actually automate the task (imagine a food delivery app that only works 80% of the time). And the methods for improving reliability are often application-dependent and distinct from methods for improving capability. That said, reasoning models also seem to exhibit reliability improvements, which is exciting.

Here are a couple of analogies that help illustrate why it might take a decade or more to build products that fully take advantage of even current AI capabilities. The technology behind the internet and the web mostly solidified in the mid-90s. But it took 1-2 more decades to realize the potential of web apps. Or consider this thought-provoking essay that argues that we need to build GUIs for large language models, which will allow interacting with them with far higher bandwidth than through text. From this perspective, the current state of AI-based products is analogous to PCs before the GUI.

The lag in product development is compounded by the fact that AI companies have not paid nearly enough attention to product aspects, believing that the general-purpose nature of AI somehow grants an exemption from the hard problems of software engineering. Fortunately, this has started to change recently.

Now that they are focusing on products, AI companies as well as their users are re-discovering that software development, especially the user experience side of it, is hard, and requires a broader set of skills than AI model development. A great example is the fact that there are now two different ways to run Python code with ChatGPT (which is one of the most important capabilities for power users) and there is an intricate set of undocumented rules to remember regarding the capabilities and limitations of each of them. Simon Willison says:

> Do you find this all hopelessly confusing? I don't blame you. I'm a professional web developer and a Python engineer of 20+ years and I can just about understand and internalize the above set of rules.

Still, this is a big improvement over a week ago, when these models had powerful coding capabilities yet did not come with the ability to run code that could use the internet! And even now, o1 can neither access the internet nor run code. From the perspective of AI impacts, what matters far more than capability improvement at this point is actually building products that let people do useful things with existing capabilities.

Finally, while product development lags behind capability, the adoption of AI-based products further lags far behind product development, for various behavioral, organizational, and societal reasons. Those interested in AI's

impacts (whether positive or negative) should pay much more attention to these downstream aspects than current or predicted capabilities.

Concluding thoughts

Maybe model scaling is over; maybe not. But it won't continue forever, and the end of model scaling brings a long list of positives: AI progress once again depends on new ideas and not just compute; big companies, startups, and academic researchers can all compete on a relatively even playing field; regulation based on arbitrary training compute thresholds becomes even harder to defend; and there is a clear recognition that models themselves are just a technology, not a product.

As for the future of AI, it is clear that tech insiders are trying to figure it out just like the rest of us, and it is past time to stop trusting their overconfident, self-serving, shifting, and conveniently vague predictions. And when we move beyond technical predictions to claims about AI's impact on the world, there's even less reason to trust industry leaders.

Acknowledgment. We are grateful to Zachary S. Siegel for feedback on a draft.

---

1 While OpenAI is known to have crawled YouTube in the past, that was a small sliver of YouTube; it won't be possible to crawl all of YouTube without Google noticing.

2 A nice analysis by Epoch AI showed that scaling could continue until 2030. But this was published too recently (August 2024) to have been the anchor for the scaling narrative.

3 We are referring to *substantive* knowledge about the safety of AI models and systems; whistleblowers did bring forth new knowledge about safety-related *processes* at OpenAI.

4 That said, we can't take future cost decreases for granted; we are also running into fundamental limits of inference cost-saving techniques like quantization.

5 We set a cost limit for $4 for all models. On a small sample, with a $10 cost limit, o1-preview performed very poorly (10% accuracy). Given cost constraints, we did not evaluate the model with a higher cost limit on the entire data.

6 o1 doesn't even have access to tools during inference in the ChatGPT interface! Gemini Flash 2.0 does, but it is not clear if this is a model that has been fine tuned for reasoning, let alone fine tuned for tool use.

159 Likes · 29 Restacks

← Previous

Discussion about this post

Comments Restacks

Write a comment...

Claude Coulombe Dec 19 *Edited*

Emergent capabilities based solely on scaling are largely a myth. While scaling does extend capacity, such as with "one-shot" or "few-shot" learning, this is primarily due to a larger pattern-matching basis, inductive mechanisms (not deductive reasoning), and combinatorial effects. Currently, LLM builders are attempting to compensate for significant shortcomings with human intervention (hordes of poorly paid science students and some PhDs) to create the illusion of progress (patching AI). This is not a viable path toward Artificial General Intelligence (AGI).

LIKE (5) REPLY SHARE

Andy X Andersen Dec 18

Historically speaking, progress is slow, but it adds up.

There's likely a lot of mechanisms to understand and modeling to do. Each time we make advances it becomes more clear what problems remain and what to do next.

LIKE (3) REPLY SHARE

1 reply

10 more comments...