# How Well Does Reinforcement Learning Scale?

Toby Ord (University of Oxford)

I demonstrate (using benchmark data from OpenAI's reasoning models) that LLM performance scales much less well with more RL training compute than with more inference compute. To achieve the same gain as one gets from scaling up inference compute by 100x, RL training compute typically needs to be scaled up by 10,000x. Such RL scale-ups have been possible due to starting from a very low base (giving the impressive gains seen over the last year) but will soon be impractical. This means that most future performance gains from the RL scaling paradigm will require continued scaling of the number of reasoning tokens (and thus the deployment costs) by orders of magnitude.

The current era of improving AI capabilities using reinforcement learning (from verifiable rewards) involves two key types of scaling:

1. Scaling the amount of compute used for RL during training
2. Scaling the amount of compute used for inference during deployment

We can see (1) as training the AI in more effective reasoning techniques and (2) as allowing the model to think for longer. I'll call the first *RL-scaling*, and the second *inference-scaling*. Both new kinds of scaling were present all the way back in OpenAI's announcement of their first reasoning model, o1, when they showed this famous chart:

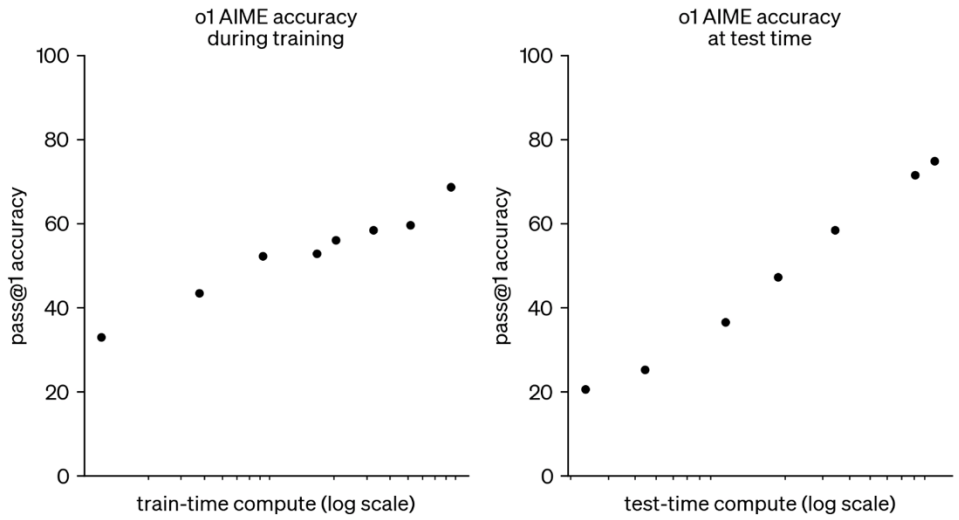


*Figure 1.* Twin graphs from the launch of o1 showing how performance improved (roughly logarithmically) with both RL-scaling and inference-scaling. Figure reproduced from OpenAI (2024).

I've previously shown (Ord 2025c) that in the initial move from a base-model to a reasoning model, most of the performance gain came from unlocking the inference-scaling. The RL training did provide a notable boost to performance, even holding the number of tokens in the chain of thought fixed. You can see this *RL boost* in the chart below as the small blue arrow on the left that takes the base model up to the trend-line for the reasoning model. But this RL also unlocked the ability to productively use much longer chains of thought (~30x longer in this example). And these longer chains of thought contributed a much larger boost.

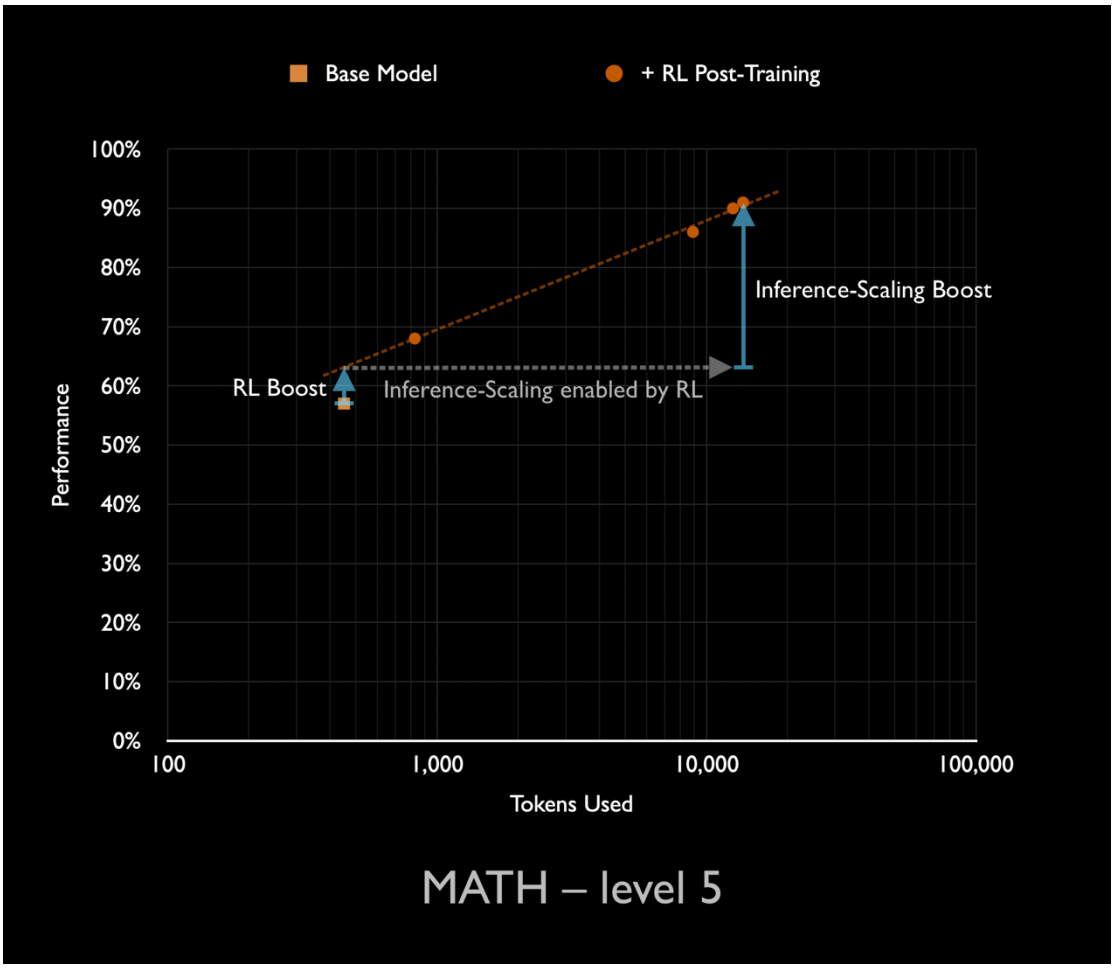


*Figure 2.* A graph showing the performance at different token counts for a base model (Claude Sonnet 3.6, orange square near foot of left-hand arrow) and a reasoning model likely derived from it (Claude Sonnet 3.7, red circles). The reasoning model's performance obeys a clear logarithmic trend (shown on this log-scale graph as a straight dashed line). The performance of the base model is only slightly below this trend, showing that most of the benefits of RL require the use of many more tokens. Figure drawn by author using data from Epoch AI (Ho & Berg 2025).

The question of where these capability gains come from is important because scaling up the inference compute has very different implications than scaling up the training compute.

In this first round of reasoning models, they were trained with a very small amount of RL compute compared to the compute used in pre-training, meaning that the total

cost of training was something like 1.01x higher than the base-model. But if most of the headline performance results require 30x as much inference compute, then the costs of deploying the those capabilities is 30x higher. Since frontier AI developers are already spending more money deploying their models than they did training them, multiplying those costs by 30x is a big deal. Moreover, these are costs that have to be paid every time you want to use the model at this level of capability, so can't be made up in volume.

But that was just the initial application of RL to LLMs. What happens as companies create more advanced reasoning models, using more RL?

The seeds of the answer can be found all the way back in that original o1 chart. The chart shows steady improvements for both RL-scaling and inference-scaling, but they are not the same. Both graphs have the same y-axis and (despite the numbers being removed from the x-axis) we can see that they are both on a logarithmic x-axis covering almost exactly two orders of magnitude of scaling (100x). Indeed, in one version of the graph (OpenAI 2025) the relative numbers for the compute scale have been included, confirming this:

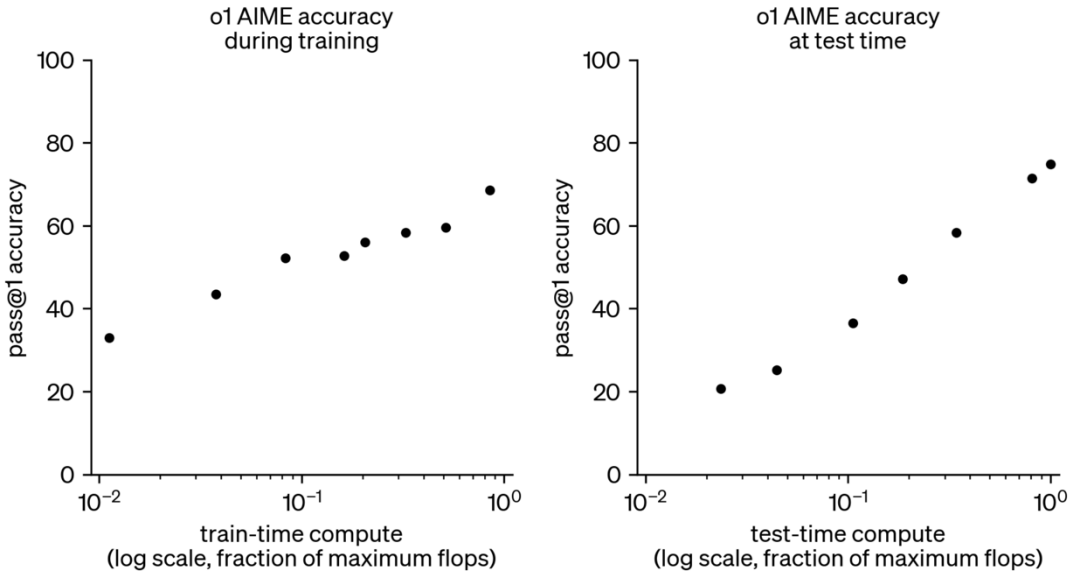


*Figure 3.* A version of figure 1 with labels (and major tick marks) included. Figure reproduced from OpenAI (2025).

In both graphs the datapoints lie on a relatively straight line, which is presumably the central part of a larger S-curve. However, the slope of the RL-scaling graph (on the left) is almost exactly half that of the slope of the inference-scaling graph (on the right). When the x-axis is logarithmic, this has dramatic consequences.

The graph on the right shows that scaling inference-compute by 100x is enough to drive performance from roughly 20% to 80% on the AIME benchmark. This is pretty typical for inference scaling, where quite a variety of different models and

benchmarks see performance improve from 20% to 80% when inference is scaled by 100x.

For instance, this is what was found with Anthropic's first reasoning model (Sonnet 3.7) on another AIME benchmark, with almost exactly the same scaling behaviour:

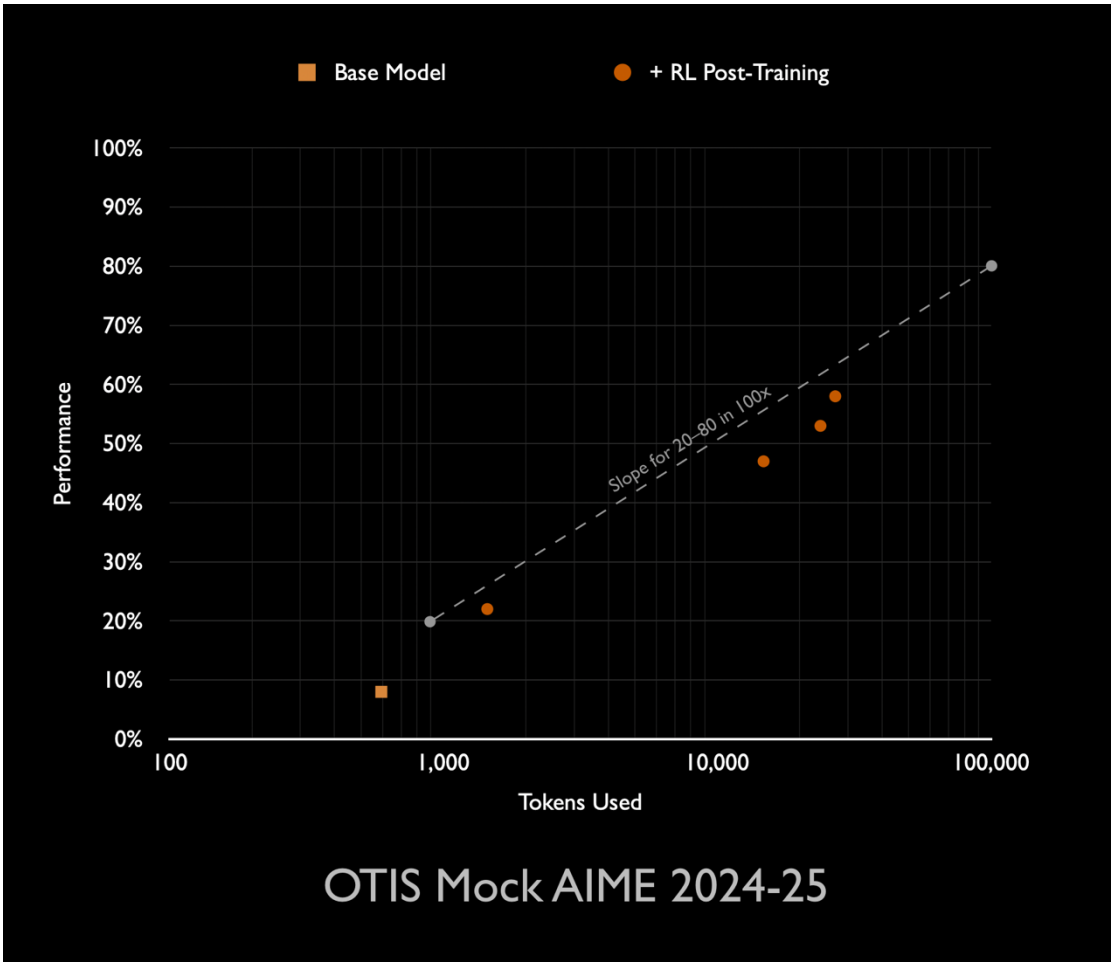


*Figure 4.* A graph showing the performance at different token counts for a base model (Claude Sonnet 3.6, orange square near foot of left-hand arrow) and a reasoning model likely derived from it (Claude Sonnet 3.7, red circles). We can see that the trend for the reasoning model has the slope corresponding to going from 20% to 80% on the benchmark as inference compute is scaled by 100x. Figure drawn by author using data from Epoch AI (Ho & Berg 2025).

And ability on the ARC-AGI 1 benchmark also scales in a similar way for many of OpenAI's different reasoning models:

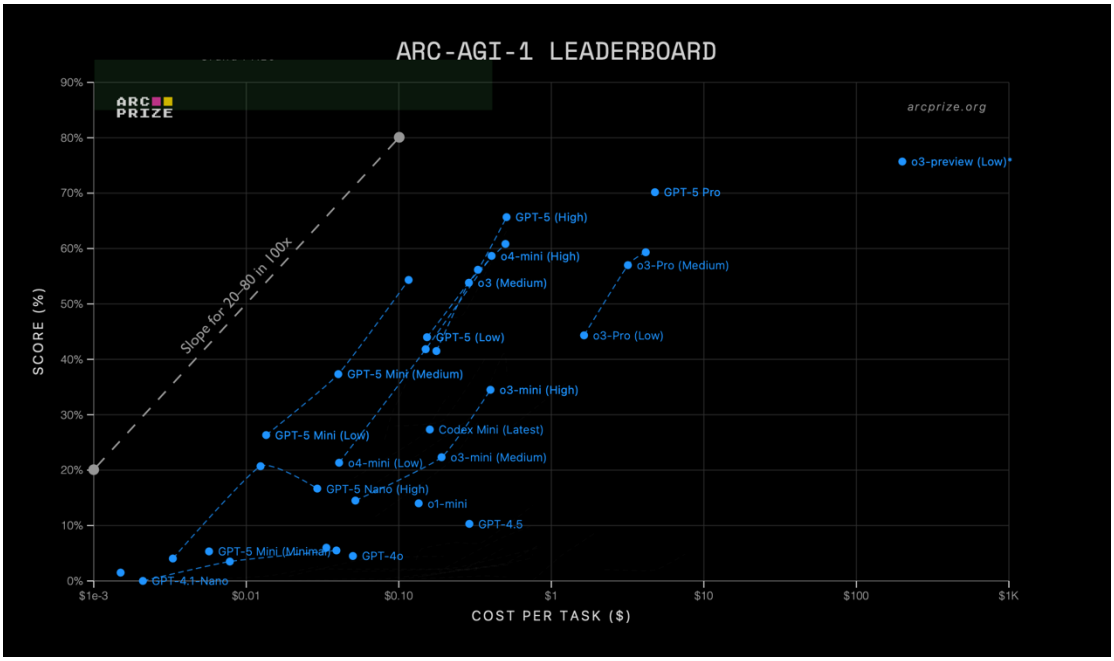


*Figure 5.* A graph showing the performance of many different OpenAI models on the ARC-AGI-1 benchmark (Chollet 2019). Families of reasoning models that use different numbers of tokens are shown connected by blue dashed lines. We can see that these are roughly at the slope corresponding to going from 20% to 80% on this benchmark as inference compute is scaled by 100x. Figure reproduced from ARC Prize Inc (2025) with required slope added by the author.

We don't always see this scaling behaviour for inference: some combinations of LLM, inference-scaling technique, and benchmark see the performance plateau below 80% or exhibit a different slope (often worse). But this climb from 20 to 80 with 100x more inference compute is pretty common (especially for reasoning-intensive benchmarks) and is happening on that original o1 graph.

In contrast, the slope of the RL-scaling trend is half as large, which means that it requires *twice as many orders of magnitude* to achieve the exact same improvement in capabilities. Increasing the RL training compute by 100x as shown in the o1 chart only improved performance from about 33% to 66%. At that rate, going from 20 to 80 would require scaling up the RL training compute by 10,000x.

We can confirm this trend — and that it continued beyond o1 — by looking at the following graph from the o3 launch video (OpenAI 2025a), with a line added showing the slope corresponding to going from 20 to 80 in 10,000x:

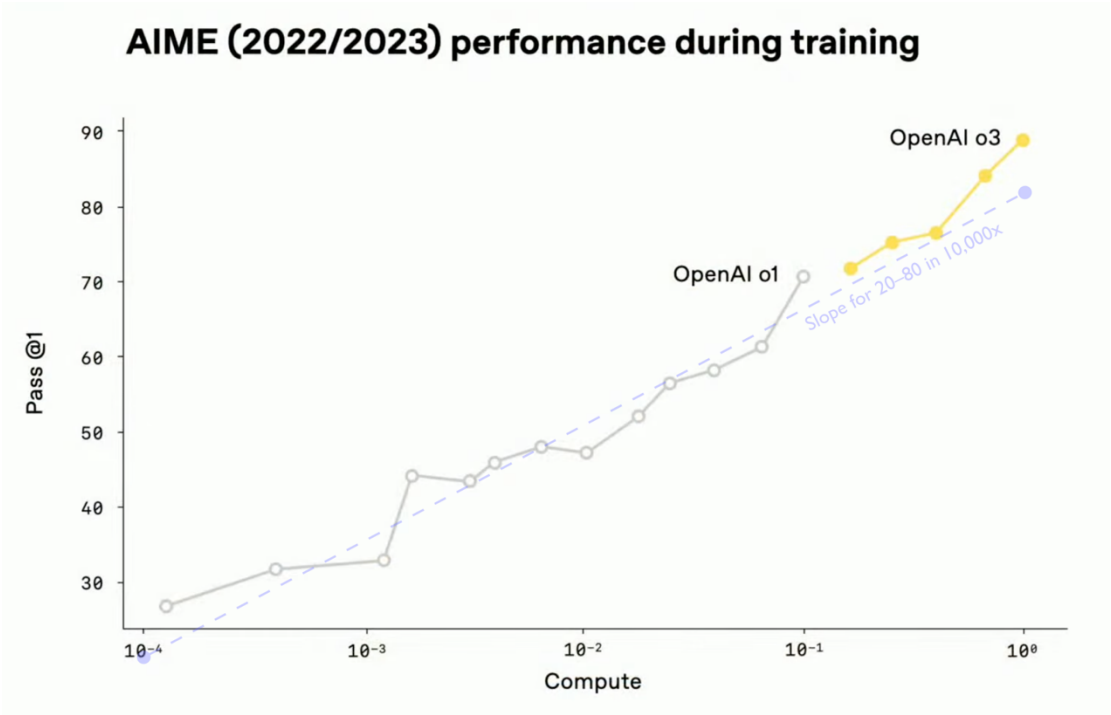


*Figure 6.* A graph showing the performance of OpenAI's o1 and o3 models during RL training. We can see that its ability improved with the slope corresponding to going from 20% to 80% on this benchmark as RL compute is scaled by 10,000x. Figure reproduced from OpenAI (2025a) with required slope added by the author.

Using another version of the AIME benchmark, this shows o1's training progress over 3 orders of magnitude and o3's training over a further order of magnitude. In total, we see that scaling up the RL-training by 4 orders of magnitude takes the model from about 26% to 88%. This provides some confirmation for the rule-of-thumb that a 10,000x scale-up in RL training compute is required to improve this benchmark performance from 20 to 80.

To my knowledge, OpenAI hasn't provided RL-training curves for other benchmarks, but they do have charts comparing o1 with o3 and o3 with GPT-5 at different inference-scaling levels on several benchmarks (OpenAI 2025a, 2025b). Given that o3 used about 10x as much RL training as o1, we'd expect the RL boost going from o1 to o3 to be worth about the same as the inference boost of giving o1 just half an order of magnitude more inference (~3x as many tokens). And this is indeed what one sees on their performance/token graph comparing the two:

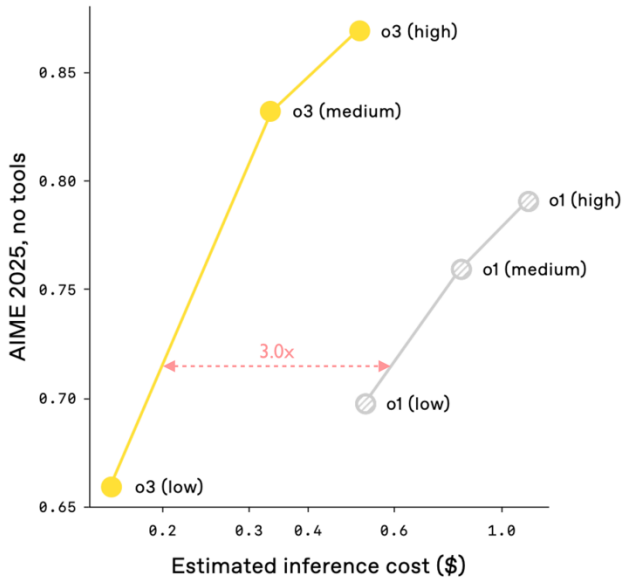


*Figure 7.* A graph showing the performance of OpenAI's o1 and o3 models when using different amounts of tokens at inference time. We can see that the trend for o3 is about 3x cheaper than the equivalent trend for o1, meaning that the extra order of magnitude of RL-scaling was worth about half an order of magnitude of inference-scaling. Figure reproduced from OpenAI (2025b) with the measurement between trends by the author.

Similarly, o3 also requires about 3x as many tokens to match GPT-5 on the SWE-bench and GPQA Diamond benchmarks. This would fit the expected pattern of GPT-5 having been trained with a further 10x as much RL training compute as o3:

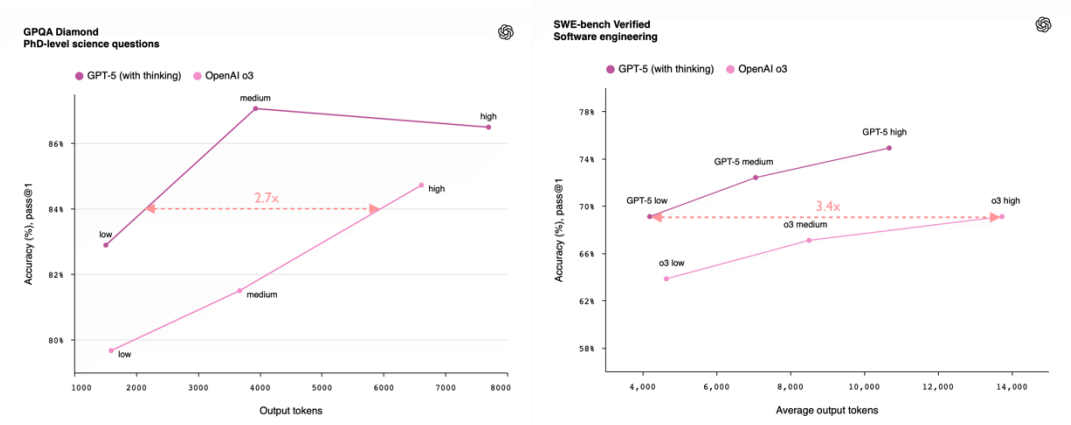


*Figure 8.* Two graphs showing the performance of OpenAI's o3 and GPT-5 models when using different amounts of tokens at inference time. We can see that the trend for GPT-5 is about 3x cheaper than the equivalent trend for o3, meaning that the extra RL-scaling was worth about half an order of magnitude of inference-scaling. Figure reproduced from OpenAI (2025b) with the measurement between trends by the author.

It is hard to verify that this trend holds for models from other leading companies, as this data on training curves for cutting-edge models is often treated as confidential. But the fact that other leading labs' base models and reasoning models are roughly on par with OpenAI's suggests none of them are scaling notably better than this.

So the evidence on RL-scaling and inference-scaling supports a general pattern:

- a 10x scaling of RL is required to get the same performance boost as a 3x scaling of inference
- a 10,000x scaling of RL is required to get the same performance boost as a 100x scaling of inference

In general, to get the same benefit from RL-scaling as from inference-scaling required *twice as many orders of magnitude*. That's not good.

**How Do These Compare to Pre-Training Scaling?**

The jumps from GPT-1 to 2 to 3 to 4 each involved scaling up the pre-training compute by about 100x. How much of the RL-scaling or inference-scaling would be required to give a similar boost? While I can't say for sure, we can put together the clues we have and take an educated guess.

Jones (2021) and Villalobos & Atkinson (2023) both estimate that you need to scale-up inference by roughly 1,000x to reach the same capability you'd get from a 100x scale-up of training. And since the evidence from o1 and o3 suggests we need about twice as many orders of magnitude of RL-scaling compared with inference-scaling, this implies we need something like a 1,000,000x scale-up of total RL compute to give a boost similar to a GPT level.

This is breathtakingly inefficient scaling. But it fits with the extreme information inefficiency of RL training (Ord 2025b), which (compared to next-token-prediction) receives less than a ten-thousandth as much information to learn from per FLOP of training compute.

Yet despite the poor scaling behaviour, RL training has so far been a good deal. This is solely because the scaling of RL compute began from such a small base compared with the massive amount of pre-training compute invested in today's models. While AI labs are reticent to share information about how much compute has actually been spent on RL (witness the removal of all absolute numbers from the twin o1 scaling graphs), it is widely believed that even the 10,000x RL-scaling we saw for o3's training still ended up using much less compute than the ~$10^{25}$ FLOP spent on pre-training (Epoch AI 2024). This means that OpenAI (and their competitors) have effectively got those early gains from RL-training for free.

For example, if the 10x scaling of RL compute from o1 to o3 took them from a total of 1.01x the pre-training compute to 1.1x, then the 10x scale-up came at the price of a 1.1x scale-up in overall training costs. If that gives the same performance boost as using 3x as many reasoning tokens (which would multiply all deployment costs of

reasoning models by 3) then it is a great deal for a company that deploys its model so widely.

But this changes dramatically once RL-training reaches and then exceeds the size of the pre-training compute. In July 2025, xAI's Grok 4 launch video included a chart suggesting that they had reached this level (where pre-training compute is shown in white and RL-training compute in orange):

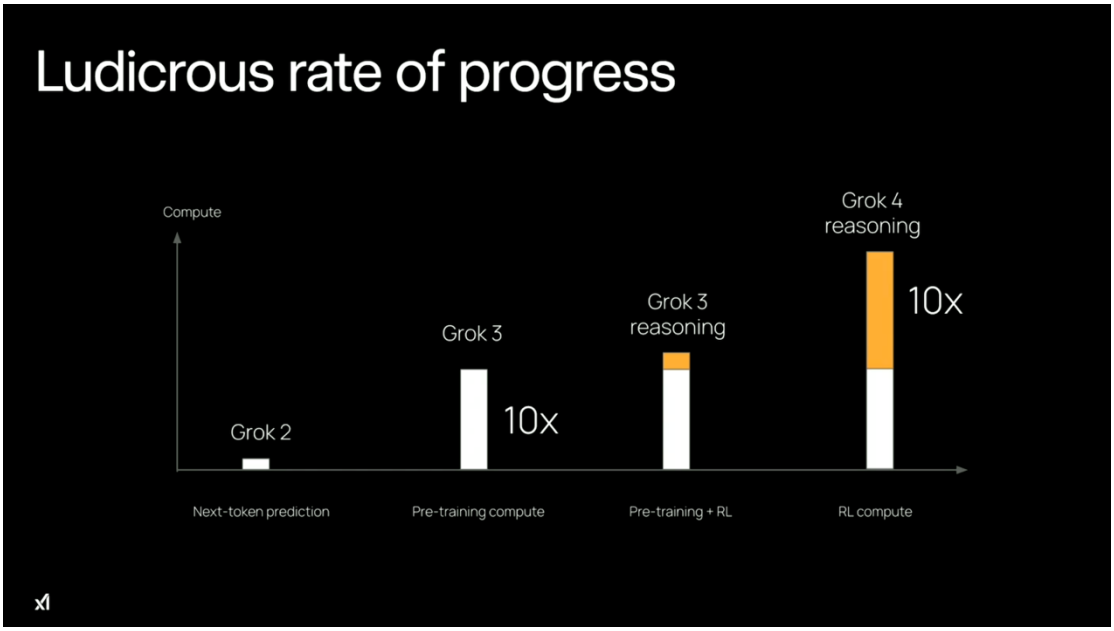


*Figure 9.* A chart showing the scale-up of pre-training compute (white) and RL (orange) in recent generations of xAI's Grok model. Figure reproduced from xAI's launch video (2025).

Scaling RL by another 10x beyond this point increases the total training compute by 5.5x, and beyond that it is basically the full 10x increase to all training costs. So this is the point where the fact that they get much less for a 10x scale-up of RL compute compared with 10x scale-ups in pre-training or inference really bites. I estimate that at the time of writing (Nov 2025), we've already seen something like a 1,000,000x scale-up in RL training and it required ≤2x the total training cost. But the next 1,000,000x scale-up would require 1,000,000x the total training cost, which is not possible in the foreseeable future.

Grok 4 was trained on 200,000 GPUs located in xAI's vast Colossus datacenter (xAI 2025). To achieve the equivalent of a GPT-level jump through RL would (according to the rough scaling relationships above) require 1,000,000x the total training compute. To put that in perspective, it would require replacing every GPU in their datacenter with 5 entirely new datacenters of the same size, then using 5 years worth of the entire world's electricity production to train the model. So it looks infeasible for further scaling of RL-training compute to give even a single GPT-level boost.

I don't think OpenAI, Google, or Anthropic have quite reached the point where RL training compute matches the pre-training compute. But they are probably not far

off. So while we may see another jump or two in reasoning ability beyond GPT-5 by scaling RL training a further 10x or 100x, I think that is the end of the line for cheap RL-scaling.

**Conclusion**

The shift towards RL allowed the scaling era to continue even after pre-training scaling had stalled. It did so via two different mechanisms: scaling up the RL training compute and scaling up the inference compute.

Scaling RL training allowed the model to learn for itself how to achieve better performance. Unlike the imitation learning of next-token-prediction, RL training has a track record of allowing systems to burst through the human level — finding new ways of solving problems that go beyond its training data. But in the context of LLMs, it scales poorly. We've seen impressive gains, but these were only viable when starting from such a low base. We have reached the point where it is too expensive to go much further.

This leaves us with inference-scaling as the remaining form of compute-scaling. RL helped enable inference-scaling via longer chain of thought and, when it comes to LLMs, that may be its most important legacy. But inference-scaling has very different dynamics to scaling up the training compute. For one thing, it scales up the flow of ongoing costs instead of scaling the one-off training cost. This has many consequences for AI deployment, AI risk, and AI governance (Ord 2025a).

But perhaps more importantly, inference-scaling is really a way of improving capabilities by allowing the model more time to solve the problem, rather than by increasing its intelligence. Now that RL-training is nearing its effective limit, we may be losing the ability to effectively turn more compute into more intelligence.

**References**

ARC Prize Inc, 13 Oct 2025. Published online at arcprize.com. Retrieved from: https://arcprize.org/leaderboard

Epoch AI. 2024. 'Key Trends and Figures in Machine Learning'. Published online at epochai.org. Retrieved from: 'https://epochai.org/trends' [online resource]

Pablo Villalobos, David Atkinson, 28 Jul 2023. 'Trading off compute in training and inference'. Published online at epoch.ai. Retrieved from: https://epoch.ai/publications/trading-off-compute-in-training-and-inference

Anson Ho, Arden Berg. 'Quantifying the algorithmic improvement from reasoning models', Published online at epochai.substack.com. Retrieved from: https://epochai.substack.com/p/quantifying-the-algorithmic-improvement

Andy L. Jones, 2021. 'Scaling Scaling Laws with Board Games', arXiv:2104.03113 [cs.LG].

François Chollet, 2019. 'On the Measure of Intelligence', arXiv:1911.01547 [cs.AI].

OpenAI, 12 Sep 2024. 'Learning to Reason with LLMs', Published online at openai.com. Retrieved from: https://openai.com/index/learning-to-reason-with-llms/

OpenAI, 16 Apr 2025a. 'Introducing OpenAI o3 and o4-mini', Published online at openai.com. Retrieved from: https://openai.com/index/introducing-o3-and-o4-mini/

OpenAI, 7 Aug 2025b. 'Introducing GPT-5', Published online at openai.com. Retrieved from: https://openai.com/index/introducing-gpt-5/

Toby Ord, 2025a. 'Inference Scaling Reshapes AI Governance', arXiv:2503.05705 [cs.CY]

Toby Ord, 19 Sep 2025b. 'The Extreme Inefficiency of RL for Frontier Models', Published online at tobyord.com. Retrieved from: https://www.tobyord.com/writing/mostly-inference-scaling

Toby Ord, 3 Oct 2025c. 'Evidence that Recent AI Gains are Mostly from Inference-Scaling', Published online at tobyord.com. Retrieved from: https://www.tobyord.com/writing/mostly-inference-scaling

xAI, 9 Jul 2025. 'Grok 4', Published online at x.ai. Retrieved from: https://x.ai/news/grok-4