

# A new generation of AIs: Claude 3.7 and Grok 3

Yes, AI suddenly got better... again

ETHAN MOLLICK

FEB 24, 2025

---

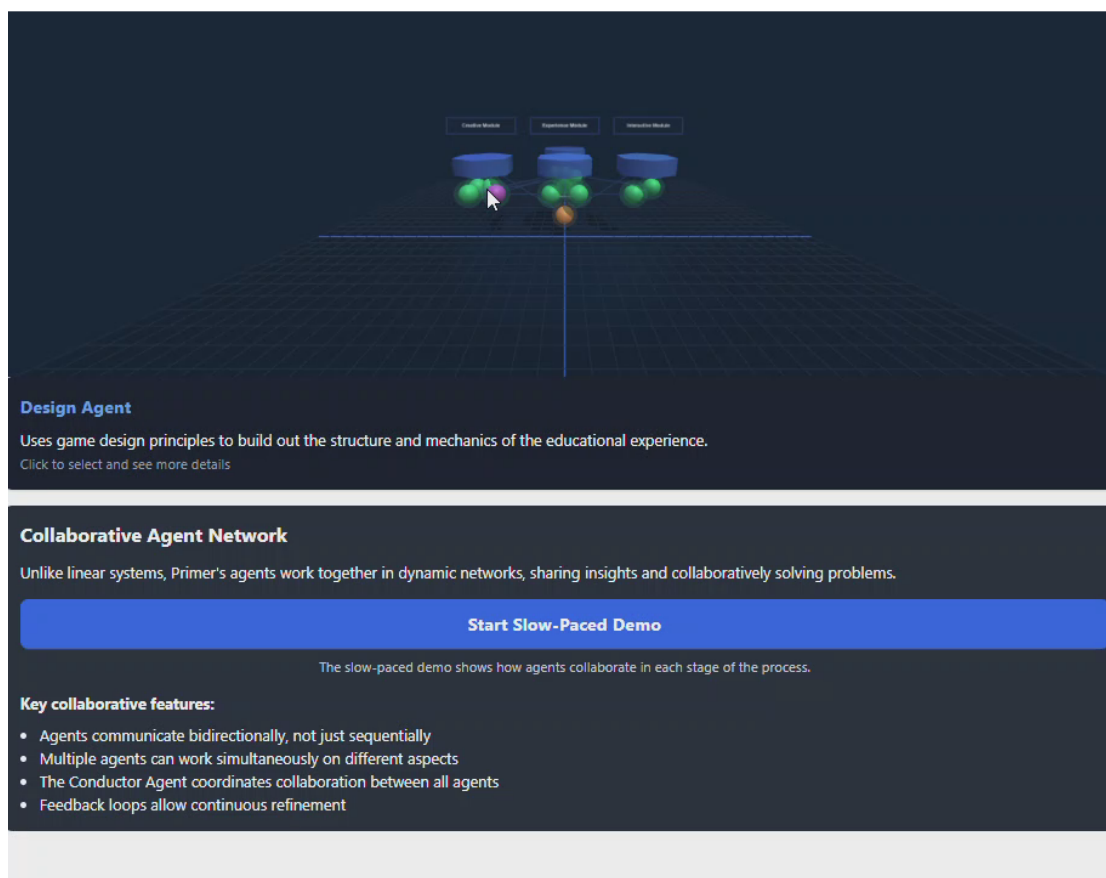
*Note: After publishing this piece, I was contacted by Anthropic who told me that Sonnet 3.7 would not be considered a  $10^{26}$  FLOP model and cost a few tens of millions of dollars to train, though future models will be much bigger. I updated the post with that information. The only significant change is that Claude 3 is now referred to as an advanced model but not a Gen3 model.*

I have been experimenting with the first of a new generation AI models, Claude 3.7 and Grok 3, for the last few days. Grok 3 is the first model that we know trained with an order of magnitude more computing power of GPT-4, and Claude includes new coding and reasoning capabilities, so they are not just interesting in their own right but also tell us something important about where AI is going.

Before we get there, a quick review: this new generation of AIs is smarter and the jump in capabilities is striking, particularly in how these models handle complex tasks like math and code. These models often give me the same feeling I had when using ChatGPT-4 for the first time, where I am equally impressed and a little unnerved by what it can do. Take Claude's native coding ability, I can now get working programs through natural conversation or documents, no programming skill needed.

For example, giving Claude a proposal for a new AI educational tool and engaging in a conversation where it was asked to “display the proposed system architecture in 3D, make it interactive,” resulted in this interactive visualization of the core design in our paper, with no errors. [You can try it yourself here](#), and edit or change it by asking the AI. The 3D graphics, while neat, are not the impressive part. Instead, it was that Claude decided

turn this into a step-by-step demo to explain the concepts, which wasn't something that it was asked to do. This anticipation of needs and consideration of new angles approach is something new in AI.



Or, for a more playful example, I told Claude “*make me an interactive time machine artifact, let me travel back in time and interesting things happen. pick unusual times I can go back to...*” and “*add more graphics.*” What emerged after just those two prompts was a fully functional interactive experience, complete with crude but charming pixel graphics (which are actually surprisingly [impressive](#)- the AI has to 'draw' these using pure code, without being able to see what it's creating, like an artist painting blindfolded but still getting the picture right).

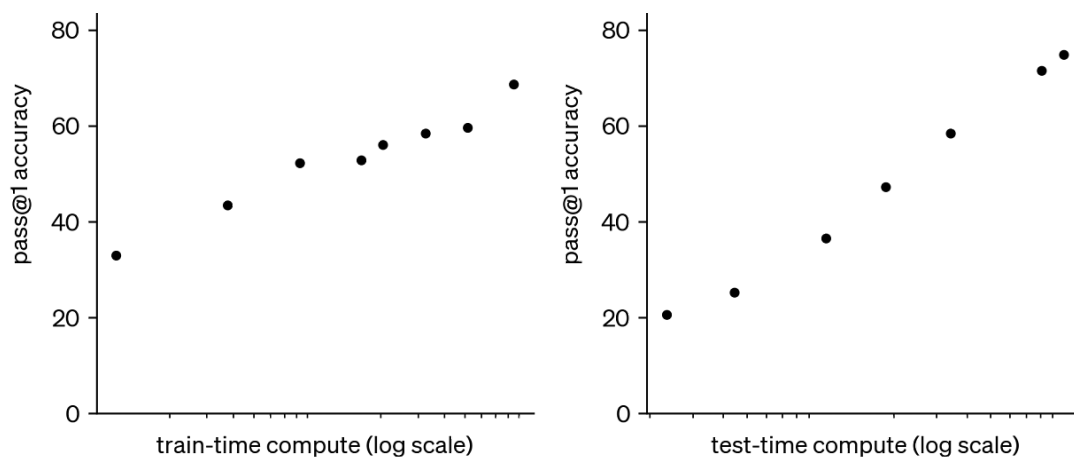


To be clear, these systems are far from perfect and make mistakes, but they are getting much better, and fast. To understand where things are and where they are going,

## The Two Scaling Laws

Though they may not look it, these may be the two most important graphs in AI. Published by OpenAI, they show the two “Scaling Laws,” which tell you how to increase the ability of the AI to answer hard questions, in this case to score more highly on the famously difficult American Invitational Mathematics Examination (AIME).





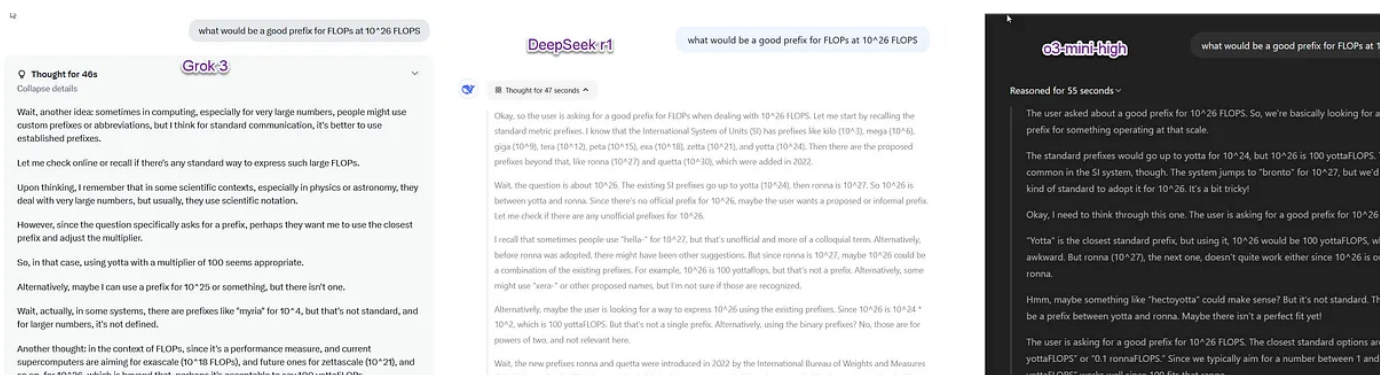
The left-hand graph is the training Scaling Law. It shows that larger models are more capable. Training these larger models requires increasing the amount of computing power, data, and energy used, and you need to do so on a grand scale. Typically, you need a 10x increase in computing power to get a linear increase in performance. Computing power is measured in *FLOPs* (Floating Point Operations) which are the number of basic mathematical operations, like addition or multiplication, that a computer performs, giving us a way to quantify the computational work done during AI training.

We are now seeing the first models of a new generation of AIs, trained with over 10 times the computing power of GPT-4 and its many competitors. These models use over  $10^{26}$  FLOPS of computing power in training. This is a staggering amount of computing power, equivalent to running a modern smartphone for 634,000 years or Apollo Guidance Computer that took humans to the moon for 79 trillion years. Naming  $10^{26}$  is awkward, though - it is one hundred septillion FLOPS, or, taking a little liberty with standard unit names, a HectoyottaFLOP. So, you can see why I just call them Gen3 models, the first set of AIs that were trained with an order of magnitude more computing power than GPT-4 (Gen2).

xAI, Elon Musk's AI company, made the first public move into Gen3 territory with Grok 3, which is unsurprising given their strategy. xAI is betting big on the idea that bigger (way bigger) is better. xAI built the world's largest computer cluster in record

time, and that meant Grok 3 was the first AI model to show us whether the Scaling Law would hold up for a new generation of AI. It seems that it did, as Grok 3 had the highest benchmark scores we've seen from any base model. Today, Claude 3.7 was released, though not yet a Gen3 model, it also shows substantial improvements in performance over previous AIs. While it is similar in benchmarks to Grok 3, I personally find it more clever for my use cases, but you may find otherwise. The still unreleased o3 from OpenAI also seems to be a Gen3 model, with excellent performance. It is likely this is just the beginning - more companies are gearing up to launch their own models at this scale, including Anthropic.

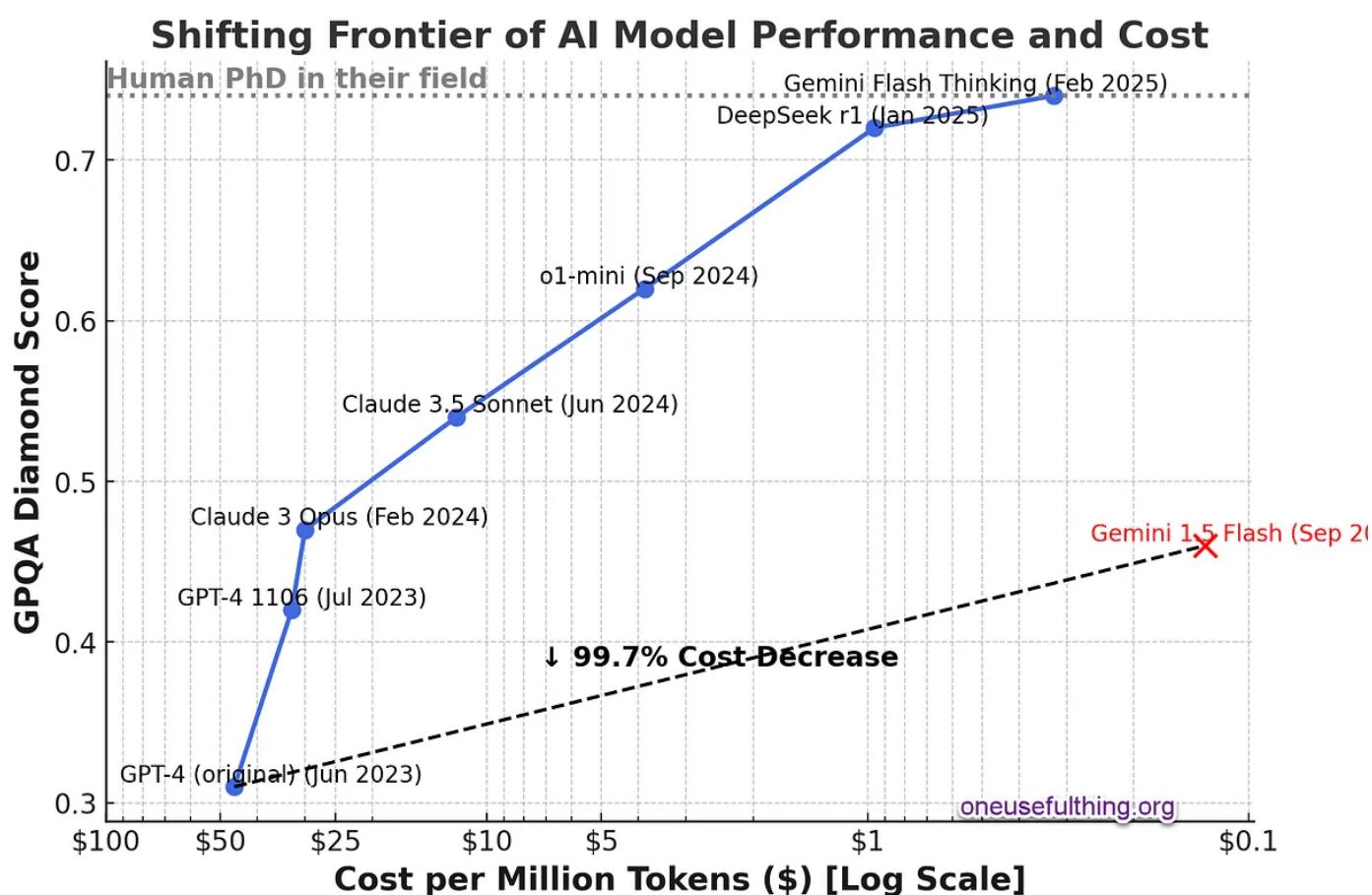
You might have noticed I haven't yet mentioned the second graph, the one on the right. While the first Scaling Law is about throwing massive computing power at training (basically, building a smarter AI from the start), this second one revealed something surprising: you can make AI perform better simply by giving it more time to think. OpenAI discovered that if you let a model spend more computing power working through a problem (what they call test-time or inference-time compute), it gets better results - kind of like giving a smart person a few extra minutes to solve a puzzle. This second Scaling Law led to the creation of Reasoners, [which I wrote about in my last post](#). The new generation of Gen3 models will all operate as Reasoners when needed, so they have two advantages: larger scale in training, and the ability to scale when actually solving a problem.



An example of three different models using reasoning

Together, these two trends are supercharging AI abilities, and also adding others. I

you have a large, smart AI model, that can be used to create smaller, faster, cheaper models that are still quite smart, if not as much as their parent. And if you add Reasoner capabilities to even small models, they get even smarter. What that means is that AI abilities are getting better even as costs are dropping. This graph shows how quickly this trend has advanced, mapping the capability of AI on the y axis and the logarithmically decreasing costs on the x axis. When GPT-4 came out it was around \$50 per million tokens (roughly a word), now it costs around 12 cents per million tokens to use Gemini 1.5 Flash, an even more capable model than the original GPT-



The Graduate-Level Google-Proof Q&A test (GPQA) is a series of very hard multiple-choice problems designed to test advanced knowledge. PhDs with access to the internet get 34% right on this test outside their specialty, 81% inside their specialty. The cost per million tokens is the cost of using the model (Gemini Flash Thinking Costs are estimated). Data based on my research, but [Epoch](#) and [Artificial Analysis](#) were good sources, and [Latent Space](#) offers its own more comprehensive graph of costs across many models.

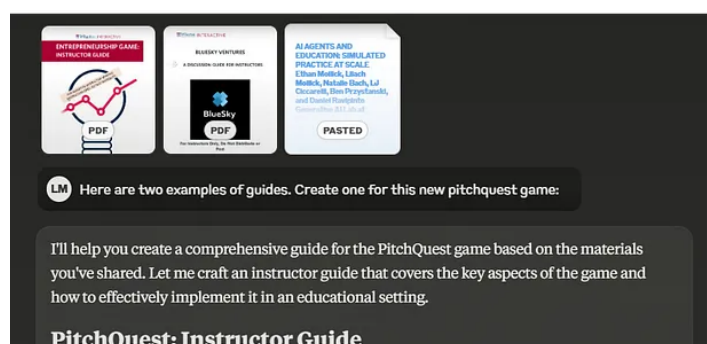


You can see the intelligence of models is increasing, and their cost is decreasing over time. That has some pretty big implications for all of us.

## Taking Scale Seriously

A lot of the focus on AI use, especially in the corporate world, has been stuck in what we call the “automation mindset” - viewing AI primarily as a tool for speeding up existing workflows like email management and meeting transcription. This perspective makes sense for earlier AI models, but it's like evaluating a smartphone solely on its ability to make phone calls. The Gen3 generation gives the opportunity for a fundamental rethinking of what's possible.

As models get better, and as they apply more tricks like reasoning and internet access, they hallucinate less (though they still make mistakes) and they are capable of higher order “thinking.” For example, in this case we gave Claude a [24 page academic paper](#) outlining a new way of creating teaching games with AI, along with some unrelated instruction manuals for other games. We asked the AI to use those examples and write a customer-friendly guide for a game based on our academic paper. The results were extremely high-quality. To do this, the AI needed to both abstract out the ideas in the paper, and the patterns and approaches from other instruction manuals, and build something entirely new. This would have been a week of PhD-level work, done in a few seconds. And, on the right, you can also see an excerpt from another PhD-level task, reading a complex academic paper and checking the math and logic, as well as the implications for practice.



Here are two examples of guides. Create one for this new pitchquest game:

I'll help you create a comprehensive guide for the PitchQuest game based on the materials you've shared. Let me craft an instructor guide that covers the key aspects of the game and how to effectively implement it in an educational setting.

**PitchQuest: Instructor Guide**

I don't see obvious mathematical errors in the tables or the reporting of statistical results. The regression specifications are clearly presented, and the coefficient interpretations align with traditional econometric approaches. The paper's claims about widening performance gaps are supported by the statistical tests in Table 5.

The paper's logic flows well:

1. AI summaries help both sophisticated and unsophisticated investors
2. But the benefits are greater for sophisticated investors
3. Therefore, performance gaps widen rather than narrow
4. This is explained by the theoretical model and demonstrated empirically

This conclusion contradicts some intuitive expectations that AI might democratize

**Introduction**

PitchQuest is an AI-powered educational simulation designed to help students practice and develop pitching skills in a low-stakes environment. Using a multi-agent AI system, the simulation provides personalized learning experiences where students can practice pitching to virtual investors, receive real-time feedback, and improve their presentation skills through guided practice.

This guide will help you implement PitchQuest effectively in your classroom, understand its pedagogical approach, and maximize learning outcomes for your students.

**How PitchQuest Works**

PitchQuest creates an adaptive educational environment using multiple AI agents that

financial markets by helping less sophisticated investors more, but the data and analysis presented in the paper support their conclusion.

One potential limitation (though not a logical error) is that the study only examines one-time use of AI. The long-term impact of AI on investment decision gaps might differ if less sophisticated investors learn to use AI more effectively over time.

Overall, I don't see significant mathematical or logical errors in the paper. The claims are supported by the data presented, and the statistical analyses appear to be conducted appropriately.

Based on my review of this research paper on AI's impact on investment decisions and inequality, the math and logic appear sound and consistent throughout. Here are the key points that demonstrate mathematical and logical integrity:

Managers and leaders will need to update their beliefs for what AI can do, and how well it can do it, given these new AI models. Rather than assuming they can only do low-level work, we will need to consider the ways in which AI can serve as a genuine intellectual partner. These models can now tackle complex analytical tasks, creative work, and even research-level problems with surprising sophistication. The examples I've shared - from creating interactive 3D visualizations of academic concepts to performing PhD-level analysis - demonstrate that we're moving beyond simple automation into the realm of AI-powered knowledge work. These systems are still far from flawless, nor do they beat human experts consistently across a wide range of tasks, but they are very impressive.

This shift has profound implications for how organizations should approach AI integration. First, the focus needs to move from task automation to capability augmentation. Instead of asking "what tasks can we automate?" leaders should ask "what new capabilities can we unlock?" And they will need to build the capacity in their own organizations to help explore, and develop these changes.

Second, the rapid improvement in both capabilities and cost efficiency means that a static strategy for AI implementation will quickly become outdated. Organizations need to develop dynamic approaches that can evolve as these models continue to advance. Going all-in on a particular model today is not a good plan in a world where both Scaling Laws are operating.

Finally, and perhaps most importantly, we need to rethink how we measure and value AI contributions. The traditional metrics of time saved or costs reduced may miss the



more transformative impacts of these systems - their ability to generate novel insights, synthesize complex information, and enable new forms of problem-solving. Moving too quickly to concrete KPIs, and leaving behind exploration, will blind companies to what is possible. Worse, they encourage companies to think of AI as a replacement for human labor, rather than exploring ways in which human work can be boosted by AI.

## Exploring for Yourself

With that serious warning out of the way, I want to leave you with a suggestion. The new models are clever, but they are also friendly and more engaging to use. They are likely to ask you questions or push your thinking in new directions, and tend to be good at two-way conversation. The best way to understand their capabilities, then, is to explore them yourself. [Claude 3.7](#) is available for paying customers and has a new feature where it can run the code it writes for you, as you have seen throughout this post. It [does not train on your uploaded data](#). [Grok 3](#) is free and has a wider range of features, including a good Deep Research option, but is harder for amateurs to use for coding. It is not as good as Claude 3.7 for the tasks I have tried, but the Xai commitment to scaling means it will improve rapidly. You should also note that Grok does train on your data, but that can be turned off for paying customers.

Regardless of what model you pick, you should experiment. Ask the model to code something for you by just asking for it (I asked Claude for a video game with unique mechanics based on the Herman Melville story “[Bartleby the Scrivener](#)” - and it did so based on a single prompt), feed it a document and ask it for an infographic summary, or ask it to comment on an image you upload. If this is too playful, follow the advice in [my book](#) and just use it for work tasks, taking into account the privacy caveat above. Use it to brainstorm new ideas, ask it how a news article or analyst report might affect your business, or ask it to create a financial dashboard for a new product or startup concept. You will likely find cases that amaze you, and others where the new model are not yet good enough to be helpful.



The limitations of these models remain very real, but the fact that Gen3 AIs are better than Gen2, due to both the first and second Scaling Law shows us something essential. These laws aren't fundamental constants of the universe - they're observations about what happens when you throw massive resources at AI development. The computing power keeps growing, the capabilities keep improving, and this cycle accelerates with each generation. As long as they continue to hold, AIs will keep getting better. Now we know that the next generation of AIs will continue to offer rapid improvements, suggesting that there is a good chance that AI capabilities may continue to increase into the future.