

# **SDS PODCAST**

## **EPISODE 916:**

### **THE 5 KEY GPT-5**

### **TAKEAWAYS**



Jon Krohn: 00:00 This is episode number 916 on the five most important GPT five takeaways. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. In today's episode, I'm providing you with the five most important takeaways from the release of Open AI's long anticipated G PT five model. My first big takeaway is that unlike the leap from GPT-3 to G PT four, the transition from G PT four to G PT five may not feel is groundbreaking. And certainly a lot of folks out there have been expressing underwhelm about the model. This underwhelm, however, is misplaced as evaluations by METR, an organization called Model Evaluation and Threat Research as the research clearly illustrates. And if you watch the YouTube version of today's episode, I've actually got a chart showing this progress and I've got a link in the show notes if you want to see the chart, if you're just listening in an audio only format.

01:07 So this chart shows an exponential, exponential rate on the vertical axis, and what it's showing is 50% accuracy on software development tasks. And what it's showing is the time to complete software development tasks at a 50% accuracy rate. So when GPT two was released at 2019, it could only replace a human task that would take a couple of seconds. GPT-3 that came out in 2020 was able to replace humans or do the same kind of task as a human at a 50% success rate on tasks that would take about 10 seconds. GPT-3 0.5, we were at about 30 seconds. That's a big jump, GPT-4, that was a really big jump. It was about five minutes there. So we're going from 30 seconds with GPT-3 0.5 to a five minute software development task being handled at a 50% success rate by GBT four. That was in 2023.

- 02:14 And so that big jump from GBT three in 2020 to GBT four in 2023 from just a handful of seconds to several minutes, that feels like a huge jump. But actually following forward through all the kind of model releases, GPT 4, 0 0 3, ROC four, and now with GPT five we're following this same curve, this 50% success rate on software development tasks, it has a doubling time, an average doubling time of 213 days, and GPT five fits perfectly onto this curve. In fact, it's doing a little bit better than you would anticipate it would do than a cutting edge model in the summer of 2025 would do. And so this is really exciting because it means that we're on this trajectory. It means that in just a couple of years will have in about 400 days from now, so actually that's not much more than a year.
- 03:17 We can anticipate that we'll have models that can handle an eight hour task, a full workday at about a 50% success rate for these kinds of well-defined problems like mathematical problems, software development problems, and so on. So it's not all kinds of problems, it's only a 50% success rate, but following on behind this curve, maybe by 18 months or 24 months is a 90% or 95% success rate. So this means that we're going to be able to have more and more really complex long tasks be handled successfully by LLMs, which means more and more opportunity to be inserting these kinds of models confidently into important processes in enterprises personally in other organizations. It's a really, really big deal. And GBT five fits perfectly on this curve. So underwhelmed maybe, but we are exactly where we should be at this time. Alright, so that's my first big takeaway.
- 04:15 My second big takeaway is that GPT five consolidates several different LLM capabilities into a single model experience. So prior to G PT five's release on August 7th, you might use GPT-4 oh. If you were prioritizing speed,

you might pick GPT-4 0.5 for high quality creative writing and you might pick O three PRO for challenging mathematical or coding tasks. And so it was kind of weird to have to guess what the right model is for a particular kind of task, and it's something that you would just kind of get used to. Now with GPT five, it's just one model experience and so it figures out what the right kind of reasoning approach is, how kind of heavy a lift in terms of reasoning abilities prior to outputting responses required. This is convenient for sure, but this isn't actually, OpenAI isn't the leader in this because Anthropic has had this capability for at least several months.

05:14 I've gotten used to this with Claude Sonnet four and Opus four in the Claude user interface for some time now. Alright, my third big takeaway is that since the advent of LLMs naysayers have been complaining that because of hallucinations, generative and age agent AI applications have limited viability in serious commercial or industrial use cases. I've already found that the hallucination rates since GPT-4, but particularly since age agentic approaches like open AI's deep research were released that hallucination rates are negligible already for most use cases while G PT five continues to make big, big strides here. So again, that you can read about, well, that you can see in the G PT five report that OpenAI released on August 7th, or that I'm showing in video versions of this podcast, it shows that hallucination rates have plummeted relative to opening eyes oh three model. So oh three had about 5% hallucination rates on particular kinds of prompt situations, and this has now dropped to 1% or less with GPT five as long as you have the thinking, as long as thinking gets engaged by GBT five, which you might as well use.

06:39 So that is a big drop from kind of 5% hallucination rates to less than 1%. That is a big deal, definitely going in the right direction there. And meaning that more and more

use cases are now relatively safe to be using with GPT five, with LLMs in general, the cutting edge ones. Speaking of safety, my fourth key takeaway, and as I reported on in detail in episode number 908, LLMs are prone to dangerous deception, especially when they're objectives are threatened like hallucinations. This is another area where GPT five makes huge strides making GPT five much safer to use within Agentic applications than open AI's predecessor models. So again, I've got a pretty dramatic chart in the video version of today's episode that you can check out. But basically the deception rates on different kinds of deception evaluations with open AI three were as high as 50% in one case 90%, but GPT five with thinking brings that down to between 2% and 16% depending on the specific deception evaluation.

07:54 So again, huge, huge step change in reducing deception with GPT five, just like we saw a big, big, big step change reduction in hallucination. Alright, so that's pretty exciting. And my fifth and final takeaway is as the great Dr. Andre Burkoff already recently pointed out in a viral LinkedIn post, which I've linked to in the show notes with GPT five performing only on par with other existing proprietary models such as Claude Opus four on Key Benchmarks LIKEWE Bench, the time to get super excited about what the next cutting edge LLM will be able to do is passed. Now is the time to get super excited about what you can be building and accomplishing with LLMs. The tools available to AI practitioners are extraordinary. What process can you now automate to a high degree of accuracy? What new capability can you improve society with? If you're not sure, as I've said many times on this show, an LLM conversation to ideate on what you could be doing with cutting edge AI tech is merely a browser click away.



09:02      Alright, that's it for today's episode. I'm John Krohn and you've been listening to this SuperDataScience podcast. If you enjoyed today's episode or know someone who might consider sharing this episode with them, leave a review of the show on your favorite podcasting platform, tag me in a LinkedIn post with your thoughts, and if you aren't already, obviously subscribe to the show. The most important thing to me, however, is that you just keep on listening. Until next time, keep on rocking it out there. And I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.