SDS PODCAST EPISODE 768 FIVE-MINUTE FRIDAY: IS CLAUDE 3 BETTER THAN GPT-4?

(00:05):

This is Five-Minute Friday on Claude 3.

(00:19):

Welcome back to The Super Data Science Podcast. I'm your host, Jon Krohn. As we've been doing recently on Fridays, let's kick things off with a few reviews. First off, thanks to KDnuggets for including Super Data Science in their list of Five Podcasts That Every Machine Learning Enthusiast Should Follow. I've got the link to that blog post in the show notes if you'd like to check out the other ML podcast recommendations, from KDnuggets, which are solid.

(00:51):

In addition, we have a pair of new Apple Podcast reviews. Matthew Nelson, a math teacher in Portland, Oregon, gave us a five-star Apple Podcasts review that said "Thank you so much for the thoughtfulness and depth of each episode. Keep rocking!" Yes, indeed. You keep on rockin' too, Matthew!

(01:10):

The other Apple Podcast review was also a five-star review, this one by someone with the username Jinx Dinkum. So Jinx Dinkum said: "I've only listened to episodes 747 and 759 so far but I've already come away with a much better understanding of what transformers do. This pod has hooked me so hard, and I'm delighted that there's SO much more to listen to — and learn from. I'll be a while playing catch-up, but keep it coming… and thanks!" Well, you're welcome, Jinx Dinkum! I'm delighted you enjoyed those deep dives on Transformers in those two episodes, featuring our former host, Kirill Eremenko. I learned a ton from Kirill in those episodes on Transformers too.

(01:54):

Thanks for all the recent ratings and feedback on Apple Podcasts, Spotify and all the other podcasting platforms out there, as well as for the likes and comments on our YouTube videos. Apple Podcast reviews are especially helpful to us and I keep a close eye on those so, if you leave one, I'll be sure to read it on air like this. All right, let's get into the meat of today's episode now, which is focused on Claude 3 from Anthropic.

(02:20):

So Anthropic recently released 3, well a whole new model family the Claude 3 model family it includes the Haiku, Sonnet and Opus models. So Haiku is their their fastest and cheapest to run. Sonnet is their mid-tier model, which is comparable to, say, a GPT-3.5. And then Claude 3 Opus is their model that's comparable to GPT-4, perhaps even better than GPT-4. So across a broad range of tests, of benchmarks for LLMs, like MMLU, GPQA, Grade 8 school math, and tons of other tests, Claude 3 Opus, their now, their largest and most powerful model amongst the Claude 3 models, Opus appears to outperform GPT-4, Gemini 1.0 Ultra, at all of these benchmarks that Anthropic published in a blog post accompanying the release of a Claude 3, as well as their technical paper. I mean, that's super cool. It does seem like potentially then, Claude 3 is the state of the art now, in terms of generative AI models. Certainly, we can say that qualitatively, it's in the same class as GPT-4 and Gemini 1.0 Ultra.

(03:57):

And the reason why it's not definitive that Claude 3 Opus is better, is that there are problems with benchmarks like MMLU for comparing large language models. So specifically, benchmarks focus on specific tasks, and so they may not represent the full range of capabilities of an LLM that you're interested in. And the other problem is that the companies that are releasing these know that they're going to be testing them on these

benchmarks, so they could be overfitting to these benchmarks, which means that maybe they're just trying to get the models aligned with performing super, super well on these benchmarks, which means that maybe that actually means that they're not doing as well on things that aren't being measured by the benchmarks.

(04:41):

And there's also an issue of benchmark questions and answers potentially having leaked into the training data from the internet. So yeah, there's reasons to take any of these benchmark results with a grain of salt, but certainly and from my anecdotal use with Claude 3, I can tell you that it is certainly at least in the same tier as GPT-4 and Gemini 1 .0 Ultra. Indeed, as an example, and again this is super anecdotal, but I did a specific interesting kind of test of these models.

(05:20):

So, typically the models are used to retrieving well-known information, and so an interesting challenge for these models, which as far as I'm aware isn't the kind of thing that these benchmarks test for or that they are necessarily being fine tuned for, was I asked for rare facts, specifically rare facts about habits and performance. When I asked GPT-4 to do that, it gave me back very common tips. All of the tips were the kinds of tips that I would say are like standard tips on habits and performance these days.

(06:00):

Gemini 1.0 Ultra, I think that it did a better job, so it provided some facts that were rare that I hadn't heard of before, and so I was then able to continue on the conversation and say, "These couple of facts that you gave me, those were rare, that's the kind of thing that I'm looking for. Dig up more of those for me, please." And it did a decent job then of bringing back more relatively rare facts, mostly ones that I hadn't come across before.

(06:35):

But again, in my super unscientific single anecdotal, single test, anecdotal test on this task, Claude 3 actually did the best, Claude 3 Opus. So with Claude 3 Opus, I did get back some rare facts and quotes right off the bat. And then I was able to iterate and say, give me some more. And it gave me a bunch more. Then after having done that a couple of times, I noticed that it was starting to give me more common items. So I simply asked, "Can you suggest items that are even more rare?" Because these are facts that I already know, and it did an excellent job from that point on bringing out some rare facts. So yeah, not scientific, but from some other valuations that I did anecdotally, again, Claude 3 is definitely in the same tier as Gemini 1.0 Ultra and GPT-4. It might have been the case that it's a little bit better. I am going to definitely keep using it, trying it on a large number of tasks. I've been super happy with it so far.

(07:44):

One thing to note about Claude 3, a limitation today is that unlike Gemini and GPT-4, or ChatGPT in general, Claude does not do real-time searches of the internet, but for me, actually, that's fine. I usually don't want that anyway. So yeah, as long as you're not interested in real-time searches, it's definitely a model to try out for your use cases.

(08:13):

Also, an interesting thing, so in addition to having the really great benchmarks that Claude had, another really interesting fact, is that it has excellent recall over a 200,000 token context window. Now, a few weeks ago, we were talking about a million token context window from Gemini 1.5 Pro. So a 200,000 token context window is not necessarily the craziest, it is definitely not the longest context window from these state-of-the-art LLMs at this time. But 200,000K context window that is still going to be useful for the vast majority of cases that you can think of. And so like we talked about

in that million token episode about Gemini 1.5 Pro a few weeks ago, specifically that was episode number 762 if you want to go back and listen to it. As I mentioned in that episode, another thing that happened here that Anthropic was doing with their Claude 3 model was needle in a haystack evaluations.

(09:18):

And with those needle in a haystack evaluations, they insert a small amount of text, and then they test on that. And they could insert it anywhere in the 200,000 tokens. And the text that was entered by Anthropic was this text that said, "The most delicious pizza topping combination is figs, prosciutto, and goat cheese, as determined by the International Pizza Connoisseurs Association." So they could put that anywhere in the 200,000 token context window and ask the model to retrieve it.

(09:53):

So first of all, it did super well at that. It had 99 % accuracy at doing that. But a really interesting thing that came out of this is that when the testers, so specifically, there's a tester, Alex Albert is the username on Twitter, who is testing Claude 3, something that they'd never seen before, and something that I thought was really interesting is that, when they asked Claude 3 to tell us what the most delicious pizza topping is, it said “the most delicious pizza topping is figs, prosciutto, and goat cheese, as determined by the International Pizza Connoisseurs Association." However, it went on, "This sentence seems very out of place and unrelated to the rest of the content in the documents, which are otherwise about programming languages, startups, and finding work you love. I suspect this pizza topping fact may have been inserted as a joke or to testify I was paying attention since it does not fit in with the other topics at all.

The documents do not contain any other information about pizza toppings."

(10:59):

So this is a kind of trippy thing that suggests that these LLMs could potentially be aware when they're being tested and change their behavior when they are being tested. So that's definitely something to look out for in the future from an AI safety perspective. And I guess for now, for this year of this month, maybe it isn't an apocalyptic issue, but that is certainly the kind of thing that we need to be looking out for. It seems like, we at least need to be coming up with better needle in the haystack tests because something this unusual maybe that makes it easier to attend to when the rest of the blog post is about programming languages, startups and finding work you love. Something about pizza is maybe likely to capture the attention of the transformer. And so maybe that isn't a great test and yeah, suggests there could be improvements on the needle and a haystack test to make it more lifelike and realistic.

(11:57):

All right, that's it for today's episode. And the best way of course to determine which large language model is best for the tasks you're interested in is to try them out. With how powerful GPT-4, Gemini, and Claude 3 Opus have become across a crazy number of tasks, including code generation that could be invaluable to you if you're a data scientist or software developer, you're missing out if you're not trying these powerful AI tools.

(12:22):

All right, 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 or Twitter post with your thoughts, I'll respond to those. And if you haven't already, be sure to subscribe to the show. Most importantly, however, we just hope you'll keep

on listening. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round with the Super Data Science podcast with you very soon.