SDS PODCAST
EPISODE 762
FIVE-MINUTE FRIDAY:
GEMINI 1.5 PRO, THE MILLION-TOKEN-CONTEXT LLM

(00:05):

This is Five-Minute Friday on Gemini Pro 1.5, the million-token LLM.

(00:19):

Welcome back to The Super Data Science Podcast. I'm your host, Jon Krohn. In our most recent episode, #761, we detailed the public release of Google's Gemini Ultra, the only LLM that is in the same class as OpenAI's GPT-4 in terms of capabilities. Well, hot on the heels of that announcement, is the release of Gemini Pro 1.5. To recap quickly, Google's Gemini family contains models of three sizes: You got Nano, which is the smallest and intended for edge devices, like phones. You've got Pro, that's their mid-size model, which is intended for most use cases, and then the third, the biggest is Ultra which is intended for cutting-edge capabilities.

(01:02):

So, with that in mind, the first crazy thing about the Gemini Pro 1.5 is that this mid-size model has comparable capabilities to the Gemini Ultra, the big model, 1.0. And that Gemini Ultra 1.0 was just released a couple weeks ago, so this is a big release already, it's crazy how quickly is coming out. If it's accurate, and my anecdotal experience with Gemini Pro 1.5 so far suggests it is, this means that the mid-size Gemini Pro 1.5 is close to the capabilities of GPT-4 and, because it's only a mid-size model, this means it's faster and more affordable to use than either Gemini Ultra or GPT-4.

(01:46):

How did Google pull this big feat off? They used the same Mixture-of-Experts (MoE) approach as OpenAI did for GPT-4 but evidently Google figured out how to use it to more dramatic effect. The way these Mixture-of-Experts architectures work is they consist of multiple different Large Language Models — one of these submodels might specialize in, say, math, while another specializes in code and a third specializes in literature. It's

probably not as clear as cut and as simple as that but I'm just saying it like that for the purposes of illustrating how you have these different specialized submodel LLMs that form the Mixture-of-Experts. Depending on the input you provide to the whole architecture, your request will be routed to one of these specialized LLMs, so again using my probably oversimplified example, if you ask a math problem, it will go to the math submodel. If you provide it with a coding problem, it will go to the coding submodel and so on. It's not known how many "expert" LLMs make up the Gemini Pro 1.5 architecture, but for a rough benchmark to aid your understanding of the Mixture-of-Experts concept, GPT-4 is rumored to have eight of these submodels. Again, it's not public how Google is using the MoE approach so effectively, but as the company that originally published on the Mixture-of-Experts approach in 2017, I've got a link to that paper from 2017 in the show notes, but yeah, as the company that originally published on the Mixture-of-Experts approach in 2017, it perhaps isn't surprising that they've managed to overtake OpenAI on implementing it effectively.

(03:27):

Beyond the high level of capabilities, the second crazy thing about Gemini 1.5 is that it has a million-token context window. That's massive. For comparison: The OpenAI model with the largest context window is its GPT-4 Turbo, which has a 128k token context. That's like an eight of the size. The foundation model with the largest context window, period, is Anthropic's Claude 2 model. And its 200k-token context window is just a fifth of Gemini 1.5 Pro's. How big is a million token context?

(04:05):

For reference, a context of a million tokens corresponds to about 700,000 words. Given that novels are seldom longer than 100,000 words, this means you could drop text the length of seven or more typical-length novels into Gemini 1.5 Pro and ask questions about all seven of those novels at once. My friend Allie Miller, super famous AI or business expert who has

1.5 million followers on LinkedIn, Allie Miller demonstrated this incredible capability, this huge context window by dropping eight quarters' worth of Amazon shareholder reports and earnings call transcripts into Gemini 1.5 Pro and the model provided insightful answers to questions like "What was an Amazon focus for 2022 that is weirdly absent from the 2023 shareholder calls and reports?" So she provided with a huge amount of context, was able to ask that question and got great answers, I've got a link to her social media post about this for you to dig into it in more detail if you are interested.

(05:10):

Ok, so that's one example that Allie had. But how can we trust that implementing a huge context window in general is not just for show but it really works? That is, how do you know a model is actually able to reliably attend to the important information from across that huge window? Well, according to Google, they used a so-called "Needle In A Haystack" evaluation, wherein a small piece of text containing a particular fact is placed within a long block of text, and Gemini 1.5 Pro was able to find the embedded text within a million-token input context 99% of the time. That is pretty damn good. Again, it isn't independently verified but quite a stack there from Google. This attention over huge stretches of text allows Gemini 1.5 Pro to learn new skills from a long prompt; for example, you can provide the model with a grammar manual for a language that is outside of the model's training data and the model will be able to translate from English into that new language at a similar level to a human learning the same content. Super cool application there. And of course, an LLM like Gemini can do this learning many orders of magnitude more quickly than a person could.

(06:27):

As if all that I've mentioned so far, the incredible capabilities, the million token contex window, if those two things as if it weren't enough, the third

crazy thing about Gemini 1.5 Pro is that it is multimodal. And we've become accustomed in recent months to the multimodal capabilities of GPT-4V, which accepts both images and text as input, but Gemini 1.5 Pro goes a big step further by also accepting audio or video as inputs too. With its million-token context window, this allows Gemini 1.5 Pro to be fed an hour of video, 11 hours of audio, 30,000 lines of code, or, again, 700,000 words of context as input.

(07:10):

So, to put this to the test, I was fortunate to have access early access to Gemini 1.5 Pro, the million token context model, and so, recently, my talent manager asked me to come up with show reel, so a few minute video with highlights of clips from podcast episodes and other things that I've done. And so I thought well, instead of watching these videos myself, how about I provide them to Gemini 1.5 Pro and have it identify highlights in the videos. So I took one episode, one particular episode, episode #589 with Hilary Mason, I thought this was an interesting one because I did it live on stage with the guest in person. And so I thought this might look cool in a highlight reel and so I provided this 54min long video into the model, so uploaded it into the Google AI Studio which has a click-and-point GUI that you can use, so selected Gemini 1.5 Pro as my model. If you are listening to this on a podcast app and sounds I'm being a little slow and deliberate here, that's because for the YouTube version, I'm actually showing a screenshare of what I'm doing, so hopefully that translates well to the audio format, I think it will.

(08:40):

And so basically, I uploaded video, it identified this 54min video has taken me up 860 000 of the million tokens that I have in that gigantic context window, and then I added a little bit of text. Specifically, I asked the algorithm to use timestamps to identify half of dozen points in this video where the host Jon Krohn showcases personality and intelligence and I'm

saying that because that's specifically what my talent manager asked for. So this is the help, identify candidate clips for highlight reel. So it has this task of identifying these points in the video, these timestamps from the video where I showcase kind of great hosting skills and it pumped out what look kind of at a glance like good results, but in fact, they definitely are not. The timestamps don't make sense. So it provides six timestamps. And remember this is a 54-minute long video. The sixth timestamp was for the five minute and 55 second mark. And it says, "Krohn thanks Mason for her time and insights. This shows his appreciation for others and their contributions." That doesn't make any sense, because the video is 54 minutes long. There's no way five, six minutes in, I'm saying thanks. And kind of wrapping it up the podcast.

(10:06):

So I wrote back to the algorithm. I said those timestamps don't make sense. For example, the 5:55 timestamp is near the end of the hour long video. And the algorithm apologizes. I'm still under development and learning to understand and respond to complex instructions. And then it basically outputs as far as I can tell the exact same output again. So I started to dig into this a little bit more. I was like, okay. Well, maybe the timestamps are wrong, but the content being output could be great. So I checked the transcript to try to identify where these outputs were and it turns out everything that Gemini 1 .5 Pro output was hallucinated, completely made up and done very confidently. And I don't even know what to do with it. I dig into this a bit more and it turns out that this is because, for whatever reason, at this time when you upload a video, it doesn't include processing of the audio of that video. I wrote to the algorithm I said those timestamps don't make sense. For example, the 5:55 timestamp is near the end of the hour-long video. And the algorithm apologizes, it says that it is still under development and learning to understand and respond to complex instructions, but then it basically outputs exactly the same text verbatim.

(11:31):

So I realized that that approach wasn't going to work. The timestamps obviously don't make sense. So I said, okay, instead of timestamps, output a sentence of the dialogue at each of the six points that you identified. And in this case, it still provides me with the timestamps but it does provide quotes so that way I could go to the transcript and try to find these quotes. It turns out the quotes are not in the episode. It is a complete fabrication so the algorithm just confidently made up dialogue between Hilary Mason and me. It all sounds plausible but in fact it is completely made up and it goes into huge detail. The stuff about patients with a risk of developing sepsis, what were the challenges you faced with those data, and it's all made up, which is wild. And I dug into it a bit more, and it turns out that the reason for this is that Gemini 1 .5 Pro, for whatever reason, at this time that I'm using it, it doesn't consider the audio of videos that are uploaded. So it can only do videos.

(12:36):

So then I thought, okay, I'll come up with a test that should work for that. And so I simply said, when is the host smiling? And this it did very, very well indeed. So it provided several dozen timestamps over this 54 minute long video. And I checked all of them and they all are indeed points in the video when I was smiling. And if you're thinking, is Jon just smiling throughout the whole video? Is this a very easy test? No, a lot of the time the camera is on Hilary and a whole bunch of the time when the camera is on me, I am not smiling. So that's pretty cool. It turns out in the end that the algorithm does work very well as long as you're not expecting to get audio results from the video alone. So I'm going to experiment with this a bit more, maybe some blend of approaches here. So trying to identify points in the video where say I'm smiling, but also then I could provide the audio separately and ask for points in the audio where, you know, where I'm putting on a good performance, where I'm doing interesting things and I

could combine those things together, cross -reference those video timestamps with the smiling, the audio where you know there's interesting dialogue and so that might be the approach that I follow next.

(13:58):

Overall, very cool experience, it's surreal to me that we can be doing this kind of thing so quickly. And, so to recap all the crazy features of Gemini 1 .5 Pro that have just been released. First, it approximates GPT-4's capabilities but it is much smaller, faster, and cheaper. Second, it has a million-token context window, which is five times larger than the input accepted by Claude 2, the closest contender on the context window front. And number 3, it's multimodal, it accepts text, code, images, audio and video as input.

(14:31):

Well, that's the end of the crazy-feature exposition for today, but looking ahead a bit, I've got more, which is Google also reports that the company has 10-million-token context-window models under development. That's 10x, the huge million-token context-window that I've been talking about all day today. That corresponds to a model that can be prompted with roughly ten hours of video, 100 hours of audio or 70 novels. It's wild how exponential A.I. developments are these days. I hope you find it exciting and have the cogs turning in your mind on all the ways you could take advantage of these emerging capabilities at work, in products you're developing or in your personal life.

(15:15):

If you can't wait to get started with Gemini 1.5 Pro, you can access it via Google AI Studio while enterprises can access it via GCP's Vertex AI, I've got links to both of those things in the show notes. And as of recording, by default you can only access a 128,000-token version of Gemini 1.5 Pro, but

I've got a link for you in the show notes for joining the waitlist for the million-token version.

(15:42):

Conjecture on my part, but it feels like a safe assumption that in addition to all the stuff I talked about today, that Google is probably working on a Gemini Ultra 1.5, not just the Gemini Pro 1.5. So remember from the beginning of the episode that that Ultra is the largest language model category that Google has been working on recently. If my conjecture is right, then I wonder if that Gemini Ultra 1.5 will be the first model to definitively overtake OpenAI's GPT-4 from a capabilities standpoint. If so, Google — who were the long-time undisputed leaders in the A.I. space — they could potentially be reclaiming their crown.

(16:25):

All right, that's it for today. 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, or if you aren't already, be sure to subscribe to the show. Most importantly, however, I hope you'll just keep on listening. Until next time, keep on rockin' it out there and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.