Welcome back, you're with Connect the World. I'm Becky Anderson. We're at the World Government Summit in Dubai. And one thing I've noticed here is that whenever a discussion about artificial intelligence takes place, the rooms here, the huge halls get packed. That is because some of the leading minds behind the technological revolution have been gathered here at the conference over the past couple of days. And in AI's race to the top, my next guest is sprinting at speeds never seen before. Jonathan Ross is the brain behind GROK, the world's first language processing unit. Now before I lose you in the technological jargon of AI, let me put it this way. What Ross created is a chip that can run programs like Meta's Llama 2 model, for example, faster than anything else in the world. Ten to one hundred times faster, in fact. And he's here with me now to explain how that is possible. Before I ask you that, Grok, why Grok? Thank you, Becky. It's Grok and we spell it with a Q and it's because it comes from a science fiction novel and it means to understand something deeply and with empathy. Of course it does. Tell us about your chip and what makes Grok Chip LPU different from other AI chips and accelerators. I have to tell our viewers that the NVIDIA CEO was here of course this week at the beginning of the week. So we've had all the greatest minds in here. What's your story? well asking me how the chip works before I show you what it does is a bit like asking how a magic trick works before showing you the magic trick you're gonna get bored but I'll give it a shot cool so most chips they they don't have enough memory inside of them sort of like if you were building cars and you use a giant factory and you need about a million square feet of assembly line space Well if you don have a building that large enough to fit that then you need to set up part of the assembly line tear it down over and over again Right And that slow and it takes a lot of time And that what happens with the GPU You have to read from that memory So thousands of times for every word that comes out it as if you setting up the assembly line over and over Right You talk about the wow moment people have when they first experience the speed of your tech. You are breaking performance records almost every week. How fast is your processing today and why is speed so crucial? To coin a phrase, so what why do I care about the speed? The reason you care about the speed is it's about engagement. Imagine if I spoke that slowly you just drift off you'd go away. Most certainly. So the statistic is if you you improve the speed by 100 milliseconds on a

website on desktop, you will get about an 8% increase in user engagement. On mobile, it's 34%. People have no patience on mobile. And we now set a speed record where we can do what's called 500 tokens per second. That would be a novel in about, oh, somewhere around 100 seconds. I want to get to this. I just got to ask you one other question before I actually try this out, because clearly, that's the only way to really kind of reveal what we're doing here. What makes you better than other large language models? So we don't make the large language models. We just make them fast. And so we take open source models. We run them. So it's the same model everyone else has. But you will have a very different experience because of the speed. Right now, AI is unnatural. All right. I asked you that question, but why don't I actually just ask Rock? All right, let's do that. So we're going to turn it on. let do it I know the answer to this Grok but what makes you different well first To another Hey I Grok You can talk to me like I a person All right I will do What makes you different from a large language model that I may have used in the past? Sure, I'm a bit different from those big language models you've used before. I'm a language user interface, or LUI, designed to have more natural... Got it! How similar are you to a human brain? well i'm not a human brain but i'm designed to understand and respond to you like one yeah i can process and generate human language in a natural way got it but tell me something most people don't know um here's something interesting did you know that octopuses have three hearts two pump blood to the gills while the third pumps it to the rest of the body and when an octopus Did you tell it? Are you sure? Are you kidding me? Are you kidding me? Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Ha! Thank you so much. I feel as if I'm talking to a human being. It's worrying me How worried are other chip manufacturers by you? Well, we've been contacted by a few of them The speed is definitely a differentiator and people notice it and should I so thank you very much. I'm sorry Yes, oh we just interrupted thank you very much You're welcome How will this tech be applied in our everyday life I get that and this is quite something because I been sitting using others you know typing away getting quite a lot of good quite a lot of not so good I have to say So this is really interesting What the answer to that question Well, let's talk about the reality for a moment. This technology is getting better and better every single day. Right now, it's at a point

where for most people when they're accessing it, it's unnatural, it's slow. This is going to make it more natural. But the model that you were interacting with, while very good, is not quite as good as OpenAI's model. That natural experience, though, changes it incredibly. What we've done is we've taken a whole bunch of open source models and proprietary models by small companies and we've accelerated them. And that makes that very different experience. So 2024 is the year where AI is going to become real and natural. Who's the customer at the end of the day? So we sell to businesses and those businesses build applications. For example, the application that you just heard is by Vappy.ai. They made all of that work and they're using our chips to do that. PlayHT, DeepGram, Mistral, all of these companies working together to build this. They build the models, and then we make it available to those who want to build applications like that VAPI company. 2024. You're excited. Totally. So we should be too. I guess. That was fascinating. Thank you very much indeed for joining us. you