Daniel Newman: Hey, everyone. We're back here at The Six Five Summit. I'm Daniel Newman. I'm going to be joined here for this next session by Groq's CEO, Jonathan Ross. Jonathan, welcome back to this year's summit. It's great to see you.

Jonathan Ross: Thanks for having me and great to see you again.

Daniel Newman: So this year we came into the year doing a theme like we do every year at The Six Five Summit, and the theme was about navigating rough waters. I think anyone that's been through pretty much November of 2021 through this period of time, whether you're a VC backed company like Groq, or you were traded in the public markets or a small business, you've probably seen a pretty seismic shift. And despite the fact that we've seen a seismic shift from endless growth to what feels like a very fast and steep drop, technology's deflationary value has been notable and companies have been looking to technology and probably nothing more than AI, Jonathan, that has been in focus, especially in the last six months, even since we came up with the concept of this event, the impact of AI and large language models.

You're at the center of this. So I want to start with you and talk about that. There's so much attention around it, and I know you've been spending a lot of your time since before you were at Groq thinking about this problem, give me a little bit about your perspective on LLMs, what you think about them, the experiences that people are having and just what's going on in this particular space.

Jonathan Ross: So LLMs probably not that important. People are going to forget about them soon, not. So I think there are very few technologies that become so obviously important so quickly. I remember when I was a kid, I was one of the few people I knew who was actually using the internet. I remember going on a family trip and hearing about the internet in the airport. Someone said that word. I'm like, "Wow, other people who aren't on the internet know what the internet is." I can't find people who aren't talking about LLMs. It's been the most rapid... It has been such a profound shift that I don't think there's been anything like this in human history.

Daniel Newman: It's directly or indirectly, but whether they talk about it through the lens of, "Oh, I'm using ChatGPT or I use GPT to do X, Y, or Z. Or even my kids are coming home from school, high school age and they're talking literally about large language models. I mean, six ago it was barely on anybody's tongue.

Jonathan Ross: We've been working on this for years, right? So this has factored into the way we designed our chips, our software, everything. And it's funny, it really took kids doing their homework with them for people to start to care. But the fact that these were going to profoundly change things have been known to a lot of technologists for a little while.

Daniel Newman: Yeah, I think we were all aware and we've watched it, whether it's been the conversational modalities of going from single turns to multi-turns. Probably the first time that you maybe entered something into GPT3 or now 4, and you saw it being able to really generate text, I think for people in Copilot, we all know that development was all being done with open source code.

Nobody was coding from scratch. And now you've got Copilot. So now you're taking other people's code and co-piloting to build.

By the way, I mean, I don't know if you have a take on this, but I'd be interested, remember three or four years ago we were telling everyone, "Well, if you learn to code, you're going to be in great shape because they'll al we'll always need coders." Does it feel to you we all have to hit the reverse button and go, "That sounds really silly now."

Jonathan Ross: So I have a take on this, which is we spend a lot of time talking about something called Jevons Paradox which is an economical observation that the cheaper things get, the more people buy, and actually the more efficient things get, the more they buy. So for example, it was based on observing that steam engines as they got more efficient resulted in people buying more coal in total because if it's more efficient, you can do more.

The amount of work that people can get done as coders means that there's going to be more things we do. My EA wrote a program to do matching for speed friendship lunches in Python. Now, she had taken some courses before, but it never really resonated. She didn't really write any programs. She wrote this using an LLM. And what you're going to see, I think is the more that this becomes available to people, and right now it's really limited based on the amount of compute, but as it gets rolled out, as people get more access, you're going to see people doing more of the things that you think that it's replacing.

Daniel Newman: Yeah. Well, it's kind of like every industrial revolution, there's that moment where people go, "Oh no, all the jobs are going to go away. And then 10 years later, there's all these jobs that nobody realized were going to be a job. I mean, you could think about what mobile-

Jonathan Ross: Like the influencer?

Daniel Newman: What mobile did social media do to this? This is going to be the same thing. I mean, we're hearing kind of the dribs and drabs now in the marketplace, companies that are coming out and saying, "We're not going to hire all these certain jobs because we're going to replace them with AI."

Jonathan Ross: What about prompt engineering that's now becoming a thing.

Daniel Newman: Or killswitch engineer? That's kind of a joke. But the point is that, I mean, listen, what we were doing before in terms of learning to use search or to generate SEO, I mean, I don't know about you, but nowadays what I do is I take the text from an article that I write, I'll drop it in Bard or I'll drop it in GPT and I'll say, "Give me the SEO keywords." And in two seconds it does what we used to pay... But like I said, that's going to be an upskilling. We've been talking about upskilling. This is that inflection point of upskilling.

Now, I want to pivot here for a second though, because you talked a little bit about compute and the need for more, and I think you hit something I've been hitting on. I think the market for AI is going to change the hardware market substantially, and the demand is going to skyrocket and the volume is going to go up. We've got challenges of performance per watt. We've got challenges of efficiency, and of course we've got challenge on price.

So as companies like Microsoft rolled out GPT4 or in inside OpenAI, inside Bing, an Edge and inside of productivity, tons of use means tons of cost. And they're not really charging the market yet for this. They have to get more efficient. So whether it's Microsoft or the average company, everyone is going to be using more compute. This is kind of the thesis of Groq and I'll put up front, I'm an investor in the company, but this is what I loved when I met you from the first time. Talk a little bit about how this inflection, this moment is the moment where Groq really can help drive this LLM narrative forward?

Jonathan Ross: So with Groq, we made a bet and the bet was that people were going to focus more on inference than training. So we've optimized for that. And training costs you money. Inference is where you make money when you're doing AI and machine learning. Right? And the observation was at Google created the TPU, but we didn't create it for training. We created it for inference because with training, we created models that worked and you can just throw money at it and it works, and you do it once, and then you have a train model.

But when you're trying to do things at scale, it's all about inference. And the compute needs in LLMs are eye-watering for inference. In fact, not only... So take an experiment, take the Declaration of Independence, paste it into your favorite chatbot, your favorite LLM, and ask it to summarize it in a hundred words and notice how long it takes for the first token to come out versus all the ones that come out subsequently.

Reading is super cheap for them, but writing is super expensive. And the reason is GPUs are terrible at writing tokens. They're okay at reading and reading is easy. Anything you could read on a CPU, there's nothing special there. When it comes to writing the tokens. It's almost like a person and how difficult it is for a human being to write versus reading. Eye-wateringly expensive. And it's actually preventing people from doing things with these models that they otherwise could do.

Daniel Newman: I'm just listening to how you're explaining it. I mean, this is what I think a lot of people don't fully understand is right now every one of these queries, there's a cost inbound and there's another cost back outbound. And so it's exponentially more than search. It's exponentially more than what we were dealing with in terms of what we all have become very used to.

Jonathan Ross: It's thousands of times more expensive based on the best analysis that I've seen than search and search is just retrieval. They pre-compute everything and then they do a lookup. There's a little bit of processing, but it's mostly just lookups. This is computing on demand. So here's a question, do you think that Google is going to spend a thousand or 10,000 times as much per query?

Daniel Newman: Well, I can't logically think of a way they would do that and monetize it and get people to pay. I mean, no.

Jonathan Ross: They don't have enough money. They couldn't if they wanted to.

Daniel Newman: Not with the volume. And by the way, I don't know. A few weeks ago, I think there was kind of a leak document though that talked about how open source LLMs is going to create a major issue. I don't know if you read that.

Jonathan Ross: I did.

Daniel Newman: It was a pretty interesting inflection as well.

Jonathan Ross: And this was also part of the thesis. So we assumed that the ML models would become available and would become cheap. And as cheap as it is to read documents and training happens, reading. Training doesn't require writing, it just requires reading. So as cheap as training was, as the lead document you mentioned refers to this thing called Allora, it basically lets you train on 1/10,000th the size of the model. So even though it was already a hundred times cheaper to train per token than to infer and to write data out, this makes it 10,000 times cheaper yet. So training is such a solved problem. You could probably do that on your laptop, but the inference is eye-wateringly expensive.

Daniel Newman: Yeah, that's really, really interesting. All right. So let me ask you a question cause you kind of started here when you said something along the lines of, "Yeah. People are already getting bored of these." But there does seem to be a contingent, whether it was Elon Musk saying we need a six-month hold, which wasn't so much a doubter, but a sort of positioning as you we're moving too fast and don't know what we're doing. And by the way, can't regulate it.

But there also seems to still be a contingent of people, and I come across them every day that think this is like a fad. We're going to put this genie back in the bottle. I'm going to tell you up front, Jonathan, I don't think this is true. I think it's only going to get faster. But is there any truth in your mind that society will slow this down or that there there's any end of this hype cycle? I can't see it, but I'm curious through your lens.

Jonathan Ross: I would not refer to it as a hype cycle. So the only people that I've ever had refer to it as a hype cycle are people who have not used LLMs to complete a task. The first time that you complete some sort of task, whether it's writing code, whether it's dealing with some sort of writing a paper, whatever you're doing, you realize that the world has changed. It's not a hype cycle, it is reality. And it's a new reality and people have to get used to it.

Now, every time someone says, "Hey, is this hype or whatever, I just drag them to a prompt and I have them do something." They go through phases and the phases are, "Oh, that's interesting. That must have been on the web somewhere. Oh wait, how did it make that change to the... Oh, wait a second. Now I'm getting scared." It always ends with them being like, "I'm a little afraid now." And that's probably where this whole like we need to slow things down is coming from and there is real concern and there should be.

Here's the problem though. You cannot stop it. How would you stop it? Even if you banned it? Then there's still access to these models and they're so cheap to train. People are doing these on their home computers now for the training. So how can you possibly stop it?

Daniel Newman: Yeah. I think that's probably right. And I agree with you. I'm a tech thought leader, influencer, whatever you want to call it, analyst futurist, Jonathan. And we've talked for a few years now. This was the most stunning pivot I've ever seen. We've been through all these pivots. Went through the mobile era, the social era, the, quote, unquote, "digital transformation era".

I've never seen tech proliferate like this and I've never seen proliferate so quickly and be so usable. And that's the other thing about it. There's things that if you were really technical and you could code, you could create. And by the way, automations, smart programmers have been building killer companies by being able to code through workflows.

But this took all that out. If you can speak it, you can do it to some extent. Now again, it'll keep getting better and I tell people that, but probably the biggest thing I get back is the people that will tell you what it doesn't do well. I mentioned once about using it for SEO. And someone is like, "Well, it doesn't give perfect 160 character descriptions, so they're too long." And I'm like, "All right. Well how long do you think it's going to take that to get fixed?"

Meaning I think the bottom line is that we are literally what four or 5 billion daily users on the internet that are daily feeding these things and making them better. The amount of info and data and training and reinforcement that we are doing is unbelievable. But I do got to ask you, because I'm guessing you probably like me, everyone has a few things. One thing I still don't like is they have these table features, for instance, and I like to fill in... I do a lot of earnings analysis on tech companies. It's always wrong.

So one of my pet peeves is I still say the LLM's lie. They lie a lot still. They're close or it'll get three quarters of it is right and then you'll find one data point was just, "Where'd this come from? It's totally made up." Like I said, it's going to get better, it's going to get right, but what's your biggest pet peeves right now in terms of using these and what do you think could drive a better experience?

Jonathan Ross: So first off, new technologies aren't judged based on what they can't do. They're judged based on what they can do. And the fact is they're very good at a large number of things, shocking number of things. In terms of their ability to get better faster. One of the things you mentioned about exact character limits, just as one example between GPT3 and GPT4, it got to that. So GPT4 can usually hit exact character limits, but there's actually something bigger coming and this is going to have more of an impact on LLMs and anything else... This is probably one of the biggest changes, these models are self-reflective and there's papers on this.

What you see is that you can ask a model a question like how could your answer have been better? So recently I was chatting with someone and they were telling me that they wanted to homeschool their daughter. Their daughter is nine years old. And so I asked it to explain something in language that a nine-year-old would understand, and it did. And then we asked it more questions and the answers were more technical.

So then I asked it, what could you have done to better answer these questions? And number two on its list was I could have continued to answer for a nine-year-old like I did with the first question. It is stunning. But to understand why this happens, realize these models, they're operating stream of consciousness right now.

Daniel Newman: Will it open the door in the end?

Jonathan Ross: Sorry?

Daniel Newman: Will it open the door? Keep going.

Jonathan Ross: So there's stream of consciousness. And so they will often write code and people can tell them, "Hey, you made an error there." But you don't even have to do that. You can just simply ask, "Did your code have any errors?" And very often they will notice that they had an error and fix it. So they're self-reflective in that sense. Now, what you can do is you can put them in a loop and have them just keep improving the output. And if you can do that, you can now have a better outcome.

So you can take a model, the best models you have today, and you can have them do what's called reflection, and you can get a one to two generation improvement on the quality. Now, it's eye-wateringly expensive to do this in terms of the inference because inference is already expensive. You also need super low latency, but you can take the existing models and get the equivalent of a one to two generation improvement just by iterating.

Daniel Newman: Yeah. It's pretty incredible. And by the way, I simply changed the prompt on the SEO to say, "Give me a meta description in 160 characters or less." And it instantly did it right. So my point was is it's not hard to just by prompting it correctly, to often get it to do things that people are suggesting are not being done correctly.

Jonathan Ross: You can even ask it for suggested prompts to help you achieve your goals and it will sometimes come up with better prompts.

Daniel Newman: And eventually the prompt engineer will no longer be needed.

Jonathan Ross: Exactly.

Daniel Newman: But all joking aside, so let's wrap up and talk about the investment in this area. I know markets have been toughed, valuations have definitely gone off a cliff, but this area is red hot. Right now with liquidity that's in available in the market with venture and investors, definitely starting to think about the next growth wave. This has to be the most interesting and immediately opportunistic place for dollars to flow.

With all the interest and all the business and all the investment coming into this space, what do you see happening next? How does this grow? What's going to discern the winners and the losers, Jonathan? And this is a moment for you, what's going to make Groqland centrally as part of this exciting LLM story?

Jonathan Ross: So I think if you have a technology that's 10X cheaper, 10X faster, and 10X lower power than a GPU, you're probably going to do pretty well in this market. Otherwise you probably won't. I think we're going to do very well in this market. Stay tuned, but we'll have some announcements soon. We have some demos that we're now showing some very selective customers on something.

I think one of the things that's frustrated me the most about these large language models is the psychological tricks that they do to try and make it seem like they're faster. Some of the stuff that they have to do for Copilot is crazy to make it look like it's usable. But when you're using one of these LLMs and you see it's almost like dial up. You see token, token, token, token, token, token, token. It's so painfully slow. It's usable, but it's not where it could be.

Now, imagine if you typed a prompt and you immediately got an answer, but not only did you get an answer, you got an answer that iterated on three times in order to fix some of its own errors.

Daniel Newman: So, Jonathan, I think we both agree this isn't all hype, but in your mind, how much bigger does it get?

Jonathan Ross: So there's a lot of obvious use cases, programming, customer support and so on. But what a lot of people forget is how much other stuff out there is language. So for example, we're seeing people using this to predict DNA sequences, right? We're seeing people use this to predict molecular compounds. There's a lot of things out there that are language that we don't even realize are language. As powerful as this is on English, it's just as powerful on all of the... Just imagine anything that you can express and how that can be a language, these models are going to affect all of that.

Daniel Newman: Jonathan, what about specialized models, whether they're vertical or whether they're trained to do one specific thing? Are we going to see models that are just going to be trained on just say proprietary data sets from an insurance company or from a healthcare group?

Jonathan Ross: So the answer is both yes and no. So you will not see specialized models that only do one thing. And there's a reason why which I'll explain, but you will see models that are tuned for more of just one thing. Every model has to be general. And here's the reason why. We had a customer recently say that their legal team told them they could not use general models. They had to specialize for a particular task like writing C++. But here's a problem: If I ask a C++ program say to write a currency exchange algorithm, well, it has to understand currency exchange.

So it has to be general enough to take any question that you're going to give it and understand it. However, you will see something called RLHF, reinforcement learning with human feedback, tuning the model to be better at a particular task such as code completion, question answering and these sorts of things. But the models that are useful will not... They must be general. You can't train them on a small data set only.

Daniel Newman: I do want to sort of wrap up here and I just do want to ask you your take, but I've come out in the market and I've said there seems... Obviously we know there are a small number of vendors. Really there's only one at this point that has a huge share of market and it's largely GPU based. Everything I'm hearing indicates that things like application specific ASICs and more efficient would make a ton of sense.

I mean, is the prediction that everything gets bigger, so demand for every kind of compute, or do you eventually think the ASIC will beat out the GPU in terms of its utilization across the AI landscape?

Jonathan Ross: I mean, it's not my opinion, it's the opinion of the godfathers of AI. So you have Jeffrey Hinton who was on a newscast recently where he said algorithms have been around since the 70s and 80s. And what really enabled all of this AI success has been the compute. He was very clear. He thinks that customized silicone is going to be required for this to work.

There's a ton of analogies. Do you use a pickup truck to drive your kids to school? Right? Pickup truck is probably not the right vehicle to use for that, and GPUs are like a pickup truck. They're sort of general. You can use them for a bunch of things, but they're not the right thing. And there's a very simple reason why GPUs are going to fall out of favor. They take way too much power.

You could be okay losing 10 to $20 on every dollar of revenue that you make hosting LLMs. Someone may give you that money. They may give you that shot, but you have a finite amount of power in the grid. And right now we're seeing hyperscalers buying data centers from other companies just to get access to the power. These are terrible data. They're old, whatever. They just need the power. There isn't enough power in the grid.

So if you have something that is very inefficient on power, doesn't matter what someone is willing to pay, there's just a finite limit to how much compute that you can fit in that.

Daniel Newman: Absolutely. I guess I'm just saying, I'm looking at the math, I'm looking at the economics, I'm looking at the shift to open source, staying away from any closed architectures, flexibility, ecosystems, and just saying, to me it makes sense that as this thing unleashes its scale, those are all going to be major factors. And on top of that, the carbon and ESG part of the story is going to be super important.

So if you can come in at lower cost, lower power, you're definitely going to be positioned well. I know that's a big part of your story, Jonathan. I know that's where Groq is heading. It's been a great conversation with you. Love talking about this. I think we will be still talking about it in a month. I don't think it's going away. And I look forward to having you back on The Six Five and having more conversations with you about this soon, Jonathan.

Jonathan Ross: Thanks for having me.

Daniel Newman: All right. Thanks everybody for tuning in here at The Six Five. Stay with us, more sessions to come.