Groq Founder, Jonathan Ross: OpenAI & Anthropic Will Build Their Own Chips & Will NVIDIA Hit $10TRN

91 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=VFIK5LFGNLK&T=5129S&PP=YGUSAM9UYXROYW4GCM9ZCYBNCM9X
https://www.youtube.com/watch?v=VfIK5LFGnlk&t=5129s&pp=ygUSam9uYXRoYW4gcm9zcyBncm9x

SUMMARY

The discussion centers on the critical relationship between compute, energy, and the future of AI, emphasizing that countries and companies controlling compute will dominate the AI landscape. Jonathan Ross, CEO of Grock, predicts significant growth in AI capabilities and economic impacts, suggesting that the demand for compute is insatiable and will lead to massive labor shifts and deflationary pressures in the economy.

- *Countries and companies that control compute will dominate AI.*
- *The demand for compute is insatiable; doubling compute could significantly increase revenue for AI companies.*
- *Current market dynamics resemble early oil drilling, with a few companies generating most revenue.*
- *AI will lead to labor shortages as it creates new jobs and reduces costs across industries.*
- *The future of AI is tied to energy availability; nuclear and renewable energy are key to meeting compute demands.*
- *Companies like Nvidia will continue to dominate but may face competition as other players develop their own chips.*
- *The transition to vibe coding will democratize programming, making it a necessary skill across various jobs.*
- *The AI market is expected to grow rapidly, with significant investments needed to keep pace with demand.*

The countries that control compute will control AI. And you cannot have compute without energy. >> So, I'm thrilled to welcome Jonathan Ross, founder and CEO Grock, back to the hot seat. And now we're going to be able to add more labor to the economy by producing more compute and better AI. That has never happened in the history of the economy before. What is that going to do? I personally would be surprised if in 5 years Nvidia wasn't worth 10 trillion. But I can't predict the outcome. The demand for compute is insatiable. If OpenAI were given twice the inference compute that they have today, if Anthropic was given twice the inference compute that they have today, within 1 month from now, their revenue would almost double. I'm sorry. Can you unpack that for me?

>> Ready to go. [Music] Jonathan, you've just been told by our team that our last show was the most successful of uh the year when it came out. So, there's no pressure at all that this is going to be the most successful of this year. But, welcome to the studio, man. Thank you. It's great to have you here, dude. Now, I I wanted to start with a understanding of where we are. It seems the world moves faster than ever before, and honestly, I think a lot of us are trying to understand where everyone lies in a new market. If we'd look at the current state of the market today, how do you analyze it?

>> Are you asking is there a bubble? >> Relatively. >> Okay. So, um, in terms of whether or not there's a bubble, uh, my answer is if you ask a question, you keep not getting an answer, maybe you should ask a different question. And so, instead of asking is there a bubble, you should ask what is the smart money doing? So, what is Google doing? What is Microsoft doing? Amazon, what are some nations doing? And they're all doubling down on AI. They're spending more. Um, like every time they make an announcement on how much they're spending, it goes up the next time. And one of the best examples of the value that's coming from this spend, Microsoft in one quarter deployed a bunch of GPUs and then announced that they weren't going to make them available in Azure because they made more money using them themselves than renting them out. So, there's real money in the market. And the best way that I I think to explain this market is like the early days of oil drilling, a lot of dry holes and a couple of gushers. I think the stat that I heard was um 35 uh companies or 36 companies are responsible for 99% of the revenue uh or at least the token spend um in AI right now. Yeah, it's very lumpy. And so >> I'm surprised it's not less when you look at No, but I mean seriously, Nvidia really, you know, having concentration of revenue with two clients so heavily.

>> Yeah. And maybe Nvidia represents 98% of that. But um when it's that lumpy, what that's an indication of is it's like the early days of the oil drilling where people didn't know how to find oil. They were going off of instinct. Uh you know, almost vibe investing. Um and uh people who had a good instinct would make a fortune and everyone else would lose their shirts over time. It becomes a science. It becomes very predictable and there's less uh lumpiness. Um there's more predictability, but um investors make less money at that point. The the good investors make less money. So right now is the best time for investors.

Right now people are making more money than they're spending. It's just very lumpy. >> I'm sorry. They're making more money than they're spending. But as an aggregate, plenty of people are going to lose their shirts, but overall less money is going to go in than is going to come out. >> But when we look at the capex spend today by the big providers, everyone is going, "Okay, okay, okay." Because there's something coming at the end of it.

>> Yeah. >> And the trouble is the capex spend is going up and up and up. >> Okay. So you're thinking of it purely financially and I think that the financial returns will be positive, but that's not why people are motivated. So I was in Abu Dhabi at the inaugural Goldman Sachs Abu Dhabi event and um I went and and um you know as you now know uh we're sponsoring McLaren and so um Zack Brown was talking I was talking and it was a fun event but I was asked uh a similar question like is AI a bubble and um I asked the following question everyone so this is like a bunch of people who manage 10 billion plus in aum right the entire like 50 plus people who manage 10 billion boss. I'm like, who here is 100% convinced that in 10 years AI won't be able to do your job? No hands went up. I'm like, great. That's how the hyperscalers feel. So, of course, they're going to be spending like drunken sailors because the alternative is that they're completely locked out of their business. So, it's not a purely economical framework that they're using. It's a do we get to maintain our leadership? Now when you look at it the next step there are these um you know scale law sort of outcomes you want to remain in the top 10 right we keep talking about the mag seven if you're not a member of the mag 7 you're not going to be able to get anywhere near the valuation and so what do you do to stay there you spend and it's worth it because the stock value stays up because you're in the top seven or 10 >> at some point the returns have to be delivered though the spend has to materialize into actual tangible revenue you back and if it doesn't whether you're in the mag 7 or not doesn't doesn't matter. Correct.

>> That's correct. But um right now AI is returning massive value already. It's very lumpy in the al uh in the applications but it's returning massive amounts of value. Let me talk about an example that actually happened for us. So I've tried a little bit of vibe coding. Um I'm not the best in the world at it. We've got some uh interns who are amazing at it. and we we had this customer visit us or um and I had a meeting with them and so they asked for a feature and I specked it out very high level viby um so I was prompt engineering the engineers and 4 hours later it was in production not a single line of code was written by a human being there was no debugging done by a human being it was all prompting um I think we even have slack integration now where you push in like you commit things through Slack. So all that was done, four hours later, it's in production.

Think about the value there. But now imagine, fast forward 6 months from now when that could happen before the customer meeting's over. It's a qualitative difference. It's not even just a dollar amount difference. Yes. Um, you know, when you're able to do it that fast, you spend less to get the feature into production. That's real uh ROI. However, qualitatively when you can do that before the customer meeting is over, you're going to be able to win deals that your competitors won't.

>> Can I ask you just going back to the Mag 7 to stay in the Mag 7? Do you think everyone realizes that they will need to move into the chip layer and own the full vertical end to end? >> Um, I don't think you're going to see too many successfully moving into the chip layer. So people look at the TPU as a big success and what they don't realize is that there were about three chip efforts at Google at the same time. Uh and only one of them ended up uh outperforming GPUs.

And when you look around the industry, you've got a bunch of people building chips. Some of them are getting cancelled like Dojo recently got cancelled. Building chips is hard. I going off and saying, "I'm going to build my own AI chip to compete with Nvidia." It's a little bit like saying, you know, that Google search is pretty nice. let's go replicate it. It's insane. Like the level of optimization, the level of design and engineering that goes into that, um, you're not going to be able to replicate it with a high probability of success. However, if there's a bunch of players out there trying to do it and you have optionality and one of them succeeds, then you have another chip. We mentioned earlier that you have to spend if you want to stay in Mac 7. Mhm.

>> Nvidia investing $100 billion into OpenAI for OpenAI just to go and buy back Nvidia chips. >> Is this not just an infinite money loop? >> Um that would be the case if they weren't spending it with suppliers to build those chips. It's not roundtpping if actual productive outcomes are occurring. So um think of it this way. How what percentage of the spend is going to building that infrastructure? 40%. So at least 40% of those dollars are actually going out into the ecosystem. So that is not an infinite loop.

>> Okay. So it's a partial loop. 60% 60% is going back to Nvidia. >> Sure. >> And then they get a bump in their stock price of a couple hundred billion dollars. >> Yes. >> How did you analyze that? >> Here's So let's analyze it in a couple of different ways from from an economic point of view. Makes perfect sense. Why not do that all day long? Um the value occurs if there is lock in right that's where like when revenue increases result in stock price increases that are greater than the amount of the revenue it's because you believe that that revenue is going to continue and that's the belief and I would actually say with Nvidia that's probably true however it's not just

because Nvidia is good and Nvidia is very good it's also because there isn't enough compute in the world.

There isn't. It's ins the demand for compute is insatiable. Um I would wager that if OpenAI were given twice the inference compute that they have today. If Anthropic was given twice the inference compute that they have today that within one month from now their revenue would almost double. >> I'm sorry. Can you unpack that for me? They are compute limited and it and it comes How would their revenue double if they had double the compute? >> Right now, one of the biggest complaints of Enthropic is the rate limits.

People can't get enough uh tokens from them. And if they had more compute, they could produce more tokens and they could charge more money. And with OpenAI, it's a chat service. So, how do you regulate your chat service? You run it slower. You get less engagement. How important is speed, do you think? There's a lot of people who think actually it's fine. I'm very happy to have latency and I'm very happy to have a prompt and then I go away do something else and something happens when I'm away.

>> Um those are those are interesting opinions. Um let let's look at uh CPG. So consumer packaging goods. So I want you to rank the CPG goods by margin. At the very top is um tobacco uh smoking tobacco. Okay, below that is chewing tobacco. Below that is soft drinks. Below that, you you keep going down, you get to water and other things like that. What is the number one thing that um a high margin correlates to in CPG? It's the speed at which the ingredient acts on you. So that dopamine cycle, how quickly something occurs determines your brand affinity.

And so um when something uh has a very quick response you associate to that brand and then you acrew brand value. This was the entire basis of um Google focusing on speed. Uh Facebook focusing on speed. Every 100 milliseconds of speed up results in about an 8% conversion rate. So that is wrong in terms of people's assessment of the future where they think, "Oh, it's fine. we'll actually just have lots of prompts going on in the background and we'll be happy to let them run for long periods of time.

>> 100% wrong. In fact, when we first started um working on getting speed on our chips, we knew what speed we could get. We even made a video example of how fast we could be. And people would look at that video example and they would say, "Why does it need to be faster than you can read?" And I would respond to that by saying, "Well, why does a web page need to load faster than you can read?" And there's just this mental disconnect where people couldn't couldn't gro the the uh the sort of visceral importance of speed. People are very bad at determining what's actually going to matter in terms of engagement, in terms of outcome, but we know this from building the early internet companies.

>> Do you think OpenAI will be able to move into the chip layer? At some point, Nvidia must be concerned that the OpenAI will want to verticalize and own the chip layer as well. Do you think they will be able to make that successful transition? Um I think one of the problems in building your own chip is it's really first of all everyone thinks that building the chip is the hard part. Uh and then as you do it you start to realize building the software is the hard part and then as you do it you realize keeping up with where everything is going starts to become the hard part.

I have no doubt that uh OpenAI will be able to build its own chips. I have no doubt that um eventually Enthropic will be building their own chips that every hyperscaler will build their own chip. One of the things that um I had this experience when I was at Google where um I I got a lab tour and this was before AMD was doing a great job, right? AMD was struggling for a little while and now they're doing great. But um they had built 10,000 servers and those 10,000 servers of AMD chips, I was walking through the lab and they were pulling the servers out of the racks, taking the AMD chip, popping it off, and throwing it in a trash can. And the funny thing was it was almost pre-ordained because everyone knew that in that generation Intel was going to win. So why did Google build 10,000 servers?

because they wanted to um get a discount on the Intel chips they bought. And when you're at that scale, the cost to design your own server because they had to design their own motherboard in order to fit the AMD chip uh and and to build that out and test it versus the discount that you get totally worth it. So, you have to think of what all the motivations are when people are building their own chips. It's not just because they're going to deploy that chip in mass production.

The the thing is Nvidia effectively has a monopsiny on HBM. A monopsiny is the opposite of a monopoly. So when you're a single buyer and there's a finite amount of HBM capacity, which is the high bandwidth memory that goes into the GPUs, the the GPU itself is made using the same process that's used to build the chip that's in your mobile phone. If Nvidia wanted to, they could build 50 million of those GPU die per year, but they're going to build about 5.5 million GPUs this year. And the reason is because of that HBM, because of the interposer that it goes on. And uh there's just a finite capacity. So what happens is a hyperscaler comes in and says, I want a million GPUs. And Nvidia is like, sorry, I've got other customers. And the hyperscaler says, no problem. I'm going to build them myself.

And then all of a sudden those GPUs are found by Nvidia to give to the hyperscaler. Um there is just a finite amount of capacity by building your own chip. What you really get isn't your own chip. It's that you get control over your own destiny. That's the unique selling point of building your own chip. And so um if >> what does that mean control over your own destiny? >> Uh Nvidia can't tell you what your GPU allocation is. It may cost you more to deploy your own chip because it's not going to be quite as good as Nvidia's.

Let's think about why Nvidia's GPUs with a slight edge over AMD's GPUs dominate. If your total cost to deploy is a huge multiple of the cost of the chips and the systems, then a small percentage increase in the cost of the chip is is negligible. So, think about it this way. If I'm going to deploy a CPU and that CPU is 20% of the bomb and I get a 20% increase in the speed of the chip, that is a 20% value increase in the entire system versus a 20%, you know, you know, increase in the the chip cost, right? It's negligible. So, you get these huge multiples when you improve the the chip performance. So small diff differences in performance make a huge difference in the value of the product.

So a small edge gives you a um massive edge in selling that product. >> Can I ask you you mentioned the monop monopsiny? >> Yes. >> Yeah. Is it possible for openai anthropic any of the mag 7 any of the other providers to move into the chip layer if there is a monopsiny on HBM market? >> It's very hard. However, there is an

incentive from those building HBM to spread that around because Nvidia gets to negotiate very good rates because they're such a large buyer. However, if you're building an HBM fab um and um packaging house and all of this other, you know, part of the ecosystem, if Nvidia comes in and writes a big check, then you're going to build the fab for them. So, Nvidia is always going to get the amount of supply that they want uh in advance. the the problem is you have to um write that check more than two years in advance. And so the where AI's gone, you know, just absolutely honking hockey sticking um even when you have the cash flow of Nvidia, it's hard to actually write the checks for the amount of demand that there's going to be in advance. So there is going to be a supply constraint and it's not purely based on being a monopsiny. Part of it is based on just the sheer capital costs and the memory um suppliers are very conservative. There's also um this situation where the margin on HBM is so high that no one wants to actually increase the supply because then the margin goes down.

>> I totally understand that. Can I ask you when you look at that and when you look at Open AI, when you look at Anthropic having their own chips, h is that why they're raising the money they are? Sam said they're going to need hundreds of billions of dollars. Is that factoring that in? >> No. Most of the spend so um buying a system is expensive. Buying a data center is more expensive. The reason is you're advertising that data center over a longer period of time. So even if a data center was going to be one-third of your cost per year, if you're advertising that data center over 10 years and the chips over 3 to 5 years, the data center is going to end up costing you more per year. So when you hear the hyperscalers talking about that, you know, 75 billion to hundred billion dollar a year investment because they're building out the capacity for data centers, they're putting a lot of money up uh for returns that they're expecting over the next 10 plus years. So it's actually not that much money when you think about it.

>> Are we thinking about amortization in the right way in a 3 to 5 year cycle if chip cycles are actually faster than that? Uh I think that the amortization like people are definitely thinking about it over a longer period than I would. We use a more conservative number uh internally. Um I think five to six years >> which would be like three years >> uh a little bit less. >> Yeah. >> We're we're looking at um upgrading chips about once a year.

>> Yeah. Now here's the the way to think about it. There's there's two phases of the the value of a chip. There's the am I willing to buy it and deploy it and there's am I willing to keep it running. They're two very different calculations. And so when you deploy it, you have to be able to cover the capex. When you keep it running, you just have to beat the opex. So if I deploy a chip today, I have to beat the capex. I have to earn all my capex back and make a profit and produce a return once I've deployed it. as long as I'm beating my operational costs, I'm going to keep that thing in production. So, you're okay with the the price the the value of that chip going down over time.

Now, the bet that everyone is making is that those new chips that come out aren't going to reduce the value of the old chips below the opex. >> That's it. >> That's right. And in our case, we actually don't think that 5 years makes any sense >> because they will be so much less performant that actually the value will be lower than the operating cost >> for the electricity and for paying for the data center. >> So what happens then? We just have this excess supply of wasted chips which are going >> because a lot of these people have entered into really long contracts and so they have a third point where they have to consider their calculation which is breaking this

contract is that cheaper than running the chip at a loss.

>> Yeah. >> Can you see this? Um so what happens then? >> Um then uh I can't tell you what happens because we're trying to avoid that situation. So, um, by having a much faster payback period in all of our, um, calculations, uh, I would not want to make a bet that long out. The the shorter the time frame that you're making the bet, the clearer your outcome is. >> So, essentially, you want to minimize payback period as much as possible and then minimize operating cost so that you can shed less performance chips faster.

>> Yes. But also, here's another crazy part, which is when you look at the math this way, you're like, if I'm approaching it as an accountant, I'm going to be like, this is a terrible idea. But if I look at it empirically, people are still renting H100s, how old are those chips? They're they're getting close to 5 years old. Um, and they're still operating well they're still earning more than their operating cost by quite a bit. You would never deploy an H100 today, but they're still profitable to run, right? They're in that second phase. And the reason is people can't get enough compute.

If that wasn't the case, H100s would be renting for a fraction of what they're renting for today. And as long as you can't get enough compute, that's going to be true. The question is, is there an alternative out there that isn't a supply constraint? And so, this is where we're hoping to come in. So let's let's talk about our value proposition. Um so you started off asking me about speed. Do you know how many customers come to us asking for speed?

>> No. >> 100%. Do you know how many customers keep asking about that once they realize um the supply constraint out there? None. So they start with speed that that they know the value of that to their end customer and then they're like, "Oh, wait a second. I can't even get enough compute." The real value prop is can you provide more compute capacity. So two weeks ago we had a customer come to us and ask for 5x our total capacity. They couldn't get that capacity from any hyperscaler. They couldn't get it from anyone else. We couldn't give it to them. No one can.

And so we couldn't get that customer. The hyperscalers couldn't get that customer. There isn't enough compute. So when you're in a market where there isn't an So your your choice is I buy this compute and I get the customer. This is where I was going to you when I said you know if OpenAI or Enthropic were to double their compute they would double their revenue. Right? So if you're someone who can't get enough compute to serve your customer then you're going to be willing to pay whatever it takes to get those customers because you feel that there's lockin value by getting that customer now.

And so the number one value prop that we have is that our supply chain is not like a GPU supply chain. You you have to write a check two years in advance to get GPUs. For us, you write us a check for a million LPUs and the first of those LPUs starts showing up 6 months later. >> Wow. So you got an 18month chasm difference. >> That's right. >> Wow. >> So I had a meeting with the head of infrastructure of one of the hyperscalers and I talked about speed. I talked about cost and all this stuff, but when I talked about the supply chain and and how we could do something in 6 months, he just stopped the conversation for a moment and wanted to dig into that. That was the only thing he cared about. Think about it this way. If you have a >> given the speed of progression

of the landscape of models, does two years make sense?

>> Well, this is um uh so uh do you know Sarah Hooker? >> No. So she she wrote this paper um the hardware lottery um where it the the my TLDDR on that one is people are designing the models for the hardware. So there are architectures that could be better than attention. However, attention works really well on GPUs. So if you are the incumbent, you have an advantage because people are designing their models for your hardware.

It doesn't even matter if there's a better architecture out there. It's not going to run well. So, it's not a better architecture. There's a little bit of a loop there. So, um if you are building two years out and you're the incumbent, that's okay. But if you're trying to enter the market, no one's going to design for for your chips two years out. So, you have to have a faster loop. >> When you see everyone moving into the chip layer, as you said, OpenAI will have their own, anthropic will have their own. What does Nvidia do in that world?

>> Nvidia still keeps selling chips because >> to Hula given the concentration of their buyers [__] >> Um so no one is successfully predicting how fast AI Okay, we started off talking about is AI a bubble? If you look uh for the last 10 years infrastructure for data centers, you're planning that out two three four five years in advance, right? And what happens is everyone everyone's predictions are wrong. They end up building too little. This has just been what what's happened for the last 10 years. So if you don't build enough for 10 years, what do you do? You try and overbuild. You try and build more than your most optimistic projections. And then once again, you haven't built enough. So you increase your projections and you just keep doing this. Um that's what's been happening.

And yet people still aren't building enough compute. And where people's um instincts are off and and this just hasn't I think been recognized yet. AI doesn't work the way SAS does. In SAS you have a bunch of engineers who go out and build a product and the quality of that product is determined based on what those engineers did. That's not the case in AI. In AI, I can improve the quality of my product by running two instances of the prompt and then picking the better answer. I can actually spend more to make my product better on each query.

I can even decide this customer is more valuable and I'm going to give them a better result. That's kind of what OpenAI announced when they said uh we're about and they did this this week where they said we're now going to release some products where we can't really afford the compute so we're going to give it to a limited set of users and we're going to charge more because we want to see what happens when we give more compute to the to the AI. We want to see what that product looks like and how much better it is. And that is going to be our future. Every time you uh give more compute to an application, the quality increases. And this is why it's not coincidental that you see people's token as a service bill almost matching their revenue because they're competing for customers and if they just spend more, their product gets better. Totally understand that. But bluntly the the assumption when you look at GPT5 and the focus on efficiency is that SAM transition from performance to efficiency because compute does not equal a parallel level of performance improvement. Do you think that is fair and true and does that not go against what you just said?

>> No. And you have to think of the different outcomes that they're looking for. So if you are open AI, you have

moved into markets that are incredibly cost-sensitive. Let's talk about India for a second. So if you want to go win India, what's the one thing you need? 99 rupees a month. That's about a$1.13 with current conversion rates. You need to charge your customer a$113 for your product. So they're going after a market whose alternative is I have no AI.

>> You've got open. I mean they can use deepse. >> This is another misconception in the market. Let's just start busting every misconception. >> Sure. Great. So when the Chinese models came out, everyone reacted by saying, "Oh my god, they've trained models that are almost as good as the US models." And we had we had a a podcast on this, right? Uh and even I was uh snookered a little bit at first. Um and oh my gosh, aren't these models so much cheaper to run? Um now that I know more about the foundation models that people are using uh versus the Chinese models, no, they're not cheaper to run. They're about 10x as expensive. Actually, let's just take the GPT OSS model that was released. It's optimized for something different than the Chinese models, but the quality is very high. Uh, and I would argue clearly is a better model for what it focuses on than the Chinese models. Now, the Chinese models focus on different things. However, the cost to run the OSS model is about uh onetenth that of the Chinese models. So why was everyone charging less? Well, when you have a um sort of a a captive market for a model because people say, "I want this model and there's only one provider of it. You can charge 10 times as much." The price was higher and people were confusing the cost with the price. So the Chinese models were optimized to be cheaper to train as opposed to be cheaper to run.

And when you see how much um intelligence has been squeezed into the OSS model versus the equivalent uh Chinese models, it's clear that the US still has a training advantage. And the economics work out such that you have to amvertise that training over um every inference, which means that you want to charge more. And so there's still a balance there. But as you scale out into larger and larger numbers of people, being able to afford to train a model starts to be a payoff. As you deploy more inference capacity, you want to spend a bit more on the training to get your inference cost down. In the US, we have a massive compute advantage. And so people train the models harder, bringing the cost down.

>> Why do we have a comput advantage in the US just in terms of access to chips? >> That's correct. >> Yeah. Um, and so >> will will China not just subsidize the inference and the running though? I understand we >> Yes. So does it matter if their cost of running is higher, but the Chinese the CCP will just subsidize it. >> Does it matter? >> There's a home game and there's an away game. The home game is we want to um build enough uh compute for the United States. The away game is we want to build it for our allies, right? Europe, South Korea, Japan, um India and so on.

And the advantage that the U so China can can win their own home game. They're going to build 150 nuclear reactors. So they're going to have enough energy even though their chips aren't as energy efficient. Uh and they can subsidize as you mentioned. But the away game is different. If a country only has 100 megawatts of power, what are they going to do? Build another nuclear power plant? Like that's just not a realistic thing. You can do that in China. You can't do that elsewhere.

So having a better chip gives you an advantage in the away game. So my expectation is that right now for the

next 2 to 3 years, the United States has a clear advantage in that away game over China. And if we move very quickly, then we're going to be able to bring a bunch of allies into the AI race. >> Do you think we should have open models to allow for China to distill in the effective ways that they have done already?

>> I think the model itself is not a clear advantage. So the the first time you had me on your podcast, I predicted that OpenAI was about to open source their model. >> You remember that? >> Yeah. And my prediction was based on their branding strength. Frankly, OpenAI could probably open like use um Llama 2, the old model from how long ago? Like two years ago. >> Yeah. And people would probably still use it. And so there's a brand advantage there. Now, they do have very good models, but they don't necessarily need it because of that brand advantage.

I think that Enthropic should be open sourcing their previous generation in order to get people using them instead of the Chinese models because if someone is willing to use a Chinese model, then they would at least be using the anthropic model and their prompts would be recyclable. And just like you have software compatibility, you have prompt compatibility. For example, when the um OpenAI OSS model was released, one of the main reasons people started adopting it over the Chinese models was they could reuse their prompts.

>> Now, of course, when someone has a lowcost application um and they can't afford the the premium for, you know, open AI, they want to use one of these open source models. Eventually, they start doing really well. They make more money. They start wanting to get access to the premium model. Their prompts are reusable. So there's a win by open sourcing these models and you're also getting all of these infrastructure providers to um uh drive the cost down on that model as well. There's a lot of innovation that goes into that.

>> Totally get that. Can I ask you there's so many different areas that want to take this but you we said that just build as much comput as possible. The energy requirements are intense. Is the only way to provide the energy required for this compute wave tsunami whatever you want to call it is the only way nuclear. >> No no no no. Um so nuclear is efficient and and cost effective but um uh renewables are efficient and cost effective. I'll give you my my simple hack. Um, so all all the allies of the United States have to do in order to have more energy than China is to be willing to locate their compute where energy is cheap.

So right now, okay, let's compare Europe to the United States. The United States is incredibly risk averse compared to Europe. >> Wow. >> Yeah. >> In energy? >> No. No. In in everything. But you have to ask what kind of risk. There's two kinds of risk. There's um mistakes of commission where you do something that's a mistake. And then there's mistakes of omission where you don't do something and it's a mistake. And the United States is terrified of making mistakes of omission.

When you are in a massive growth economy, missing out is more expensive than fumbling something. And so the Europe is incredibly willing to embrace the risk of omission. So the way that Europe is trying to compete is through legislation by saying things like I want to keep this data in Europe or I want to keep this data in this country.

If Europe wanted to compete in AI, all you'd need to do is say Norway, please deploy an enormous number of wind turbines. Why? Norway has about an 80% uh utilization rate of wind. So like 80% of the time you can be generating energy. Um they have enough hydro that if you deployed an uh 5x the wind power of the hydro, Norway itself could provide as much energy as the United States and could do it consistently.

The entire United States, that's one country in Europe. How much other energy is there out there that could be unlocked that isn't nuclear? And by the way, let's also deploy nuclear. Nuclear is incredibly safe these days. Why do we not then? >> Fear. >> Is that really it? >> Yeah. When you speak to European governments, what do they say to you? >> I don't bring up nuclear because I'm not going to push an energy source that everyone's going to push back on. But um when I was in Japan recently, they were talking about bringing their nuclear reactors back online.

Japan um has a reputation of being uh very slow. There's a lack of subtlety and nuance in that um perception. The reality is Japan is slow to make a decision, but when they decide something, they move really fast. Um let's take an example. Japan decided to build a 2nmter fab. Um when I was there last they were showing off these two nanometer wafers that they had produced. Now the yield's not where it needs to be. This is not production grade but they built a 2nanmter fab and they are producing wafers out of it and they're going to start getting that defect density down.

They're going to move quickly. Uh they've allocated $65 billion for AI and they're going to spend it and they're going to spend it quick. They're going to turn their nuclear reactors back on. when Japan is going to turn their nuclear reactors back on, Europe needs to listen to that and go, gosh, we need to catch up in energy. >> Catch up is exactly what I was thinking because what I'm thinking is the speed it takes to build out. You said about kind of noise like latent capacity of wind and how we could utilize it. Dude, it takes years to build huge huge supply of turbines.

>> Does it? >> Yeah. Why you think you the Norwegian government is going to let you shell out and have 10,000 wind turbines on the ground? >> Why does the Norwegian government need to pay for it? >> Who should how about the hyperscalers? How about um other governments that want to locate there? In Saudi Arabia, there are gigawatts of power and they're building out data centers for that. Why doesn't Europe work with Saudi Arabia to say, you know what? So Saudi Arabia wants to do a program of data embassies where you have sovereign um oversight over your data, but you get to use their energy.

Why not use that? Problem solved. They're going to build out 3 to 4 gawatt in the very near future. So the hyperscalers would pay Norway to use their renewable energy sources and then leverage that. The complaint that the hyperscalers have is all of the the um paperwork and the slowness. I was talking to someone who was on the board of a major energy company that builds nuclear power plants. He said they spend three times as much on the permitting in the United States than on the nuclear power plant. And I don't know about Europe, but typically the United States is better than Europe on this. How much does it cost to build a nuclear power plant in Europe versus like versus the actual cost of the infrastructure versus the permitting? Here, here's what everyone needs to walk away from this with.

The countries that control compute will control AI and you cannot have compute without energy. How far behind is Europe? Is there a way for us to get back? Like, is it too late? I don't want to be negative. I'm not overly pessimistic, but is there a chasm which we can catch up on? >> I don't think there's a problem right now if if Europe acts now. I mean, um, China is ahead in action, but there are 500 million people in Europe. There's over 300 million in the US. And if you start bringing all the allies together, South Korea, who by the way knows how to build nuclear power plants. The power plant in um the UAE was built by South Korea, they could build power plants here.

France knows how to build power plants. How about a little bit of a Manhattan project for building enough energy? Um when I'm walking around in Europe in the summer, it's incredibly hot. And when I'm walking around in the winter, it's incredibly cold. That is not an experience you have anywhere else in the world. Build more energy. >> I'm with you, Jonathan, but I'm also realistic. I know how slow we are as governments, both singular and in collaborating together. It's not going to happen at the speed of which this needs to be done. What happens if that does not happen in the speed with which it needs to be done?

>> Um, then Europe's economy is going to be a tourist economy. People are going to come here to see the quaint old buildings and that's going to be it. you you cannot um you cannot compete in a new economy if you don't have the resources that the new economy is built on. And the new economy is going to be AI and it's going to be built on compute. >> Is model sovereignty enough to win? If you look at a provider >> because if you don't have compute, you can't run the AI. Doesn't matter how good your model is. Um you could have a model that is 10 times smarter than OpenAI's model. And if you have 10 times the compute, OpenAI's model is going to be better.

>> So for a Mistral who say, "Hey, we're going to have sovereignty within Europe and the German healthcare system and the Croatian transport ministry are going to use Mistral because we're a European alternative. That's not a reason to win." >> What's the USP? What's the unique selling point? >> It's a European model and it doesn't have ownership in the US under a Trump administration. What does it have to do with giving you enough compute? What what you're solving for there is removing someone's else someone else's ability to control you.

>> Yeah. >> But what you're not solving for is having enough of it. And by the way, I'm not saying don't use Mistral. We have a partnership with Mstral. We we love Mistral. The the thing I'm saying is build enough compute so that ML can compete. If you listen to this, are you not just like, "Shit, I should just buy the [__] out of Coreweave." Seriously, like when you look at what they provide on demand, like yeah, Core is a great company. Um, but they have a finite allocation of GPUs.

Everyone has a finite allocation. >> When we chatted before, you said to me that GPUs are not the best infrastructure for inference. >> Correct. and that we are moving more and more into a world of inference as we move further along the maturation cycle of training models. >> Yes. >> Does that not mean that Nvidia's power hold weakens further? >> No. Nvidia is going to sell every single GPU that they build. And even if we end up um supplying 10 times as many LPUs as GPUs, all that's going to do is increase the demand for GPUs and allow

them to charge an even higher margin.

>> Why is that? Sorry. >> Because the more inference you have as as mentioned before the the more you need to train the model to optimize for the inference and the more training you have um the more inference you want to deploy to optimize for the cost of that training to amvertise the cost of the training. There's a virtuous cycle between the two. >> Is the inference market playing out as you expected it to in terms of maturation deployment speed?

>> What I never expected was that AI was going to be based on language. And what that's done is it's made it trivial to interact with AI. I thought it was going to be more like Alph Go. I thought it was going to be intelligent in some weird esoteric way. The fact that it's language means anyone can use it. So I expected AI to come sooner and grow slower. It came later and it's growing faster than I ever imagined. It is so easy to interact with AI that anyone can do it. 10% of the world's population is a GBT weekly active user.

>> Isn't that astonishing? >> Yes. But um you know what's holding it back? >> Compute. >> So compute is holding it back for the quality of it. But more people would use it. They just wouldn't get as much out of it. But more people would use it if more languages were supported. Well, >> this is the number one complaint we hear around the world. You know what would solve that? More compute, more data. If you have more data then you can train more but you need more compute and by the way if you have more compute you can generate more synthetic data so you can train more every one of these so you have data you have algorithms and you have compute if you improve any one of them it's not a bottleneck it's not like if the compute doesn't get better I can't use more data or if the data doesn't get better I can't use more compute any one of these that gets better improves AI and that makes it really easy to improve AI because you can improve improve one dimension of it. It just turns out the easiest knob to improve an AI, it's not the algorithms. Algorithms rarely improve. It's not the data because it's really hard to get more data and we haven't fully figured out synthetic data generation. We're we're good at it, but but we're not we're not at the point yet where we can just directly turn compute into more data. We're getting there.

Compute is the easiest knob because it just keeps getting better and better and better every year. And if I write enough u you know if I write a check for enough money and I'm willing to wait a little while I'm gonna get more compute. It's the most predictable part of the pipeline >> given it's the most predictable part of the pipeline >> and yet we still underestimate how much we need. >> Do you think we are dramatically underestimating how much we need today?

>> Yes. Yes. >> By what scale? Going back to what I said about how every time you add more compute a product gets better. Um there is no limit to the amount of compute that we can use. It's different from the industrial revolution. In the industrial revolution um you couldn't use energy unless you had the machinery to use it and you had to build machinery and that took time. If I wanted to um if I wanted to, you know, have more cars on the road, I had to build the cars. It wasn't enough to just pull more oil out of the ground. AI is not like that. Yes, if I make my model better, it I can actually do more with the same amount of compute. But if I double my compute, I double the number of users. I improve the quality of the model. This is different. I can literally just

add more compute to the economy and the economy gets stronger.

We've never had that before where it wasn't a bottleneck. It was more of a rubberneck where you could just force more of one component through and then everything improves. You said the economy gets stronger. When we think about kind of what that's predicated on, that's predicated on the $10 trillion labor spend in GDP uh shifting um to AI and us taking a portion of that. Do you think that we will see significant shifts in the GDP or the spend on labor moving towards AI in the next 5 years? I believe that AI is going to cause massive labor shortages.

Yeah. I I don't think we're going to have enough people to fill all the jobs that are going to be created. There's there's three things that are going to happen because of AI. The first is massive deflationary pressure. Um this cup of coffee is going to cost less. Your housing is going to cost less. Everything is going to cost less, which means people are going to need less money. >> So, how is it going to cost less to have a cup of coffee because of AI? because you're going to have uh robots that are going to be farming the coffee more efficiently. You're going to have better supply chain management. You're going to um it's just going to be across the entire supply chain. Um you're going to be able to genetically engineer the coffee so that you get more of it per um watt of sunlight, right? Just across the entire spectrum. So, you're going to have massive deflationary pressure.

That's number one. And what that means is people will need to work less. And that's going to lead you to number two, which is people are going to opt out of the economy more. They're going to work fewer hours. They're going to work fewer days a week, and they're going to work fewer years. They're going to retire earlier because they're going to be able to support their lifestyle working less. And then number three is we're going to create new jobs and new uh company uh new industries that don't exist today.

Um think about 100 years ago. 98% of the workforce in the United States was in agriculture. 2% did other things. When we were able to reduce that to 2% of the population working in agriculture, we found things for those other 98% of the population to do. the jobs that are going to exist a 100 years from now, we can't even contemplate. 100 years ago, the idea of a software developer made no sense. 100 years from now, it's going to make no sense, but in a different way because everyone's going to be vibe coding, right? Um, and influencers, that wouldn't have made sense 100 years ago.

Uh, but now that's a real job. People make millions of dollars off of it. So, what jobs are going to exist 100 years? So, number one, deflationary pressure. Number two, opting out of the workforce because of that deflationary pressure. And number three, uh, jobs and companies that couldn't exist today that were going to exist and are going to need labor. We're not going to have enough people. >> It's fascinating the counternarrative, isn't it? Everyone being like, "Ah, millions and millions of people will be unemployed." And you're like, "No, we're actually not going to have enough people for the jobs."

>> Well, what was the famous um prog prognostication 100 years ago that there was going to be massive famine because we weren't going to be able to feed ourselves? People always underestimate what's going to change in the economy when you improve technology. >> When you think about the requirements from an energy

perspective and then also what you just said there about kind of labor, do you think Trump and a Trump administration is doing more to help or to hurt the advancement of AI in the US?

>> Um, definitely help. Uh, all of the moves that have been made are things that are going to help with AI. Um for example um you know the permitting issues right um overall it's been a very positive experience on AI >> you mentioned vibe coding I do just have to ask about it do you think this is an enduring and sustainable market when you look at um a lot of the use cases today they're quite transient >> do how do you analyze the future of the vibe coding market having played with it a little bit and having seen also interns as you said who are very good at it internally use it Well, >> um, vibe coding is going to be, so let let's take reading. Reading used to be a reading and writing used to be a career.

If you were a scribe, you were one of the small percentage of people who knew how to read and write and people would hire you just to record things and you you did much better than the average person in the economy because of that because it was a specialized skill. Coding has been the same thing. Very small percentage of the population did. It took you know a couple years to learn how to do it well. Uh some people were really good at it.

Now everyone reads, everyone writes. It's not a special skill. It's expected in every job. And coding is going to become the same thing. For you to be in marketing, you're going to have to be able to code. For you to be um in customer service, you're going to have to be code uh be able to code. Uh, I was having dinner with someone who runs a chain of 25 coffee shops, has never coded in their life, and they vibecoded a a supply chain tool that allowed them to check inventory.

They didn't write a single line of code. They got it to work. And it was funny because they discovered all the problems that we software engineers uh discover over time. They started getting feedback from their employees like this feature doesn't work. it doesn't this thing doesn't work when I do this all the little edge cases and then he just started fixing them and all through vibe coding >> do margins matter in a world of exponential growth when we look at the demand for your products when we look at the demand for a lovable or a replet both bluntly have bad margins does it matter having bad margins when growth demands are so high >> um I would say that margins first of of all, you do have to have profitability in the end or at least break even, right? To be an ongoing concern. At some point, you can't just keep raising money. Even Amazon had to start making some money.

But the real reason why you need higher margins is volatility. Because if you have a razor thin margin and the market moves, you may not be able to raise more money. You may not be able to get a loan. And so what a margin does is it gives you stability and staying power in the market. On the other hand, um what it does is it also gives competition the ability to enter, right? Your margin is my opportunity.

And so what you're trading is stability um for um for a competitive mode. That's the decision that you have to make. >> How do you think about margin internally today? Um, I think you want the ability to have margin and you want to give it to your customers and you want to give them an advantage. And if you have the ability to

take that margin when it's needed, then um you're in a great position. So I remember talking to a so we hired this amazing CFO recently but I remember talking to a previous candidate and when we were talking about um margin they said that we should price uh so that our supply met our our demand.

In other words they wanted to increase the price in order for the demand to come down. >> Makes sense. >> Does it >> economic sense? Yeah. >> Economic sense. >> Logically and rationally. Yes. >> But then logically uh why not um use up your brand equity? Why not use like the trust that your customers have to sell them things that aren't good? Brand value, brand equity has value.

You want to keep your brand equity as high as possible because trust pays interest. And similarly, you want to keep your margins low enough that you're building up this sort of equity value with your customers where they know that you are giving them a good deal. When you charge a high margin, you are at odds with your customer and you want to do everything that you possibly can to align with your customer. I want my margin to be as low as I possibly can make it while keeping my business stable. And I'm going to make my cash flow by increasing the volume. And one of the things that I love about the compute business is that the need for compute is insatiable. It's Jevan's paradox. If we produce 10x the compute, we will have 10x the sales.

That's just the way it works. As long as we keep bringing the cost down, people are going to buy more. And so I want to keep bringing that cost down. I want to keep increasing the volume. And I want to keep selling more for less so that people get more value out of their business and they buy more and that cycle continues. >> How far are we on the journey to bring the cost down? You know, I remember I look back at some of the shows, dude, and I I would cringe at myself because I'm talking about like, oh, Canva implementing AI and it's hurting their margins because they're implementing AI and it's going to cost them more. And it's just such a naive approach to ask that question even because now the cost of implementation has gone down by 98%.

How far are we in terms of that cost reduction cycle? >> Well, let's step back and and use your Canva example. >> Yeah. >> Um successful businesses don't watch the bottom line. They watch their customers. They they solve problems that their customers have. If you are competing, you are doing it wrong. You want to differentiate. You want to solve a problem that your customer has not solved yet and can't solve any other way and then they're happy to pay you money.

And that's how it works. You solve their problem and then your cash flow is solved. So someone's spending on AI, if you just look at the balance sheet, that doesn't make sense. But when the customer is very happy and they're solving a problem that they couldn't solve otherwise, first of all, you're increasing the TAM usually with AI because it makes the product so much easier to use. Did you use Photoshop two years ago? Impossible. Now, if you want to generate an image, you just explain what you want. That increases the TAM. You may be able to charge less per photo, but your total revenue increases. Your total market increases.

Forgive me for this financial question, but we see the S&P about to hit 7,000. We see this ripping of the MAG 7 like we haven't seen a concentration of value in many, many years. And people suddenly start to feel like, wow, it's getting toppy. I listen to you and I hear all of this and I think it's just the start. How should I think about the

duality of those two thoughts? There's two components to the value. Um, one is the weighing machine and one is the popularity contest.

And there are some products that are pure popularity contest like crypto. I have never bought a Bitcoin, you know, I missed out. Why? Because I can't play in the popularity contest. I'm not good at it. I don't know what's going to be popular and what isn't. All I can do is I can see value. When I look at AI, I see real value being delivered. Best example, PE firms are all over us. They want access to cheap AI compute because every time they get more cheap AI compute, they can bring the the they can change the bottom line of their businesses. It has real value. When PE firms go after something and see value in it, it's not a popularity contest. It's pure value. And so what happens is the the reason companies get a large multiple is people see that the valuation is the actual value is going to acrue or they get hype cycled on it and it's pure popularity contest and there are different participants in the market. Some of them are just playing the popularity contest.

Others are looking at the value and they may come to the same conclusion for different reasons. coming at it from the value point of view, the weighing machine point of view. The most valuable thing in the economy is labor. And now we're going to be able to add more labor to the economy by producing more compute and better AI. That has never happened in the history of the economy before. What is that going to do? Do you worry that if we have a speed bump in the short term, it will derail significant parts of the economy given the concentration of value? Everyone rips today. But if Nvidia, Mata, Google, Microsoft suddenly hit speed bumps and the AI speed train is just slowed down, the consequent multiplier effect is mega. Do you worry about that?

>> Yeah. And this is this is independent of the value of AI. This is the sort of control system um theory of what's going on, right? So a stock market could inherently be on an upward trajectory. it can overheat and that overheating causes it to run away. People bid things up. They realize they've made a mistake and then it has to come back down and then it dips below where it should be. Spending um retreats and then people don't have the funds they need to build their businesses. Uh a lot of good businesses can die during one of these downward trends. But this is also where the best businesses are made. How many times do you see a downturn? Um, and a ton of amazing businesses come out of it.

>> Do you think we will have a downturn in the next year? I >> I can't predict whether or not there'll be a downturn. There are things. So, the ability to predict something is largely dependent on whether or not predictions affect predictions. If a prediction affects the prediction, you cannot predict it because you are whatever your prediction is changes the outcome. The only things that are predictable are things where the predictions don't change the outcome. If an asteroid is headed towards the earth and we see that um if we don't have the technology to stop it, then it's going to happen. But if we see that happen and we can predict it, then we might develop the technology to stop it. Do do you see the problem? I do.

>> And so in the economy, you don't have to do anything other than move dollars around. So you have these very sort of fast twitches in the economy based on people's ability to predict which makes it unpredictable. I can't tell you what's going to happen in the economy. All I can tell you is that right now the biggest problem I see in AI is if you see a good engineer, one that you would have hired before, they can go out and they can raise

10, 20, 100 million, a billion dollars, and then rather than contributing to one of the other AI startups, they go create their own, which means that you have difficulty in getting critical mass of talent in any one of these AI startups. On the other hand, AI is making everyone at one of these startups more productive.

So in terms of whether or not the the economy is overheated, I think one of the best predictors of that is is the economy getting in the way of the success of the companies. If it's not getting in the way, then I don't think it's overheated. >> Do you not think it is getting in the way? Because fundamentally the capital supply side is so uh large that we are actually preventing you from being able to get great engineering teams together because we're funding talent to the extreme where they can raise huge amounts of money rather than join Grock.

>> Yes. Please stop doing that. Um no no but but AI is making people more productive. So, it might be possible for um the economy to keep ripping and for all of the companies to continue being very successful. We don't know. We've never been through this before. >> Is the war for talent insane today? >> It's it's definitely um much more aggressive than it's ever been in history. Uh but only in tech. When you look at sports, um sports have always been insane or at least recently been insane. Like you look back 20 years, 30 years ago in sports, the salaries looked a lot like tech salaries.

>> Sure. >> People are just realizing the value. The problem is in in sports, um you have a limited number of um teams. You have a you know uh you might even institute a salary cap and things like this. In technology, we're not doing that. And you have an unlimited number of teams, an unlimited number of startups, right? Just imagine if anyone could go create their own football team. What would that do to salaries? What would and and what would that do to the the value of the franchise?

>> Which incumbent are you most impressed by and which are you most worried or concerned for? >> Um I would say Google has probably done the biggest turnaround and they had a structural advantage in that. So Google historically has depended more on their engineers to come up with good ideas and as long as management gets out of the way great things happen at Google and so I just think from a cultural perspective that's a systemic advantage. Um and for for them >> you think Gemini has been a success for them ultimately.

>> I do. I mean you just look at the numbers of the adoption it's been great. >> How do you feel about the implementation into consumer products? um less so. I mean, you you see like random Gemini introduction into each product. It's like it's in Gmail, but it's practically unusable. It's in pretty much every product and it it seems thrown in kind of like half thought through. But you shouldn't judge that yet because at least they're getting exposure to how people are using it and they can use that to figure out what they should actually do. I mean, what happened with Google Chrome, right?

Like it was originally Google TV. it was a total flop and then they iterated and they turned it into Google Chrome. Uh this is the classic um problem where where someone puts something out there, everyone throws darts at it and you don't realize that they're just willing to take those darts in order to build a better product. >> And it's fine to take those darts as long as the window of distribution advantage remains. But what's challenging is it doesn't open AI has closed that chasm so significantly.

>> That's true. Um Google may be too late. >> Do you see what I mean? It's like a classic like you know can the incumbent attain uh innovation before the startup acquires distribution and it's like the startup's acquired distribution 10% of the world. It's pretty impressive. >> Yeah. At this point it would be hard to imagine a scenario where open AI goes away. I just I don't see how that happens. And so at the very least you have two competitors from this point on going at it. But >> which is OpenAI and Anthropic or Open AAI and Google?

>> OpenAI and Google. Anthropic does something different. Anthropic is doing coding, right? Um, OpenAI is doing a chatbot. Google's doing a chatbot. Google's also doing coding. Google's doing everything. >> Well, I mean, OpenAI is doing coding, too. >> That's Well, um, yes. And actually, um, our engineers recently started using codecs more than using the anthropic tools. >> Wow. >> Yeah. And it's funny because it's almost on a monthly basis. So, we have a philosophy. We don't tell our engineers what tools to use. We do tell them you must use AI because otherwise you're just not going to be competitive. Um but we saw them um using source graph. We saw them then using Enthropic. We saw them then using codeex. Next month it'll probably be source graph again. It just keeps going around and around in a circle.

>> Do any of these have enduring value then if the switching cost is so low and if they're just bluntly being used so promiscuously. >> Um our engineers are cutting edge engineers who will switch to the best tool the moment it's the best tool. Not everyone is like that. A lot are like that though. >> A lot of a lot of the people you interact with are like that. Enterprises make these long-term deals and they stick with whatever their deal they made a year ago.

>> Would you rather invest in OpenAI at 500 billion or Anthropic at 180? >> I'd want to invest in both. >> Would you? >> Yeah. They're both undervalued. Highly undervalued. You You're still Okay. you're still looking at them as if they're competing in a finite market for a finite uh outcome when they're actually increasing the value of the market with the more R&D that they do. >> Play this out for me then. If we do the bullcase for them, what does that look like? I think the current tech companies can increase their value significantly, but I don't know why they couldn't increase their value significantly while the AI labs catch up to where those current, you know, AI the current technology leaders are. The Mag 7 is going to increase in value and what's going to happen is the the AI labs are going to achieve the same amount of value as the current Mag 7, but the Mag 7 is going to be more valuable. The question is, will the AI labs overtake the Mag 7?

>> What will determine that? >> I don't know. I frankly, I think they're just going to become the Mag 9, the Mag 11, the Mag 20. >> Do you think the AI labs move very significantly into the application layer and subsume the majority of it? >> That is the natural tendency of a very successful tech company. um they start to do what their customers do and they move up the stack and then they create they subsume what their customers did and then there are new people who build on top of them, right?

And um OpenAI uh you know I think on your show Sam Alman said something about how um if you're just uh doing something like a small refinement on top of OpenAI you're you're going to get overrun or whatever. Um he was just being very honest that's what they do. In our case, um, we found an area where we will not compete with our customers, which is we will not create our own models. So, we just won't do it. And by putting that line

in the sand, we're saying it's safe to build on our infrastructure, right? Because we're not going to go after what you do. And that may be the wrong call. We may find that we're subsumed by one of our customers. Um, but it also means that you can trust that you can build on us. I could be making a huge mistake on that call.

>> You could be. You would also need a lot of cash to do that to build our own models. >> Yeah. And speaking of cash, how much did you just raise? >> So, we raised $750 million. >> $750 million at uh what was it? 6 billion. >> Uh yeah. Uh almost 7 billion. >> Okay. Got you. >> This sounds really unfair. And that's amazing. Is that enough money? >> It is. In fact, um we were only going to raise 300 million. Um uh you you brought up, you know, the question of profitability and all that.

Um the hardware companies are are in a good position because unlike uh these other companies, um we actually make money off of what we sell. So we we have um when we sell hardware, those hardware units actually have positive margin. >> I thought you had negative margin. um when we sell hardware uh now >> versus when you sell software. >> When we sell software, it depends on the model. So our most popular models um on the chip that um we're ramping up now um are positive margin. So um but we do have some models that we run that um beat the opex but we're not happy with the capex. Others would be happy with the capex but we're more conservative.

And so it's just easier to say when we sell hardware we have positive margin because you know it at that moment. We might have positive margin on on even our least profitable models because we just don't know how long the hardware is going to last. >> Like what are the margins and where do they go over time? >> Well, one of the benefits of being private is I don't have to tell you. >> You don't. But it'd be lovely if you did.

>> It's the only advantage of being private. >> No, no, no. There's many, many advantages. Um you don't have a lockup period. You can sell much more easily. >> Yeah, but I don't sell shares. So, >> you've never sold a share, have you? Never. >> No. >> Yeah. You clearly don't understand how this game works. Uh, don't worry. I will teach you. >> Um, but like margins over time, do they like get significantly sign like how how do you think about that? I'm not asking necessarily.

>> No, no. I I I'm going to say what I said earlier, which is I want our margins to be as low as our business remains nonvolatile. So the like I said, the only reason for a high margin is because you you want to have the ability to bring in cash when you need it. And all you need is the ability to price higher if you need to in order to be able to lower your margin. The demand for compute is so high that if someone came to us and said, I need this compute and we have it, they will pay a higher margin which allows us to charge a lower margin. Can you help me understand what the chip market looks like in a five-year timeline? You said there we'll have OpenAI, we have Anthropic, will have all the providers having their own chip infrastructure. You'll also have Nvidia, they'll also be What does that look like?

>> My prediction is that in 5 years, Nvidia will still have over 50% of the revenue. However, they will have a minority of the chips sold. They might have, you know, minority share. They might have um 51% of the revenue and they might have 10% of the chips sold. Um >> can you help me understand that? >> Yeah, there

there is huge value in being a brand. Um you get to charge more. However, uh it makes you uh less hungry and you're you're going to start charging high margins and some people are going to pay it because no one's going to get fired for buying from Nvidia. it's a great place to be in.

That business is going to remain incredibly valuable. Um, if you're invested in Nvidia, you're probably going to do okay. However, um, if you're looking at it from the customer point of view, when you have customer concentration like we're seeing where, you know, 35 36 customers are 90% 99% of the total spend in the market. They're going to make decisions less on brand and they're going to make decisions more on what makes their business successful because they're going to have more power to make those decisions.

So, you're going to see other chips being used because those companies are going to have enough power to make decisions themselves. You said you won't do badly if you're an Nvidia investor. One of my u friends says the thing I love about Harry is that you know he's wonderfully charming but at the end of the day he goes that's great that's great but what about me which is very true over under on Nvidia in a 5-year timeline 10 trillion I personally would be surprised if in 5 years Nvidia wasn't worth 10 trillion the question you should ask is will Grock be worth 10 trillion in 5 years possible.

We don't have the same supply chain constraints. We can build more compute than anyone else in the world. The the most finite resource right now, compute, the thing that people are bidding up and paying these high margins for. We can produce nearly unlimited quantities of. >> What do you think the market does not understand about Grock that you think they should understand? >> Oh, it changes every month. Um, it used to be we couldn't have um, uh, it used to be we couldn't have multiple users.

Um, and then we demoed multiple users to people uh, on the on the same hardware, right? They used to think that we >> this is because of the SRAMM structure. >> Because of the SRAM actually here's another one I still impressed with my learning from last time. Thank you so much. Yeah, I dude I learned so much from you genuinely. I was like genuinely learning so much. But okay, >> the the question I get asked the most is, isn't SRAMM more expensive than DRAM?

>> The answer is yes. Um, SRAMM, a good way to think of it is SRAM is inherently three to four times as expensive per bit inherently. Putting all the like, you know, >> and just for anyone who doesn't know again, SRAMM is versus DRAM. Super simple. >> So, um, SRAMM, I'll keep it super simple, but this isn't technically accurate. SRAMM is the memory inside of a chip. DRAM is the external memory. It really has more to do with how you design it. Um but but anyway, um so SRAMM has three to four times as many transistors or or capacitors just transistors for SRAM than DRAM. DRAM is a capacitor and a transistor. SRAM is six to eight uh transistors.

And so SRAM is inherently larger per bit, which means it uses more silicon, therefore it's more expensive. You're also deploying it on a more expensive chip like a 3 nanometer chip. So it costs you more per unit of area than DRAM. So So there's a multiple. Maybe it's 10 times as expensive uh per bit. The thing is, when we're running a model like Kimmy and we're running it on 4,000 of our chips and you're running that Kimmy model

on eight GPUs, we're using 500 times as many chips, which means the GPUs have 500 copies of that model, which means they're using 500 times as much memory, which means that their cost is higher because they even if it the SRAM is 10 times more expensive, they're using 500 times as much memory in the DRAM.

So this is one of those classic problems of looking at it from a chip point of view rather than a system point of view. Everything that we did was actually system point of view and now it's world point of view. We actually load balance things across our data centers. We're now at 13 data centers. We have data centers in the United States, in Canada, in Europe, uh in the Middle East. When you have a worldscale distribution, you don't just make decisions at the data center level. We actually will have um more um instances of some models in some data centers with different um compile optimizations for input or output based on what's going on in a geography.

We may not even have an instance of a model in a particular data center. We may have it elsewhere and we can load balance that. And so we're optimizing at the world level, not at the data center level. What would you do if you weren't scared? Jonathan, >> I'll I'll rephrase that to where could I increase risk in the business? >> Yeah, same question. >> And where we haven't um we could double our our orders in our supply chain. Yes, we have a six-month supply chain, so we can respond to the market faster than anyone else. Um but >> how overweight demand are you in supply?

>> Like I said, last week someone came to us and asked for five times our total capacity. Here's the only reason we don't just completely double down on >> if you're not a supply concern. Why can't you just do that? >> Because um there are thresholds. So for example, if we had double the capacity, we wouldn't have won that customer. They needed 5x. So it's not enough to have twice as much. We have to have enough.

And so if we double the capacity, do we have enough for those customers? >> And so what you the risk that you could take is to what? Sorry, just specifically, >> we could just double um the rate at which we're building out supply. I mean, with this fund raise, we we ended up raising um you know, more than twice what we were, you know, expecting to raise. And then we were 4x overs subscribed over um over what we did raise. And so we could have raised a lot more money. It would have been more dilutive. Um, and I'm trying to be dilution sensitive for investors and everyone else. Um, but on the other hand, we could have just raised more money and we could have just built a ton of compute. Um, the other advantage that we have is versus anyone else, our cost per token, especially given a a given speed, um, is very advantageous. So, we know that we can charge less than the rest of the market. um which matters when you're trying to build these businesses, not because people are are spend conscious. If we lower um what we charge 50%.

People are going to buy twice as much. They're spending as much as they're making because whatever they spend increases the quality of the output. >> Do you think about going public at all? >> Uh our focus is purely on execution right now. um whether or not you go public, you know, that's um like that's a completely different game than we're playing right now. Right now, all that matters is can we satisfy the demand for compute? >> Why do you think Cerebrus decided to go public?

>> Well, they recently decided not to go public. >> That answers that question. Um dude, I could talk to you all day. I do want to discuss um a quick fire around. So I say short statement, you give me your immediate thoughts. Does that sound okay? >> Yeah. >> What's the biggest misconception about Nvidia today? >> That Nvidia's software is a moat. >> CUDA lockin is [__] >> Yeah. Um it's true for training, but it's not true for inference. I mean, we have 2.2 million developers on us now.

That's how many have signed up. >> Wow. >> Yeah. >> How many do CUDA have? >> They claim six million. If you were founding Grock today with Nvidia at 4 trillion and the AI boom in full swing, what would you do differently? >> I wouldn't do chips that that ship has already sailed. It takes too long to build a chip. The the bet that we >> does it. So So for the chip providers today that are coming out, we we are seeing new chip providers come out where they're raising like a lot of money from good people.

>> It's too late. >> Yeah. So the reason that I decided to go into chips. So um I did the Google TPU but um also before I left I set a record on the uh best classification model like ResNet 50 um uh with someone in Google Brain. We we did an experiment. We we beat everything. Um and so I could have gone in the algorithm side. And the reason that um and and actually when we were fundraising I wasn't even 100% sure that I was going to do chips. I was like thinking maybe we we do something on the algorithm side especially in formal reasoning. Uh which is good that I didn't but um the the main motivation to go into chips was the the moat the the temporal moat. So, a question we get asked by VCs a lot is what prevents someone from copying what we're doing?

And the answer to that is if you copy what we do, you're three years behind us because it takes that long to go from the design of a chip to a chip in production if you execute perfectly. I've done um three chips now that are um in production or ramping to production. All three were a zero silicon. Only 14% of chips that are taped out the first time work the first time are a zero silicon. So that means there's an 86% chance each time that you're going to have to respin it. When we built our V2 chip, we actually um uh already scheduled a respin for it and we ended up not having to do it because to our shock the first one worked. Like you shouldn't expect that.

So that 3 years is if everything goes perfect. Nvidia um typically takes 3 to four years per chip and they just have multiple being done at a time. Uh Gro is now in a one-year cycle. So a year after our V2 is our V3 and a year after that is our V4. >> How do you evaluate the meteoric reaceleration rise of Larry Ellison and Oracle? um brilliant business uh decisions and and the willingness to move fast. So um most people right now keep asking themselves, is AI overheated? Should we double down on this? Um they just went for it. They're just aggressive. And that's what it takes to win. When everyone else is fearful, um you should be greedy. And when everyone else is greedy, you should be fearful. Right now, there's a lot of fear in in in around AI. What you're seeing though is there's a couple of greedy, really smart people and they're making tons of money and it it looks like there's a lot of greed out there. It's just a handful of people that are moving fast.

>> Where should I be greedy and where should I be fearful? I'm an ambassador today. Obviously, >> wherever there's a moat, you know, um Hamilton Helmer, seven powers, right? Wherever you see a moat, you should be greedy. >> Very few people have a moat. >> Yeah. And especially at the stage that you invest in. Yeah. >> Yeah.

So, you have to predict that there's going to be a moat. >> And if there is a moat, it's it's a billion valuation for a preede.

>> I mean, there's a billion, you know, valuation for a preede primote. That's what you you should call it primote. That's what the investors should denote it as. Primote. >> What have you changed your mind on in the last 12 months? >> Oh my gosh. Um, I mean, it's not so much that I've changed my mind, it's that I've changed how much what percentage of our business doubles down where. So, every month we become more focused. We we um say yes to fewer things and what happens is the business just does better. So the the I would say I used to think that the most important thing was preserving optionality and now I think it's focus. However, I think having that optionality early on was crucial so that we could play where we would be most successful and now it's about focus.

>> We've spoken a lot about open anthropic. Do you think Elon Musk is able to pull it off with Grock and Axe? Um yes, although it it's probably going to be different. Um when whenever a new area emerges, a bunch of people think that they're competing and they're not. All of these people creating foundation models think that they're competing for the exact same thing. What did Enthropic do that was brilliant? They decided to stop competing um by doing everything and focus on uh coding. And that's worked great for them, right? So, uh, if you look at, uh, XAI, um, they have a social network and they've integrated their chatbot with that. I'm not going to use that chatbot for, um, you know, solving deep uh, analysis or or, you know, deep research problems. I'm not going to use it for coding. Now, they do have a coding model, but they don't have a coding distribution. Can they use that distribution to get into coding? Maybe, but then they're not going to be as focused. So what are they doing?

Eventually the markets will diverge. Mag 7. All of those companies uh have some overlapping business. But the primary business of each of those mag seven companies is different. If you do not differentiate, you die. >> When you look at Google, Microsoft and Amazon, you can buy one and you can sell one. Which do you buy? Which you sell? Um, it depends on the time frame. So, in the short term, I think Microsoft is resetting a little bit um because of the the OpenAI relationship. Uh, long term, they're probably going to do fine again.

I think >> Do you think that's a material damage to them? >> No, that that's why I'm saying in the short term I I think it's going to hit them and then the long term it's not. >> Have they not done majestically well from that? they have the financial ownership of open AAI and then they have the flexibility to use anthropic for most of suite >> and they've deployed an enormous amount of compute. So if open AI diversifies and gets their compute elsewhere, they have that compute now. Um compute is like gold, right? If you have it, you you have AI. Um and then Amazon I think um doesn't have AI DNA and they like if you compare them. So you didn't mention Meta, right? But um Meta and Google always had the AI DNA and uh Microsoft bought it with OpenAI but that bought them time. Amazon still doesn't have that DNA but they do have compute.

Final one. What are you most excited for when you look forward? I like to end on an element of positivity. What are you most excited for when you look forward over the next 5 to seven years? I think the things that scare most people are what excite me. And what I mean by that is, you know, everyone's afraid of what AI is going to

do. And I think there's a good historical analogy here, which is Galileo.

So, um, couple hundred years ago, Galileo popularized the telescope, right? And, um, he got in a lot of trouble for that. And the reason he got in so much trouble was the telescope allowed us to see some truths and allowed us to realize that the universe was larger than we imagined. And it made us feel really, really small. And over time, we've come to realize that while we may be small, the universe is grand and it's beautiful.

I think over time we're going to realize that LLMs are the telescope of the mind. That right now they're making us feel really, really small. But in a hundred years, we're going to realize that intelligence is more vast than we could have ever have imagined and we're going to think that's beautiful. >> John, dude, I I always end up taking copious notes in our conversations. Thank you so much for doing this with me, man. So lovely to do it in the studio and you've been fantastic.

>> Thank you.