How Export Controls Helped Not Hurt China & Power is the Bottleneck to AI | Perplexity CEO

95 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=0XFYVCO1YOW
https://www.youtube.com/watch?v=0xFyVcO1Yow

SUMMARY

Aravind Shrinivas, co-founder and CEO of Perplexity, discusses the rapid growth and impact of his company, which has reached a valuation of $20 billion and 45 million users in just three years. He emphasizes the importance of motivation driven by impact rather than wealth and shares his aggressive business philosophy of always being on the offense. Shrinivas critiques the current state of AI and data centers, arguing that the future will rely heavily on orchestration across models and devices to maximize efficiency and value.

- *Perplexity has grown rapidly, achieving a $20 billion valuation and 45 million users in three years.*
- *Shrinivas believes motivation should stem from making an impact rather than financial gain.*
- *He argues that Perplexity has influenced Google more than any product manager at Google, particularly in redesigning its search interface.*
- *The future of AI lies in orchestration across various models and devices, not just in building better models.*
- *Shrinivas identifies power supply as a critical bottleneck for data centers, predicting resistance to their development will continue.*
- *He suggests that companies will need to adapt to a landscape where open-source models and local computing become more prevalent.*
- *The conversation touches on the potential for significant shifts in the job market due to AI, emphasizing the need for a positive narrative around technology's impact.*
- *Shrinivas envisions a future where companies can operate efficiently with fewer employees, leveraging AI to enhance productivity.*

I have nothing to lose.
I came from nothing. I never even
imagined myself to be doing all this.
>> A $20 billion company,
>> $45 million users,
>> over a billion searches a month,
>> built in 3 years by 400 people.
>> These numbers like doesn't motivate me.
It's hard to get motivated by wealth.
You want to get motivated by impact.
>> This is perplexity with co-founder and
CEO Aravven Shrinus.
>> No one's ever in a comfortable position

that no one can relax.
>> They forced Google to redesign their homepage. then bid $34 billion to buy Chrome.
>> More than their own valuation.
>> Perplexity changed Google.com more than any product manager at Google has ever done. Now you look at AI mode, it looks exactly like Perplexity.
>> He doesn't do defense. He doesn't do comfortable. His words, attack, attack, attack. That's my motto. Go all in and try your best. Be on the offense all the time.
>> You know what I hate with podcasts? When people sit on the fence. Aravind has really strong opinions in the show today. He says that Micron will be more valuable than better. He says that the resistance to data centers will continue and get worse. He says the biggest problem today is a lack of power. He claims that perplexity has changed Google more than any Google PM. You want opinions? This is the show for you.
Ready to go
Ara. Dude, I am so excited that we get to do this. We've done one remote and then we did one at Founders Forum last year. So, thank you so much for joining me in person.
>> Thanks a lot, Harry.
>> Dude, I It's a weird start, but just roll with me on it. I asked this of the best founders that I meet. Are you motivated more by the fear of failing or by the thrill of winning?
>> Thrill of winning.
>> Why? Because I have nothing to lose. I came from nothing.
Like I I I never even imagined myself to be doing all this. So my life has

already been extraordinary u beyond any level of imagination. Um I was just in India like you know doing my undergrad and you know just just training neural nets with graphics cards that people in the labs were using for playing video games. It was all for fun and um you know my path led me all the way here. It was never like a mo my for my mom just getting a job was success because we were not we were financially lower middle class in India which is not even like lower middle class in UK or the US and so from there all we wanted to do was get a job in Google being an engineer at Google was considered a win and so I'm I'm already doing remarkably well compared to that ambition we had as a family so there's really nothing for me to lose. That's why anytime I try to act like I'm trying to avoid failure, I'm being on the defense. I remind myself that like that's the stupidest thing to do. Like you know, you you it's better go all in and try your best. Be on the offense all the time. Attack, attack, attack.

>> When you review then what are you not being aggressive enough on today? Maybe in the early days we be very very loud on social media talking about perplexity versus Google and I used to do that myself a lot. So and some people don't like me for having done that. Today I'm a lot more measured in how I talk about our products or competitors and stuff like that. But it's not a lack of aggression or anything. Um it's just that like it's that that is boring. People already heard that enough from me.

>> Do you regret the being so bold in your

messaging?
>> No. So it's not a nuance and maturation of message. It's a that stale and I need something new.
>> Not just that, I I kind of don't think it's a relevant framing anymore. We worked on search. Perplexity started out as search. We built the first answer engine in the world that people know perplexity even today if you mention the name perplexity people would think oh that's an answer engine. We built a lot more things after that. We built a lot of agents, browser agents, deep research, computer. We built so many products after that but we're still known for that first product and um the mark has already been made the we changed the road map of Google. You could argue that I or the company Perplexity changed Google.com more than any product manager at Google has ever done.
>> Make that argument for me.
>> Well, they never that nobody ever wanted to ship an answer engine at Google. Nobody like nobody wanted to tinker anything on on the interface that made them $250 billion a year. And uh and then now you look at AI mode, it looks exactly like perplexity. There's there's not even any difference like the font, the citations, the specific bolding of inline text, inline hyperlinks, um suggested follow-ups, the whole experience is literally looking like perplexity except it's still not as good. And so
>> is that bad or good for you that they learn from you and adapt?
>> It's it's it's it's both good and bad in the sense, you know, you have to

obviously we I knew this like around end of 2024 this is going to happen. So, it never caught me by surprise at all. Um, it was just a matter of time. I still am I'm am surprised that the quality is still not there cuz I I regularly test every product out there. And u but I'm I'm happy that honestly uh they they they changed Google to be what it should be. And um I believe that the frontier is where the money is. The frontier in AI is not about answering questions anymore. It's about actually going and doing work for you. We, you know, like we still have the state-of-the-art deep research the world will. And that's actually where people subscribe to pay for our pro or max products is not for getting answers in in the in the traditional way. They're asking for sophisticated research reports. They're asking for agents that go and do things for you. And so we wouldn't have been able to do all that if we were sitting in 2024 thinking we have everything settled here. We're we're good and comfortable. No, we it the answer engine was always a lead genen for the frontier products we build. You need something, right? Like think about it. Every company needs to have one successful product to build the next set of products. And in AI, nobody can sit comfortably thinking they have it all sorted out, including Anthropic. If Anthropic thinks cloud code is already a win, in 6 or 12 months from now, they won't even be around. And so that's the uncomfortable it it's it's it's it's an uncomfortable fact about the whole field. Would you argue today, you just told me, if you don't mind me quoting

you here, um, you just told me before we started, that you think OpenAI isn't ready for an IPO.
Um would you have believed you would be in a position to say this two years ago when nobody h wanted to deal with any product other than chat GPT.
Think about it. So anyone even in such a massive advantages position can be in a can be put in a position where they're no longer the kings. They're fighting from behind. Right? So that's the state of the field. That's just it's less about perplexity or anthropic or open AI not having modes or having modes.
>> Can I push back on you that I would I I would stand by two years ago even when they were a do and they are still a dominant consumer product but I would stand by it because I don't think they are financially ready when you look at the balance sheet of that that
>> maybe maybe I'll decouple that. I'll decouple that.
>> Let's decouple that being like financial readiness for an IPO
>> versus perception of a dominant leader.
Yeah.
>> Do you perceive them as a dominant leader right now?
>> Yes.
>> In what?
>> Consumer search.
>> Well, except there's no money there, right? Because it's been commoditized.
So, it's it's always a legion. Like, for example, why why why are they going all in on Codeex? Cuz that's where the money is. And uh we're doing the same on computer. Anthropic is doing the same on cloud code. Google doesn't yet have a product in this category, but I'm sure

they're going to come after that. Meta is trying to launch Hatch for $200 a month. You You see that? You see what's happening, right? So, nobody
>> But there has to be more money than just code codeex claw.
>> It's not about It's not about code. That's the main thing. The the money at least in non-advertising.
I'm not talking about advertising revenue. In non-advertising subscription or usage based revenue, the money is in whatever is the frontier. And today the frontier is about doing going out there and doing things for you. And uh
>> do you not think then that there will be a 100 to20 billion advertising business for open AI? It's
>> yet to be proven. Let's let's work through the categories of advertising. Um who's the number one advertiser on Google? Amazon. It's a number two. Booking.com.
Number three or four I think Expedia.
So, um, how much do you think Booking.com spends on Google? 16 billion, something like that. Something some some some crazy amount like that.
Um, and, uh,
um, how do you book your hotels or flights today? Do you book it on chat GBT or do you book it on Google?
>> Google.
>> Why is that?
>> For me, I actually like discovery. I would like to see the options.
>> Exactly. Right. So the interface the interface is less about conversations and more about exploration.
So when when the decision making is more subjective and vibes based, you don't need an objective

answer engine. And and so it's it's it's and and and you think about the other category of advertising, direct to consumer products, fashion. Where is most of that advertising budget going into? It's going to meta, Instagram, because you're just browsing. You're just like doom scrolling or whatever they call it, right? And so, uh, the chat interface doesn't capture that user intent, that user behavior right now, which is why it was never a great fit for advertising. And u, it also fundamentally corrupts the trust that people have when they go into a product and they want the accurate answer, which is what, you know, perplexity is known for. Um, and then you're like, "Hey, by the way, you know, you ask for the most um highest like like mo best protein shake, but by the way, these are good protein shakes you can check out." Like it it it kind of like hurts the trust that people have in your platform and your product. And so, um, that's another reason why, if you think about it, like like what Meta or like I think some other companies in the past have tried to put ads inside, um, messaging apps and emails and it's never really worked out. Um, it it works out in China in WeChat because there's no other way for them to fund the whole thing, you know. So the the the whole economy and and and user sentime user behavior has been optimized around gamifying. It's not how things work in America. So I I'm I'm I'm bearish on advertising to really take off in in the chat interface. I I'm happy to be proven wrong there, but I'm bearish on that.
>> There there are two areas that I want to

unpack there. The first and just taking them kind of chronologically and how you said them, money's in the frontier. The more I hear this kind of the more I question it because I think that we dramatically overestimate how important frontier models are to do quite basic work.
>> Yeah. So frontier doesn't mean a frontier model. Frontier just means whatever is the frontier outcome you can have right now with AI.
The Greg Brockman recently tweeted the model is no longer the product, right? Um, and it's funny because you you know that as as a leader of a frontier lab, he has all incentive to say the model is the product and and that's what Google people tell. I think one of the Google people keeps tweeting that model is the product. I forgot who. Um, and so the reason Greg is right is because um, if you take codeex or perplexity computer clot code, what is that? It's it's an orchestration system, right? It it takes a model, pairs it with an agent harness.
And what is an agent harness? Think of it the simplest way of describing it is like rules for how the agent loop should run. What are all the skills and sub agents and connectors and tools it accesses? And uh without the harness you don't necessarily capture and convert the intrinsic intelligence in the model into valuable output tokens.
The output tokens if you're if you're literally just a reseller of model tokens you have no business because the model will get commoditized. So even if you're a model builder, you don't have a business. As an infra layer, you have

some business on serving those output tokens. But as an application layer or a model builder, you don't really have a business. If you're just a reseller of tokens that come directly out of the model, you have business if you know how to take the model ground it in valuable context, orchestrated with a really good agent harness
um connected to the right set of tools and connectors whether it's personal connectors or business connectors and provide the experience to people in one single unified system.
And uh the way we differentiate ourselves at perplexity is we don't just orchestrate across tools and files and connectors. We also orchestrate across models.
That is the differentiation that anthropic and open AAI cannot claim because you wouldn't find GPT5I inside the cloud code harness. You wouldn't find claw opus 47 or 8 inside the codeex harness. These are competing with each other, right? Whereas you would find both these models inside perplexity computer and that way we can bring down the we can increase the token value per watt per user. If you assume that
if you assume that whatever decides the dollar like the price the dollars is the power watts fundamentally that's that's the thing that nobody else can subsidize other than the government. Um you know that whoever pro you know provides the most valuable output tokens with the least amount of power expended to produce them generates the greatest value to the end user and has the most pricing power has the most value and so

that that that is the orchestration problem to solve who the one single the most important metric in AI is token value per what per user. What does it mean for the value of open AI and anthropic? If model is not the product and it becomes a utility, something you can switch into and switch out
>> interface.
Everyone thinks we're all building the model layer or the race. We're not actually. Um, I would even argue that building models is a way to stay at the frontier, but you have to own an interface
in which valuable AI output tokens are generated, the most valuable tokens. It doesn't have to be the product. This is the single most important thing to like you know unlearn for most founders and I had to do it too which is to be successful in AI product layer whether you're a model builder or not it's not about building something that gets a billion users that mentality has to completely shift
there are a few power users who are propelling this token economy right now if you look at like all these um crazy stories of how there's this one engineer who got Amazon to spend like half a billion dollars in a month because of some stupid way they set up like agent loop inside cloud code. Okay, maybe that's a mistake, but there are real engineers in meta in in other companies spending like 10 million a year per engineer on on on these, you know, coding tools. There are users in Perplexity Computer. Um, there's one user, I think, who spends upwards of like $10,000 a month, something like

that. Crazy. And and and not like wasting it. They're not wasting money. Their business runs using agent loops that are running inside these harnesses. And they use these products in sophisticated ways that I I I couldn't even conceive when we were building the product ourselves. Even internally inside our own company, there are some people who have set up these kind of like multi- aent hierarchy and agent loops that looks like its own software architecture.
And I often just ask these guys to come explain to the rest of the company, hey, like what are you doing with these tools? Like you clearly are consuming it way over, you know, what we thought the average person in the company would do. And the single biggest differentiation between those who use agents a lot and those who don't is whether they run repetitive cron jobs
like whether you use AIS as one-off tasks. You just delegate a task and then it gets done. That's like kind of using it for like deep research or like whatever, right? Like one single task versus the AI is like continuously monitoring something for you. The AI is continuously like triggering based on certain events and going and doing certain things, giving you alerts like you set up workflows that keep running for all the time. Every time you get an inbound email like it triages or every time there's a latency spike, it has to identify which part of the codebase caused that, it has to go and do the root cause analysis and then identify the right engineer. all these things the this is where the frontier is and and so

uh going back to my main point.
These products are not going to be used by uh you know 100 million people but they will generate revenue that's going to be higher than the advertising revenue of Google or Meta. It's going to happen.
>> Completely understand what you say there. I do just want to focus in on a specific element there when you were saying like the power users because I think one of the core numbers is actually Mark Benov said they spend 300 million on anthropic which works out to be about
>> it'll be interesting to know from him if that 300 million came from you know what is the distribution across employees
>> so it works out to be that was on developers within Salesforce so it's about 3.8% 8% of developer salaries. What percent of developer salaries do you think will be spent on tokens in 24 months time? Cuz that fundamentally changes the value of open air and anthropic. If it stays at 3.8%.
They will not be $5 trillion companies. But if it's 100% like Brandon at Mccor said it will be in a year, they they'll be 10 trillion companies. Well, um I think they can certainly beat $10 million companies whether it's going to be um a full percent of the developer payroll today or not because there's a lot of non-developer
work that'll also be done with a agents. Um and that's actually what we focus on for Perplexity Computer. We're not going after the developer market. We're going after anything that developers don't non-developers do. basically um your your finance department or your corp dev

or your like um sales reps or your data science teams um your your research analysts I think that's actually even bigger market that it's it's it's not even like like think of it as like clot code multiplied by 10 that that's the size of that market
>> if I push you on developer salary spend what percent of token spend as a portion of salary do you think We'll see in 24 months.
>> It's hard to say.
I think the costs are going to go down. That's why it's hard to say.
>> You think the cost will go down? Cuz this is the kind of the challenge that we've had. We thought when we went from chat to agent that costs would go down and token costs would go down. They've gone up.
>> Yeah. For now.
>> Help me understand that and how that changes.
>> I think in software um you kind you kind of want to pay for the frontier. Um it's kind of like if you know some engineer is awesome. If you know you have like the next Jeff Dean,
would you rather hire that person and not hire five five people who are medium engineers but not Jeff Dean level with the same amount of budget you have? Yes. Right. Let's say you had a million dollars. You could hire five people worth 200K or you could hire one Jeff and pay them a million. What would you do?
>> One Jeff.
>> Yeah. So I think you would pay for the frontier. Um but what stays frontier keeps changing. Um in in 12 months from now let's say thought experiment there

is an open source model as good as Opus 48.
>> Mhm.
>> And um you still have to pay for inference. You know nothing is truly free but it's going to be like let's say 10 times cheaper than Opus 48. And when you pair it with the right agent harness,
you know, and all the connectors, GitHub, everything, all your developer workflows work fine, why would you um assume that the token spend is going to be still high? It's not going to be for the same things you're doing today. It's not going to be. But there might be a different set of things you might do with the frontier that you're not conceiving today. Uh my prediction would be soft a agents that are like completely autonomous software engineers. Today I think we we're all using tools like cloud code or codeex to write code but not as literal software engineers. There is a large sway of people that is now bearish on your frontier models who open eyes and your anthropics because they're realizing that you can actually do a lot with open models for a fraction of the price. What you're saying is actually that is true but
>> we will still pay for the frontier and so they will still ac
and and I think this distinction it feels like a contradiction. It's not though. It feels like two things cannot be true simultaneously.
But that that's not quite the case. In fact, I would argue that the frontier is increasingly going to be a thing that um very few individuals might even want.

Like you could argue that after a point like it's not even interesting that AI can write software. You you we've normalized it, right? Let's say let's say that that's going to be the case. Instead of companies being built with like tens of thousands of software engineers unlike the past, there'll be a lot more companies with smaller software teams and each of us will be using a lot of AIS. So um that's actually good for the world. We'll be seeing a lot of different businesses. We'll be seeing allocation of software labor in places that was never even possible. and and uh whatever is a frontier is going to be things that kind of like AI is going and designing chips, AI is designing drugs, AI is figuring out how to build robots, AI is figuring out how to cure cancer. These are applications where you don't have like 10 million users. It's like a few companies, but the effect of that work will touch a lot of human lives. I think to me that that's where the frontier is headed. Um you could also see that from the moves that Frontier Labs are making. Anthropic bought um a wet lab could be for the talent could be for the infrastructure to run like wet lab experiments but imagine taking all those tokens and putting it in the mid training instead of just tokens from GitHub right um so then that's going to produce something interesting.
>> Don't laugh. Is there an asimtope to frontier problems to be solved? I know that sounds ridiculous, but if you are continuously on the chase for the next frontier problem, you get to cancer, you get to climate change. And my word, I hope they solve both in like heaven,

that's a huge amount to solve.
But if you're on the treadmill of
continuously, is there an asmtope to
that? Do you see?
>> There's no there's no mathematical
argument
to there being a cap on the amount of
economic value
one can create with with with um AGI or
ASI like systems. Um and Elon Elon has a
good argument for this like like where
he says money loses all meaning in a
post AGI economy
because you'll be producing
an abundance of energy and labor
and fundamentally the economy is
grounded to energy and labor. If you can
produce an abundance of them,
well what what meaning does money have?
Um and um and so I I don't think we run
out of things to solve at the frontier.
I think we're always going to be creat
like like why why would why did people
even want to understand the universe?
Like like why did we want to understand
subatomic particles, quantum physics,
black hole theory, um you know the
origins of the universe like what what
what is the purpose? But we still went
ahead and did it because that's kind of
what the purpose of humanity has always
been to understand the unknown. You
know, David Deutsch is famous for saying
this, right? Like we are the only
species capable of being curious about
what is already familiar. Like you can
stare at a fruit and you know that it's
a mango and like you know exactly like
how it tastes, you know how it looks,
you know the shape, you know what
seasons it grows and and and stuff, but
you can still look at it and ask one

more question about it that you haven't asked before. Other animal species cannot. Once they kind of once they have it in their mental model, what it looks like and touches and feels like, they're going to ignore it. It's not it's no longer interesting to them. Kashi, you mentioned about agent usage and you said if you do repetitive tasks versus one off say chron jobs, you know, I think Sam said it's we're going to have 24/7 AI and um yeah, they've talked about a hardware product that's going to come out.
>> Do you think we will have continuous agents running?
>> Yeah. In so
>> and I think that's kind of why I believe
>> the orchestration problem I I talked about maximizing the token value.
>> Can you just help me? Sorry. when you say the orchestration problem.
>> Yeah. So, so okay. So there are like four objectives
um accuracy, intelligence and accuracy and then privacy and cost.
You know these are all competing with each other. So you can you could argue that um you could max out on intelligence and accuracy by building giant giant data centers and spending a lot of power to uh you know run them and u you could miss out on privacy and costs cuz everything will be centralized and and and you're going to be paying a lot. Um, you could argue that everything can run locally and so that'll be good for privacy and cost but may not be frontier intelligence, may not be frontier accuracy. So the solution is to figure out a sweet spot. You know, use local models when necessary, use server

side models when necessary and orchestrate across local models and serverside models. Uh, grounded in valuable personal context. Sometimes the intelligence might already be there but the system might not work that because the harness isn't grounded in the right set of tools right so build a worldass harness that can even make an okayish model appear great and be able to use the right model for the right task and the right part of the task sub aents and and even like utilize the compute we all have in our own devices all that that you know doesn't need to be always on a server that is an orchestration from a router, an awesome router, a master orchestrator router. Now, um, if you do that, you can realize the vision of a 24/7 AI without people freaking out about going bankrupt
because no one's going to be able to afford a 24/7 AI, Frontier AI, running on the server. Imagine you turned it on and you you could never switch it off unless something crazy happened. Um, you're the the the the the thing that most people worry about those AI is like, "Oh, what if it does something crazy?" But the real concern actually is the cost. Um nobody's going to be able to afford it a cron job at the fidelity of few seconds, you know, um that that runs all the time. And so, um the bottleneck there is actually orchestration and local compute. And so I believe like like um one needs to build a continuously learning
local model um that can save you on like compaction
context windows. So you and and and try to preserve as much compute locally and

rely on the server side frontier only when necessary and keeps learning, keeps adapting, keeps evolving. And that model um is not just a model. It's a model plus the harness plus the local chip and the compute and the ecosystem of devices it controls. That system um is going to be your own intelligence. Essentially the data center moved to your local device and you you get to control it. You get to own it. You don't get to worry about somebody like you know spying on you or looking at all your tokens. Very valuable personal tokens. Imagine you have like very sensitive deal materials. Let's say you're doing a deal um and then um a Frontier Lab has all your tokens that you use to like write a memo.
Imagine somebody could hack into that server and steal your deal from you. You wouldn't want that, right?
>> I'm going to be honest. I think there's much more valuable things for people to steal from
London based VC.
>> But yes, I can figure you're not just yet another London VC. You have like a $400 million fund last time read it. So
>> imagine like you know you're
>> already making your moves for the $4 billion fund right so so everyone has certain levels of like you know sensitive stuff and and so I think that's where I believe that the 247 always on agent is going to be realized by the company that wants to play the role of the orchestrator not the model builder not the frontier model builder but the orchestrator and uh and I think that's what that's what we want to do um computers has been positioned explicitly

as the agent orchestrator. The the musicians in the orchestra are these sub agents that utilize these different models. Think of them as the instruments and uh the tools, the connectors, the models. These are all the instruments and the musicians are the sub aents and the symphony is the work and the system is the orchestra and and computer is the orchestra conductor. that that that's how it's been positioned. So what it orchestrates
keeps evolving, right? It it changes. It changes from,
you know, models to files to tools to chips to devices. But but it doesn't even matter like you don't care as long as it orchestrates things correctly and and and and maximizes the token value for what per user. If you can solve this problem, you will capture the most economic value in AI long term. Shortterm it might look like oh like this other lab's revenue is growing you know exponentially this that but long term this is the one objective that truly matters.
>> Who is best positioned to do that?
>> I believe it's us cuz you have the incentive of not token maxing you have the incentive of delivering the most value to the user like we we every time any part of the AI stack improves our product improves. Um since the beginning of the year, Anthropics models have made tremendous progress. But what's also true is that our revenue has more than tripled since the beginning of the year. Tripled since this beginning of the year and uh we and a lot of thanks to model progress made by anthropic and we also brought our

burndown thanks to OpenAI competing with them and bringing down the cost of the same capability. And now with progress in open source and local models and local chips, we're going to move some of the inference back to the local devices and bring down the cost even more. So every time any part of the AI stack, whether it's chips, models, harnesses, any of these gets better, our system improves tremendously. And if our system improves tremendously, our users love it and they pay more. They spend more and so our business grows. So I think to your question of who's best positioned to win in that world for that objective of being an orchestrator is the one whose product or business benefits from other people's progress at any layer of the stack. And so if Jensen produces a better chip, it's great for us. If Dario produces a better model, it's great for us. If Apple produces a better device, it's great for us. And like I I love the fact that we are able to be a very positive player at every layer of the stack and not have to rely on any one person to win. When we look at the different providers that we said kind of server side versus you on device when we look at server side a lot of people talk about an AI infrastructure bubble which I think is funny stupid and moronic. To what extent do we have a data center supply problem today from what you see? I think the biggest problem is actually in power. So what let's break down what is a data center. Is it like that you just buy like a bunch of chips from Dell or Super Micro and No, that that's just one part of it. You actually have to go secure land or

you have to lease something, lease a property and uh you have to buy a bunch of turbines to generate power or you have to work with like power suppliers, grid suppliers and uh you also have to work on cooling. So there's a lot of other work you got to put in that is far far slower. you have to get permits to do all these things and uh and so usually what's happening is um there's a lot of lead time to doing this and um the models that are um already in use today these have been trained in the hopper generation so the blackwell generation model I think the first model that's blackwell generation category is um mitos
and it's already scary like people are already like freaking out about it So imagine that um everyone pre-trains a model um on like a million or like you know hundreds of thousands of black wells now those models are going to be far more powerful than what exists today and then the ver rubins are coming next year in full capacity like like all the data centers of wear rubons will be in next uh you know used next year that model will be even more powerful. So I think we there is a certain physical buildout time
that always bottlenecks frontier capabilities. That's why there's a value in that layer. Whoever knows how to do this puts it puts together a bunch of GPUs and chips and networking and power and cooling and actually like orchestrating all this software layer on top and you know is able to convert that into frontier output tokens. that that that vertical integration has a lot of value. So that's why the markets are

pricing infrastructure companies with a higher uh PE ratio
>> than companies like Meta for example. Even though Meta builds a lot of infra is valued as a software company. When we see like you know Meta's capex spend and it wanting to increase in the last few days and thinking about raising more and more money to increase capex spend I get it with a lot of the AI providers like your open eyes or Anthropase because they are making money from their AI products. For Meta, the capex spend correlates to increasing accuracy on ads, which is like a six to eight% bump in revenue. I get it. But for the capex spend, it doesn't make sense.
>> Well, um I I I believe like they they are understanding
what the market's saying. You know, I don't think they're
dumb to not see what what what's being said. I think they're introducing a lot of subscription products um from what I'm reading. So they're definitely going to like basically the company needs to not just be a social platform maximizing engagement and turning that into ad revenue, right? And I think um that requires them to launch a lot of like agents subscription based products and maybe even a cloud meta cloud that that rents out servers like what Elon's doing at SpaceX and and maybe once they do that
the the narrative might change, right? But um to go back to my point, it might not be inconceivable that
um
Micron, the supplier of HPMs, might be more valuable than Meta in the next 6 to 12 months. It's already at like a

trillion and Meta is like 1.3 to 1.4 trillion.
>> Can you help me understand that? Because memory is already a massive bottleneck. It's increased 5x in price in terms of the cogs, right? Um, but people are going, "Wow, Micron is fully priced at this point." Why is it not fully priced?
>> Because it's still the bottleneck. Whatever is the bottleneck will command the price.
Um, AMD is doing really well because CPUs became a bottleneck again. Agent loops, agent harnesses are all running on CPUs. The tokens are produced by the frontier models on GPUs. But whatever work like let's say like claude generates a coding script that decides to download 500 files from different websites and then you know munches a lot of data and transforms it in certain ways and generates a plot and then hosts it on a website that you can you can share with your people. All that computers is running on CPUs.
Agents are using CPUs more than humans, right? And so suddenly there's a rise in enterprise CPUs and the beneficiaries of these are like Intel and AMD. So then they get to be the bottleneck like like whoever's going to be the bottleneck will win and and so infra is the bottleneck right now because there's a lot of demand and we just don't have the supply and so whoever supplies memory SSDs for storage CPU compute suddenly these are all like interesting like um they're more important than companies that are just building data centers and not knowing how to turn that into a valuable outputs. Do you believe your Nebius and your Core Weaves will be a

sustainable
multiund billion company in the future
or is it solving a short-term supply
problem?
>> Um I certainly think they can be
sustainable.
>> Yeah.
>> Um I think there are some I don't like
look I don't know particularly which of
those is going to win and there's also
other players like Cruso and um Firebird
and a bunch of companies. It's all about
being resourceful. You got to take power
from areas where there's a lot of
natural resources
and the cost to bring up the data center
is pretty cheap and the time to bring up
the data center is cheap and your
service is reliable. Like if somebody
commits to buying 100,000 GPUs from you,
um the service should be pretty good. Um
and uh you should be able to secure the
supply ahead of time. Plan well. Um and
I think some companies are even
innovating at the power layer. You know,
um generating their own power is one way
to bring down the margins. Uh and so I
think there's certainly like value in
that layer because um it's hard to
replicate work. That's how I see it. You
could argue that OpenAI can do all the
work that Core V was doing and that's
kind of what they wanted to do with
Stargate. But why is Corev more
successful at building data centers than
OpenAI? It's
>> hard to do. It's operationally
intensive.
>> Yeah, operationally intensive. You got
to focus. You got to like spend most of
your time um securing permits like
figuring out power, figuring out like

bottlenecks in the supply chain here and there um and constantly plan ahead and like test all these systems carefully. deal with like random physical issues that you know arise in like you know running a data center there's something called TCO you know cost of operations
>> you got to factor that in so uh that said I I I don't think there's value um if you're just like a server renter if you're just a GPU server rack renter if you're just leasing it to different companies on certain hourly pricing rates there's not a lot of value you have to actually build some software on top kind of like how AWS did. It's called Amazon Web Services, not Amazon servers, right? So, um you have to have some software orchestration on top that allows you to get software margins on top of what you're doing. And I think that's why you're seeing moves like NBS um
like like going for the AI model inference like you know taking open-source models or hosting your models and and and um that's a business model of certain other companies like fireworks and you know um ben and all that but you could imagine neocloud just going for that business.
>> That was exacting me my question. So, I just had the co-founder of Nebius on the show and the really clear takeaway was the the challenge that he has, which is there's a huge amount of money that wants just capacity and compute.
>> Yeah.
>> With the awareness that he needs to build a full stack product if he wants to have a long-term sustainable business. That was the core realization

for me. When I look at the inference layer, like you said, fireworks or base 10, how do you think that plays out? Do we have standalone hundred billion dollar companies in inference alone or do we see that commodity?
I mean you it's just it's all about working backwards like what does it take to build a hundred billion company assume like
>> 10 billion in revenue
>> exactly 10 billion in revenue 30 to 40% gross margins good amount of net income good cash flow okay 10 billion in revenue
um is not that inconceivable for for a company that can both do AI hosted inference and server capacity and data center buildouts
very operationally well. It's all about like you know there there are some factors beyond their control like open source models continuing to be awesome. If open source models stop to actually be good where the gap between them and the frontier is like more than 12 months or like 15 months 18 months then I don't think these companies really have a business model because um they're not going to be able to host they're only going to be able to
rent capacity to open Athropic and so um that's exactly what Roman and Nebia said he said if consolidation happens and there's anthropic and open AAI or two or three dominant providers that is the biggest threat.
>> That's correct. Yeah. And so um but you got to make a leap of faith assumption that you know like the the models from China or Nvidia is making good progress on their models in Limatron. Um you got

so there's going to be enough factors in the market to keep u consolidation as an outcome from from like stopping from happening. But you don't control your own destiny if you're those companies. That that that's basically the problem. Totally get that. Okay. So, we can have standalone companies that are hundred billion dollars in inference alone. So, I'm just pillaging you for your knowledge. When we look at the model selection companies like an open router or like factory AI just released that kind of model selection or model rooting product which did very well on launch, is there hundred billion dollar companies in the model selection and routting business?
>> Probably not.
Um I think you can't just be a provider of router. You have to use the router to produce something meaningful. Um actually most of the business value of open router is less than the router even though the product is called open router. It's not routing across models there. It's actually just routing across different endpoints of the same model. So um why okay so maybe let let's let's let's ask this question. If you wanted to use claude opus or um I don't know like GBD55 developer why would you not want to just use it with your own API key versus using it inside open router? Number one argument. The single simplest argument as to why you would want to do that is model fallbacks. Sometimes your API keys might not have the rate limits or even if you have the rate limits it might there might be an error on open AI servers that you know don't guarantee

you the response time you need to run your application and uh open router would go and earn the uh you know they would pay for capacity for like one year ahead uh with the funding they have and secure the rate limits and multiple endpoints across multiple different providers of openi models be it Bedrock or Azure or OpenAI themselves. And so that routing is valuable. It's essentially an infra problem they're solving which is reliable token supply. It's not actually oh like they're lowering the cost by deciding if this prompt should go to like GPT or clot or something like that. That's not what they're actually selling to the developer. That's not actually the business model. And uh and then for a lot of these Chinese open source models, there's not you probably don't want um your API tokens from going to like let's say you don't want your API tokens going to China. Um and so you and and let's say you don't have the bandwidth to work with like different inference providers or verify who's good and you know who's not. You're just thrusting open router to take care of all that and then you know they're going to like supply the tokens to you. So it it's it's routing not at the level of like oh like deciding which model is cheap for what task. It's more like um a reliable token supply and I think there's some value in that layer definitely uh otherwise they wouldn't have these many um users and these many trillions of tokens being routed a month but it's um it's not like you know high gross margins business it's the way the business model works for them is actually um they would

secure a discount from the model providers by guaranteeing a lot of supply but they would still charge the user their listing price on the API and that difference is their margins. Do you understand?
>> I I I totally get you. We spoke about bottlenecks and you said about HPM, high performance memory and micron and the value that you know they have to say and what it can be. What bottleneck will we have in 3 years that we're not discussing today?
>> I think power will remain the bottleneck. Yeah,
>> I think it's it feels like that to me. Unless something dramatically changes in the way data center buildouts happen. Uh I actually believe that there will be a lot of resistance to building data centers. It's because people incorrectly think that data centers consume a lot of water or eat up a lot of power which is both both are untrue. Satya even made the statement that it it's like a can of water or something uh in terms of how efficient these companies are.
>> Do you think that's why they're putting up resistance to them? I don't. I think it's cuz it's a symbol of job losses. Uh increasing wealth in
>> it's a lot of things. It's a lot of things. Um it's a lot of apprehensions um fear about like what's going to happen channelizing in so many different ways. Um, sometimes it's channelizing through hatred for wealth inequality and like wanting to tax people. Sometimes it's channeling through like concerns for the environment and like climate change. Um, sometimes it's uh channelizing in a way where um you're

all like oh like the price of the grid is going up because you guys are building all these data centers and then or like I'm paying more for my phones and laptops now because the RAM prices have gone up because you guys went and bought all of it. So I think there's a lot of different ways in which it's getting channelized but the common sentiment is um like like a pretty bad sentiment about AI.
>> Do you think it will be meaningful to the development of those data centers? I think right now 40 out of 100 are not being developed because of public resistance.
>> Yeah. So um that that's where the power bottleneck is and um you could see maybe certain countries sees the opportunity for this and um
um allow these model builders to go build data centers there. Um Elon's going to space to do that. Um so that's going to be an interesting experiment. Um because there's a lot of energy from the sun that can be harnessed there. and uh there's a lot of natural resources in other countries. Regulations might be more friendly.
So, we're still going to see data center buildouts. It might not happen in the US. And um but but the fact that you have to solve physical problems like you actually have to deal with the supply chain, the permits, securing power, like making sure like things work and getting the lead times lower and lower. You're not solving problems like cloning some SAS apps here, right? You're or like you're building a go to market team or like um doing better marketing against the competitor's products. Like yes,

those are also hard problems, but these
are like much harder problems where like
you're not in full control of your
destiny and you need a lot of capital
and connections and like the right
people uh sometimes even like political
help to unlock progress. And so that's
why this will continue to remain the
bottleneck in my opinion. And there's a
lot of risk as well because if you do
encounter another Deep Seek moment here
where there's a vastly more efficient
model that's been built with a very
different vertically integrated
architecture
and you built out all this capacity and
you're like, damn, that's I overbuilt.
there's something far more efficient
that can run on on on people's local
devices, their MacBooks, their Windows
PCs.
Yeah. Like you're probably freaking out
then. And so you hope that
>> How likely do you think that is though?
>> It's probably like 20% 30% chance. The
reason I think there's some possibility
is that because of the export controls
um you are so the deepseek is not
building with the Nvidia stack. they're
building with the Huawei stack
and because there are export controls on
not just the Nvidia GPUs but also on
HPMs
these um architectures that that DeepS
building are far more like memory
efficient they made innovations on the
KV cache to be really small enough that
you can host it on the SSDs
and you don't need high bandwidth memory
for inference time and they're going to
have a completely different architecture
for inference, completely different

architecture for storage cuz they're not allowed to use the 3D nans. So their architecture is going to look it's not just the model architecture. The model architecture is already pretty different. They've made innovations on the attention layer. They made innovations on like the the training algorithm so that it doesn't consume a lot of interconnect capacity. So they they've made a lot of they basically their whole stack is getting vertically integrated to their hardware and their chips and their fabs and so on and so that's a very different bet from what America's making.
>> Do you think the export controls have helped or hurt us? jury still a lot shortterm it's helping because the only reason I my belief the only reason where why there's even like a 12 month gap between open source and frontier is export controls it's definitely helped and and and definitely like companies like anthropic lobbyed very hard for it but um there is a chance that because of that they now get really good at the physical layer and One advantage they have is they can actually build data centers
a lot a lot a lot faster. Power is not a problem. Permits are not a problem. People are not a problem. Labor is not a problem. Expertise is not a problem. And so by forcing them to go out there and build all this, you're converting them into a far more
like potent competitor.
>> Do you think we still dramatically underestimate China's capabilities?
>> I think so. Because if AI is like not just digital, that's also physical AI.

You got to build fabs, robots, chips and harness the energy really well and um package it into local devices.
I think they have a lot more advantages than America.
>> How important is it that we have our own TSMC in the US? So TSMC is actually there is a fab of TSMC in Arizona. Like not a lot of people talk about this but TSMC is investing like $150 billion into that into into building American fabs and um they've already invested $40 billion or something like that. 60 billion last time I checked. So there is a TSMC in Arizona that's coming up. There's also um Intel and that's why you know American government owns 10% of Intel. U Nvidia and SoftPank own 5% each. So there is a lot of investment going into an American fab as well as TSMC is investing into its American fabs. Elon's building terra fab like I think it people have woken up to the importance of building fabs but um this is also why China is particularly very very competent because given the capabilities of China that we just mentioned there really articulately I know it's a ridiculous question but sort it um if I were to say to you your job is to make sure America stays competitive
what would you do to ensure that you retained competitiveness in an increasingly strong China.
>> I think I think take physical infrastructure a lot more seriously and continue funding it um and not like have all these
you know I I wouldn't say meaningless. It's more like not propagate fake news around data centers

um about how data centers are polluting and contaminating water or like they're sucking up all the water um and and and actually be fact driven and so you know I hope our product helps there like you you can you can go to perplexity and ask any question and get fact checked on your assumptions but yeah like it's very important that we educate the public um about what's actually going on in in a language they easily understand and not fear-monger, okay? Like not be like, oh, like all their jobs are going to go away like this, that like there's going to be lots of amazing companies that are going to get built with far fewer people getting multi-billion dollar, multiund million valuations with like 20, 30 people and propelling like trillions of dollars of new GDP. Like let's talk about how to enable that. let's talk about how to build that and create a more positive future together, right? Uh instead of, oh, like 90% of the jobs are going to be gone. Like you're all going to get screwed over by our models and like and and it's it's our it's our moral duty to tell you all this like blah blah blah. Like that doesn't make any sense to me. Like you can't win by saying that and also like complaining about not being able to build data centers fast. Do you think we've done a complete disservice by having the marketing message that Dario has had that all jobs are going and it's all doom and gloom?
>> Yeah,
I think so. I mean I think you know they have contradictory messages in their own like
uh different social

engagement so far where the most recent one I heard was there is no evidence that AI is taking over jobs.
But so I I I think there there needs to be a consistent communication around this. And
I also think that um very little is being spoken about how AIs can help you build companies in a very very different way like the current AIS AI.
It it's already true that so many things you would hire people for you can do it with agents. But one way of looking at it is like oh like what happens to all the jobs. But the other way of looking at it is like, hey, like I can I never had the chance to go build out a company on this idea that I've been having all this all this while and maybe me and a group of friends can come together and build this and can you guys figure out a way to give us compute credits or you know Amazon gave a lot of compute credits to a lot of startups like when we started Perplexity
we had like around $200,000 worth of Amazon credits and GCP credits and Azure credits.
um that almost like together cumulatively this was worth like a million dollars in computer credits. Now in today's world it's going to be like a million dollars of computer credits. And we're doing that like we we're funding this thing called a billion dollar build where we're giving a million dollars of computer credits to any group of people who have a credible path to building a billion dollar company and I want like thousand such companies to be built.
>> What did you think of Sam Alman giving $2 million of tokens to YC companies

initially?
>> I think we should do more of that. Yeah, that that's the right thing to do. Like we should do a lot more of this because you want new companies to be built. Um and and and even if they're worth multi00 million, right, it's good. If there are thousands of them, like that's a lot of new GDP. I I spoke to an Betski before the show and she said how AI pilled the team is for you. How big is the team today?
>> It's like 400 people.
>> 400 people. How big will it be in two years time?
>> I don't know. It's hard to say. Maybe 800 or,000.
>> So, will companies follow the same headcount trajectory that they have always followed and we will just solve new problems. Or will they be dramatically more efficient with a much fewer number of people?
>> Definitely, they'll be dramatically more efficient. Right. and and that's why I I I am a believer in building a lot more efficient companies
not and being an example for all these companies ourselves like like people should look at perplexity and be like oh like with 400 people um you can build like a multi like like I don't know like 20 billion $20 billion company um and so that means with like 40 people I could probably build a billion dollar or $2 billion
you know, and that that that's totally doable. Totally doable. And um and and so for us, maybe that means is with 4,000 people, we could be worth 200 billion. We we could be worth $2 trillion with like 10,000 people.

You know, I think I think that doesn't mean it's bad for all the um hundred,000 people we did not hire for a typical $2 trillion company.
I would rather have those 100,000 people be split into groups of like hundred thousand groups like that and each of those thousand groups are worth a few billion dollars. That's awesome. And I think a lot more people need to be entrepreneurial.
Um there are people who would be bad employees in any company because they're just like difficult to work with. They they don't listen to like instructions. So like they don't follow like road maps or not they're not like easy to collaborate with. But maybe the the flip side of that is those those are the kind of qualities that founders typically have.
>> Ain there is a population and a very large population that are not AI native people that are not using AI to improve workflows, improve efficiency.
What would you advise them?
>> Get started. First steps get started and and and channelize your curiosity. Right. Um, you don't need to use AIS to do your existing work. If that's if your existing work is boring to you, you probably won't enjoy it even if you use AIs to do it.
>> You got a lot of heat for saying people don't like their job. So,
>> I I didn't say if you actually listen to my interview, I did not say that. So, people want clickbait articles and they take something I said in one sentence and out of context and make it into a headline. What did you say? I I specifically said this. Hey, like there

are a lot of people who don't enjoy their jobs. Does any like By the way, the fact that that that that thing went viral is not because I was completely wrong. I think a lot of people resonated with the fact that I was actually honest in saying a lot of people don't enjoy their jobs. And that has nothing to do with your economic position or standing in society. You might even be like really wealthy but doing a job that you completely don't enjoy and like destroying like the peak years of your adult life working on something that is horrible or like like depressing. So like my point is that if that's you and if the reason you could never leave your job is because you were always worried if you how would you build a company from scratch? Like there are all these things to figure out how you have to hire a lot of people. Oh, you have to like set up an office this that. Like that's changed. For the first time in history, you can get started on an idea with like one or two other friends and and and and maybe have a real genuine shot at building a billion dollar company. Totally get that. Everything that we've discussed today has been on the back of unprecedented demand up and to the right. We need more memory. We need more data center supply. We need on demand and server side. Everything's like up and to the right. Seeing some cracks in and Uber's saying, "I'm not sure I'm getting the productivity gains that I thought." Microsoft aligning with them, putting a $1,500 token budget. Do you think we will have a continuous up and to the right acceptance that productivity gains are unwavering? We

have to do this, or will there be falterings along the way? I mean, I'm sure there's going to be falterings along the way and people are rightfully freaking out about token maxing, which is why I think you need some form of hybrid agenic inference. You need some amount of inference compute to run locally that you're not paying for tokens on um unmetered intelligence essentially.

>> How will the best companies of the future structure token budgets? My hope is that they don't have to understand that
they will be able to work with an orchestrator who does it for them. It's not going to be easy for you to constantly keep track of like which models are the best at what things and how do you allocate oh this is the budget for coding, this is the budget for finance or like like how do you even understand like which models are good at each of those things and like how much do you spend on each of these divisions? You're not going to be able to keep track.

>> I I had a friend on the show the other day say that Google will be the token king. They can produce the lowest cost tokens out of anyone. They own full stack TPUs, data centers, networking, power, procurement.
Do you think that's true that they will be the lowest cost token producer?

>> They have advantages all all advantages one needs to have to be that. But they underestimated the importance of coding models and so they far behind the frontier right now. So again, they could catch up, totally capable, totally

competent team, but today they're not quite at the frontier.
>> I was shocked the other day. I saw the Cloudflare announcement that um now agent traffic has overtaken human traffic for them.
>> Why? Why are you shocked?
>> It was quicker than I thought.
>> Okay.
>> Personally, I thought that would happen, but in two years, maybe not now.
How does the world change when agent traffic far exceeds human traffic?
>> I think people are just going to have a lot more agency.
That's it.
>> Do websites go away? Does design not matter? Does the advertising model of the internet die completely?
>> No, it doesn't. Because my belief is that
the advertising model around like travel or shopping or like u fashion are not getting disrupted by agents because the judgment is not objective.
Any anything where the judgment is objective, the transaction is based on objective judgment that's going to get disrupted by agents.
anything where the transaction is more subjective like the decisions are more subjective like like what is the best piece of furniture inside this this this podcast like why this particular table or like those kind of things
>> probably for the mic you would buy an objective decision the table
>> you probably are caring about the aesthetics of the room I think I I think that's kind of how I I feel the world will split and subjective things will still be ad based objective things will

be agent based
I watched your commencement speech on the back of speaking to Samir at Excel and he said I had to watch it. So obviously I watched it. Um and one of the points you made was the defining skill of the area is asking better questions.
>> Yeah.
>> What question is no one asking today that maybe everyone should be asking?
>> I think people need to ask more about like okay assuming I have a lot of agency available to me what do I do? Imagine like I gave you a headcount of like 100,000 people or 10,000 people and and and
you know enough computer credits to run those agents. What would you do? Like let's say I ask you Harry like you know let's say I you have suddenly like 10,000 agents at your disposal. What would you do? Like I I remember you telling me or not me but but in some episode of yours where
>> you said you're only you only did this podcasting because you felt like you didn't have an arbitrage to go win deals
>> 100%. Yeah. It's why I still do it. I mean I love what I do but yeah.
>> Okay. So you've gotten some amount of distribution. So now assuming that you can you you have let's say you could spend $100 million on a genenic inference and grounded with all the connectors and stuff and it's all working. What would you do to to with with that capability to um further your goals? Like what what what should your goals even be then? I I think that's the question I would ask assuming that in the next three to five years you're

going to be able to like delegate whatever digital task you want and and with the right harness and agents and like be able to delegate that.
>> Fundamentally it would be to build aic infrastructure to be able to find, identify, outreach, set up, win great investments and have the media sit on top and power that. that is intensely difficult to do and would be the holy grail to investing but like
>> that would power what my end goal ambition is.
>> Yeah. So your goal is to be the you know run like a 10 to 100x larger fund right that that's basically what I'm hearing from you.
>> So that's like let's assume it's like a $40 billion fund from 400 million. Um then all you got to ask is like assuming I have all the headcount I need to do this like how much faster can I do it? I I think that's how I would frame this question. I think Elon has like a similar thing he spoke about once where okay assume that a task somebody tells you a task is going to take 10 years. Um ask the question what would it take to do it in 10 months?
Maybe it's impossible to do it in 10 months, but you'll probably get pretty far asking those questions compared to somebody who takes it for granted that it's it's going to take 10 years.
>> All right, interviewer, we put it on you. What's your 10year and how does that look in a 10-month time frame?
>> I think our our mission beyond any level of capitalism is to make the planet more curious.
You know the product is always intended

to helping people ask the next question
and uh my goal is to truly realize that
like that level of agency
that needs to exist in this world is
quite not there.
>> I think I think that needs to be
grounded in numbers dude to make it like
possible. It's like me saying oh I want
the best investments
like which is why a $40 billion fund is
helpful.
>> Sure. I can say the same thing like 2
trillion you know it doesn't matter
right like 100x 10x 1000x these are all
like um motivational milestones
>> do you think will be a trillion dollar
company
>> yeah anyone can be a trillion dollar
company SKH and Samsung are worth a
trillion last last couple of weeks did
you know Samsung started off as a
grocery store did you know that you
didn't know that okay so it's true they
started selling dried
Seriously. Um, Heinix was um SK the SK
group started off as um um textiles
company. So, anyone can be worth a
trillion dollar company and like you
just have to work your way towards that.
I mean, the exact same logic for you
that you laid out for how can a company
be worth um hundred billion. Okay, you
said you need to make a $10 billion in
revenue. Isn't that the same for
trillion? Like you need to make a
hundred billion in revenue.
>> Mhm.
>> And there was actually some very
interesting data that CO2 revealed. I
don't know if you saw it recently, which
basically says about the probability of
reaching the next level of value is much

higher. So when you're at a billion, it's like much more likely to reach 10 billion. 10 billion much more likely.
>> Yeah, that's true actually for even people. Like it's way more likely for a person with $100 million in in liquid net worth to become a billionaire than someone with $10 million.
>> Are you not worried about the wealth inequality? Aaron, if we being blunt, we both are very lucky now to live in kind of nice worlds and rarified airs. Are you not worried by just how much money a very small number of people have and how [__] hard it is for everyone else and that gap is getting bigger?
I think the way to like ensure that that's not that doesn't remain the case is to distribute the benefits more widely. You got you got to let anyone by the way the people who are using our tools like I had an Uber driver I'm not even like making this thing up so um as honest as it can get. an Uber driver in San Francisco um once told me that he watched one of my uh YouTube interviews where I explain how you can build a product or a web app with an AI from scratch, went on to do it and um um used AIS to add like billing and all that and that makes more passive income for him than um driving Ubers and so he actually reduced the amount of time he's driving Uber because He he loves wipe coding new apps and u that that already tells you that
for the person with agency and a positive outlook for the future, anything is possible. And so if you keep communicating
all the negative things you can about AI and wealth inequality all the time and

that's the only thing news uh and press writes about,
I think it'll perpetuate and people will only think the bad things. And so it's it's it's very essential that if you think you're already doing well, it's very essential that you talk about what are all the things that can go well and give hopes to people who were once upon a time like you like you you were you didn't you you started this um podcasting circuit like when you had nothing, right? So
>> nothing.
>> Exactly. So it's possible. So you got you got to talk more about that than be like oh I feel so guilty that I made it and now I'm like you know what about all these people who haven't made it like you can also make it.
>> I I think I have a more pessimistic view of actual general public which is I don't think that many people have agency. I think a lot of people have victim mentality.
>> You got you got to help them. Like I think that's that's the most important thing.
>> I think they got to help themselves.
>> Sure. But people will help themselves once they see that okay like I kind of want to be like this guy. let me let me work hard. You need an example, right? Um it's not like nobody can become um get in shape. Like it it takes discipline.
>> Takes discipline. You got to get rid of bad habits. And so
>> and now is the best time ever to change your life in 12 months. Like the ability to go from nothing to to actually billionaire in 12 months is now possible

in some respect.
>> Yes. And so I look, I'm not saying everyone's going to make it and everyone's going to be worth a billion dollars.
>> Isn't that the caption from this this show, Arvin? Everyone's going to make it.
>> Anyone has the potential to make it.
>> So it's it's it's as likely for Perplexity to become worth $2 trillion as as a founder who's yet to secure your funding to be worth a billion dollars. So it's it's equally hard. Equally hard. And um and I think you just have to give yourself, you know, shots at the goal and um be curious. That's that's the message from the commencement speech. Be curious.
>> We have SpaceX. We have Anthropic. We have Open AI going public. It feels like someone's kind of shot the gun and the race is on. Is there enough money to fund three such large IP?
>> There will be some reallocation for sure.
like um there there might be some holders of like SAS stocks who would put it into anthropic or something. Let's say you believe that enterprise AI is going to take off. You might want to hedge between having a lot of Microsoft stock and Salesforce stock versus like putting some of that into anthropic. So let's say like Vanguard or Black Rockck own like you know cumulatively they own like $200 billion of Microsoft and Salesforce. They might be like, "Okay, I'm going to take 30 40 billion of that and put it into anthropic."
Fine. You know, not a bad bit to make.
>> What happens to all the enterprise SAS

companies that are public growing? Yeah. Fine.
>> They have to they have to weather the storm.
>> Is it a storm or is it a continuous precipitation?
>> I think you have to bring down the costs and produce new value.
Salesforce has done well because they always went and bought the next thing.
If you're just selling the same software, you're probably not going to be around. IBM is still around because they went and bought Red Hat and Hashikarp and now they're buying Confluent. So, there are ways for these companies to stay alive and extend their lifespans and stuff. It's obviously going to be hard to preserve a brand that's as relevant
like like I don't think the IBM brand is that relevant anymore in terms of like evoking an emotion and people to go use their products but as a business it's going to be awesome you know it's going to be fine.
>> I have to finish on you said IPO in 2028.
I had to ask this. I woke up to this in my like you know um group. We have a team WhatsApp and it's like ask IPO 2028. I hope I hope it can be sooner than that.
>> When do you know when you're ready? Like is there like a billionaire? Are you at 500 million er now?
>> More than that. Far far more than that actually.
>> Really?
>> We we're not ready to share it, but um growing really fast.
>> Revenue growth matters much more to you

than profitability
>> today. I think in general, by the way, you can look at public markets. Um people want topline growth.
more than um bottom line efficiency right now because it's very hard. It's it's rare.
>> Well, you definitely need one.
>> Of course, sustainable businesses,
>> you need to have a model in place to get the bottom line efficiency when that becomes the objective and and you need to also like have a path to getting there.
>> Where are you cost inefficient today where you expect to be significantly better in 2 to 3 years? We're training our own models, post- trainining it on top of amazing open source models, and that will bring down the cost that we currently spend on Frontier model tokens.
We expect to continue to use Frontier models for designing new experiences and new capabilities that do not exist today in our products. But whatever exists today in our products right now, we expected to completely re rely on like models we own and serve ourselves. And that's
going to be the best way to bring down the costs and increase our margins.
>> Will the largest enterprises in the world all be fine-tuning open models to have tailored models that are much more specific to them?
>> Absolutely. Because it's in your incentives to bring down the cost.
>> Does that not provide another bad case for the large frontier model providers?
>> Frontier model providers will only remain relevant if they remain at the

frontier. If for 6 months you're not seeing a new capability, it's bad for them.
And so that's the uncomfortable nature of this field. You no one's ever in a comfortable position. Like I said in the start, no one's no one can relax. This is
>> [__] a hard business. It's got harder.
>> It's going to get even harder. And and um that's the nature. This is the the price is too big. Like you've never seen like like take entropic. I think it's um worth like one to one and a half trillion some something in that range. That's basically the valuation of meta and and and this all was created in like 6 years.
Meta took like 20 years to build. So the price is so big and so no one can um no one can be comfortable
and and and and and anyone who's winning today can lose tomorrow including including the mod providers.
Previous this year, there was like a three-month period where people were like, "Oh, Py, what's happening with Pacity?" Do you pay attention? Do you give a [__]
>> Of course, I pay attention to all that.
>> Do you care? There was one in particular in San Francisco. Do you remember where they were like, "Oh, what's the company you had short?"
>> Yeah, we were voted the most likely to fail. Cursor was voted the second most likely to fail. Open AAI was voted the third or something. Um,
>> you didn't give a [__]
>> I I feel like we're all doing well. Curser, I think, is getting sold. SpaceX

Open AI
>> is
>> going public soon.
>> We tripled our revenue since that
>> since that uh judgment was made. So
brought down the burn by more than 50%.
So I I don't know like my my my sense is
that um I also feel most of those people
who sit on these like meetups and vote
don't actually build anything useful.
Yeah.
>> Uh, okay. We're going to do a quick fire
around because, uh, I could talk to you
all day. Um, first one, what's one
widely held belief that you think is
completely wrong?
>> I think a lot of people are obsessed
about like, you know, identifying a mode
in the first year or two of their
company. But, um, I think like the only
shot you have is move fast. like veloc
in my mind like moving fast is a way of
expressing humility because you you're
constantly making contact with the world
and trying to question your assumptions
all the time.
>> Where are you still moving too slow
internally today?
>> I think we can be even more AI.
It's insane I'm saying this because we
are building some of the most
interesting AI products and internal
adoption of our own products, our
competitors products can be even higher
and and um this is despite us being
extremely
um agent build internally and trying to
delegate as much to agents. Yeah, that's
where that that's like a big area for my
my hope is that we can turn this company
almost into an AGI and and um have that
doesn't mean no humans work here, but

there will be an AGI that has all the
context it needs to run different
divisions of the company
in a semi-autonomous way with some
scaffolding provided by humans here and
there. And and that's not that's going
to feel that's not going to feel scary
at all. We'll normalize that that
feeling very fast. It's just going to
feel like 10, you know, uh the 10x
engineers running certain aspects of the
company.
>> If I gave you unlimited money, what
would you do today that you're not
doing?
>> I would build data centers.
>> You would?
>> Yeah.
>> In space?
>> I don't have expertise to do that, but I
would I would start with land on Earth.
You know, I think there's a lot of land
and and maybe you can be resourceful in
securing permits and power in different
countries, but I would start there. I
think it's, you know, I like I said,
like I think physical infrastructure
buildouts is like the return of the
industrial age again
like like the the the forefathers who
built the industrial revolution,
um oil pipelines, steel bridges,
factories producing cars, all these
things that we take for
granted today were built by people who,
you know, spend a lot of time thinking
about how to scale these things in a
costefficient way and so um we need to
do that a lot for AI and um yeah that's
what I would do of course you you you
cannot just be building infra you need
to be able to utilize all that infra to

producing valuable output tokens to the user
but um we're already good at doing that so infra is the thing I would focus on
you can buy and hold for 10 years SpaceX anthropic or open AI the three IPOs coming in the next few months which you buy and hold for 10 years and why
>> SpaceX
why it's an end of one company like Enthropic and OpenAI can claim they do whatever each other does but u um SpaceX is the only company building space infrastructure for connectivity.
Have you been on a flight with Starlink?
No,
>> you should.
You will hate being on a flight without Starlink after that. Imagine we can record this. I can watch this podcast while flying on a plane.
Starlink lets you do that. That's just one aspect of the business. That's just one aspect of
>> one small aspect of the business.
>> Yeah. Like there's a lot of like I'm excited about possibilities to travel from Australia to like uh San Francisco in like 30 minutes. You know, all this feels like sci-fi, but I'm excited about all these possibilities.
>> What job does not exist today that will be incredibly common in 5 years time?
>> I think it already exists. So, if the forward deployed engineer is definitely on the rise, I guess like people with a really good sense of like um quality control.
Maybe a better uh way to answer this is like most jobs that exist like valuable jobs that exist are usually like reincarnations of something that already

existed.
Like so I don't think we're going to see completely new things. It's just going to reincarnate in different ways.
>> You can advise your little sibling who's finishing university today and just done a computer science degree. One thing, what would you advise them?
>> Stay curious.
Don't don't don't give into like FOMO and trying to max out on something here in the short term. Like don't go to Twitter and feel like a loser that people on Frontier Labs are getting so rich and like you everything feels hopeless to you or something like there is so much more to build like we are just getting started like the the the application layer era or like infrastructure buildouts there there's like a lot of opportunities we are seeing more spinouts from open AI anthropic you name it every single day do we have hundreds of these neolabs in vertical models.
>> No, not not a big believer in too many of them. I think you got to produce some differentiation. That's the most important thing. Um like like I I if would you call Deep Seek a Neolab?
>> No.
>> Why?
>> I think very stupidly for me I don't call it a Neolab because I I attribute Neolabs like spinouts from larger labs.
>> I see. and and kind of verticalized which is probably wrong on both par like axes
>> but it's horizontal and it's not a spin out.
>> Yeah. I mean I kind of like the idea of labs taking a differentiated bet. Okay.

If somebody really questions the transformer architecture itself or somebody really questions needing to build on Nvidia GPUs or something like that like like foundational bets makes or or somebody questions or somebody goes out and builds for robotics models. I think I think that's like somewhat uncorrelated and different and that makes sense for a lab, but I feel like they're like just labs for the sake of being lab and I don't think they're going to make it.

>> Can you paint for me? I I do like this one. What's the most plausible story whereasty becomes a trillion dollar company? What do you do then? The orchestration layer.

>> I mean, accuracy and orchestration is is is like two goals that have been consistently true since the beginning of our company. So I think we'll continue to do that. We'll be orchestrating across devices, chips, models, tools, files, connectors, everything. Right. So what would I do once that happens? I don't know. We we'll chart our path to 10 trillion.

>> Are you happy now? But are you are you enjoying this?

>> Of course. I mean, I wouldn't like there's so many things I could be doing if not for this. And um I think the process is is is what motivates you. So you asked me I think somewhere in between you need to give me a number of where you want. I don't I don't work like that actually. I I like for example like the these numbers like getting to two trillion or 20 trillion are are exciting but like that that doesn't motivate me. It's it's hard to

get motivated by wealth. You you you want to get motivated by impact. Who's the smartest person you've met? Final one. You've met Jensen Hang. You've met the best of the best. I've been fortunate enough to Who's the smartest?
>> People are smart in their own ways. It's hard to compare. Like I met Jensen, Elon, all these guys and like Bezos.
>> What was it like meeting Elon?
>> Amazing. I mean Elon's like a very uh focused person like like that. He might not appear that way on Twitter but you know with a lot of like random tweets but he's extremely laser sharp focused on whatever he's doing at that moment in time. Actually the the one skill that as an entrepreneur that I would really like to take like build from somebody like him like take from somebody like him and have it for myself is that ability to just zone out of all the other things that's happening in your business or other businesses and just focus on that limiting problem right now like the bottleneck problem and and and and ignore everything else. It's very hard to do like even within perplexity I cannot just focus on like one part of the business alone. It's very difficult like I'm I'm always looking at other things simultaneously. And um his style is to just always look at the limiting problem and just ignore everything else. And uh that's very hard to do because you you actually have to be really good at concentration.
You you have to be really good at ignoring
even important things which are distractions to your core objective right now.

>> Was Jensen hang what you thought he'd be?
>> Far better.
>> Really?
>> Yeah. Jensen is so truth seeeking. It's insane.
I think he or somebody else told me or read in a book that he
um is so intense that he wakes up every day and tells himself that he sucks
and goes in like like he's so intense
that he tells everybody around him that they're 30 days away from going out of business. Think about it, right? $5 trillion
um guaranteed to make $500 billion in revenue in the next two years. Um
and um has the most advanced chips in the world and and and he operates with that mentality that he could be 30 days away from going out of business. That is what it takes to be Jensen Huang. And uh there's so much to learn from these guys. There's so much to learn. I think uh there's one aspect of like you know being comfortable where you are thinking you made it uh you know that that feels good to get here so far but these guys are not stopping like I I don't think Elon wants to stop at uh if you look at his pay package for SpaceX it's structured around um creating a colony in Mars with a million inhabitants and uh building enough comput in space so that's why it's not like motivating to be
worth a 10 trillion in net worth or something. you know, if if he does these things, I'm sure he's going to get there, but um it's more motivated around like making the impossible things happen and and and having like that long-term

outlook like you um I think that has
been the biggest thing to learn from
maybe these two individuals in
particular is um
a lot of people view this like
entrepreneurship as like oh if it wins
if if I win and I have a great outcome
and I sell my company. I would have like
generational money. I don't have to work
ever again. And then what you end up
like like just staying at home and like
your your your kids will obviously have
like trust funds and they're not going
to get inspired watching their dad play
paddle.
>> Yeah. You know, you're not going to set
the right example for them. and they're
not going to be able to take your wealth
and multiply it cuz they they didn't
watch somebody who actually did that.
You they you did it before
they they they were like adults. And so
um I think you always need to be doing
something. Um like like Jensen said some
recently that he hopes to die on the job
or something like that. Like that's the
attitude you need to have. Like you got
you you need to work forever. I was so
upset though when Jensen said, "If I'd
known how hard it was going to be, I
wouldn't have done it." When he did, I
don't know if you saw that into I was
like, "Oh,
>> yeah. I think it's pretty hard, but you
you don't do it because you do it
despite that." I think I think that's
how it works.
>> Arvin, listen, this has been so
fantastic to do. I so appreciate you
taking the time while you're in London.
So, thank you so much for joining me.
>> Appreciate it.