Inside India’s Sovereign AI Ambition | ft. Pratyush Kumar and Vivek Raghavan of Sarvam AI

55 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=NHUFWLPRPV0&T=996S
https://www.youtube.com/watch?v=NHufWLprpv0&t=996s

SUMMARY

Saram AI's founders, Vive and Pratush, discuss their journey in building AI technology in India, emphasizing the need for the country to develop its own AI capabilities rather than relying on foreign models. They highlight the importance of creating foundational models and infrastructure to empower various sectors in India, aiming to democratize access to AI and ensure sovereignty in technology.

- *India must develop its own AI capabilities to avoid becoming a "digital colony."*
- *Saram AI focuses on building foundational models and infrastructure for AI applications, particularly in Indian languages.*
- *The company aims to train both small and large models, with plans for a trillion-parameter model.*
- *They emphasize the importance of having a full-stack approach, integrating model training with application development.*
- *The founders believe that AI can significantly improve lives and should be accessible to everyone.*
- *Saram AI has already contributed to open-source models and aims to create a robust ecosystem for AI development in India.*
- *The company is positioned to leverage the growing demand for AI, particularly in sectors like education and healthcare.*
- *They acknowledge the need for significant capital and infrastructure to support their ambitious goals.*

Sometimes with pride we say we have the largest number of chat GPT users but you know what are we doing really we are exporting our data and like uh importing intelligence can we actually afford [music] to do that >> he's saying India should build a 1 kilometer bottle >> this technology is so fundamental and it impacts so many different things that uh this is [music] not an optionality for us to have >> given that uh India has not participated in value creating platforms we built services we built apps [music] we have built use Sage but you've not built platforms that acrew value in a major way. Our bet is that the model layer does that and and you can't restrict yourself to the model layer that's why you do full stack. What are the things that we have done so far that allows us to say serbum can do this? to the best of my knowledge outside North [music] America and China and if you leave a scite mist nobody has pre-trained from scratch a 100 billion plus parameter model >> from here to a trillion parameters what's the capital do you think it'll [music] require in the billions or tens of billions or hundreds of millions and what's the number of GPUs you require and how much do we have right >> [music] >> Welcome to another episode of lunch break with light speeded. Today we have a very very special guests both founders of Saram AI. Pratush and Vive are in the studio with us. It's been very hard to get them out of

their uh product technology and uh company building cycles and then you know get them into a room with us to talk about what sermon does. Thank you both for your time and absolutely delighted and talking to you guys today.

>> Yeah, thank you for having us. >> Thanks. >> Amazing. Uh you guys are like now in a place where everybody in the country knows about what serum does but they all hear about it from WhatsApp. Like my mom has sent me a bunch of these forwards from Instagram about is this the company that you would talk about like yes it is the company uh is this what they do like that's not all what they do they do a lot more so today we'll try to figure out and you know explain together what does stand for why does a company like this need to exist in India and for the world so we'll get started maybe we'll start with you Ve tell us about what were you doing before Sarbam what was your life's mission what gravitated you towards starting a new company and how did you meet Pratush and how did you guys come together?

>> I think I've had an interesting life let's [laughter] have >> but uh you know I I you know I started off the regular way I uh you know got uh went went to IIT and then uh went to you know the US to call and got a PhD spent about uh 20 years you know uh multiple startups but actually in something very very different. we were actually building software to design you know semiconductor chips and that's that's really what uh I was in that industry for 20 years and uh then you know suddenly in 2007 really almost on a whim I decided to come back uh to India and and I think that uh um after that you know I you know I was I was still in the same industry for a couple of years and uh you know I said you know I need to do something different and uh and I had no idea what that different would be. And what it ended up being is really I spent 15 years of my life and and actually uh you know building or being a small part really of what we call the digital public infrastructure of India right and and as a full-time volunteer and and frankly uh you know and uh I worked on Aadhaar for a very long time. I've been involved with the GST. I've been adviser to UPI. I've been adviser to the the courts and I actually about five six years ago I got interested in Indian language AI and and actually that's where for the first time Pratush and I kind of uh got together to try and build things in the open source for Indian languages. But frankly my view really and I and the way I saw and I I really didn't think that uh I was going to do another uh startup you know I I I thought that I'll I'll be working in kind of you know in the open source or working with uh government helping government with various things and I really thought that was really what I was going to do for the rest of my life but you know uh you never know what happens in life. I think I you know uh you know was actually stunned at you know what how this really can change right how how this really can change the world and how it's and and you know I had the experience of building things at scale for India and things like that and I said that you know this is so fundamental this technology that we need to do uh something about it but nevertheless the you know my first instinct was you know how do I get you know philanthropic money or how do I you know help asked the government to kind of you know build something there but then after a while I realized uh this is not something that can be achieved uh in that sense and and I think that's that's when you know um I asked that you know we need to do this you know we we need to do this >> so so a full-time >> volunteer working with a full-time professor I mean that combination is I think one of the most unique founder combination that we've ever seen >> starting from philanthropy a word that venture capitalist tend to not really like so much and going to deep venture capital. So maybe yeah, tell us more about like what >> what made you feel like you know what let me leave this nice >> professor job at I Midas I think at the time and then uh let's go drop all of that then jump in the middle of this whirlwind. Uh V is right. He was the uh instigate >> was I was

actually doing a dual role both at Madras and Microsoft research.

>> Oh yeah. Okay. >> Also seeing the hey the first wave of uh identity crisis of researchers. >> What do you do in this post AI word and I was also recognizing that it's a very big step change. So we have to be allin, right? And it was not clear what all-in meant but then Vive was like hey okay I know many people have caught up with various things but why don't you why don't we just do this right because we knew we need to do something allin and we knew that you need a structure we had we had tried other things in the past uh and so this just felt right and I think frankly speaking uh the the only kernel of clarity was that we should build this ourselves right I think that thing right of building things ourselves doesn't matter what the market share is doesn't matter what you know when is it going to happen do we even know how to do it no answer is no but it is clear that this is a step change and it is clear that we'll have to be able to build this ourselves right now the word sovereign AI etc was not there uh the India TAM of AI etc nobody knew right but uh the I think that is what I guess is the builder's instinct you say that this is a hard enough problem and valuable enough problem that we should have a company to do it and there were uh luckily good investors like yourself who said that we should consider that >> take us to the first day in a room, right? I mean, you just raised money.

Now what? Now, now it's all real. Now you got to, you know, you got to leave your volunteering role, which I hear you never took salary for, which I really admire. And then you you were doing part-time Microsoft and full-time role at IT Madras. Now, now what was the first month look like? It must have been like complete chaos. Something that I think a lot of the founders really relate to is you know how does it start for companies that eventually become successful like you guys still early but still it's an important step for these guys. So tell us what that was like for you.

>> It's it's interesting that you know uh you know even though today we are mostly based in Bangalore the start of the company is actually in Chennai and in terms of you know where people were. So we actually started with a boot camp I think uh I mean uh we we we actually had an office in the IT Madras research park and you know because we were a company where where you know we needed to start with you know normally a startup starts with you know two three four people we started with close to I don't know 15 people at the very beginning >> and free rent I'm [laughter] guessing frugal engineering >> almost free almost free rent >> we shall not speak about that anymore [laughter] okay >> but but but I think that and you know we just said that you know I think the most important thing was to get everyone in one place and actually just kind of uh get get together and do things as as hey this is how we're going to start about where we want to go. Did you guys know like you know what this is what this is a direction like this is day one we got to build something big in AI but models hardware agents not even like clear idea what would actually work at that point in time this is 2023 mid to late how what gave you the direction this is where we should start paddling this is the direction we should start swimming >> how did that come about was it a journey for you >> as Vick said I think it's about the team first you need to assemble a team that can carry the ambition no matter what direction you're going right and uh that's where we were lucky the work that we did have done in the past lives created let's say good karma that people wanted to come and join us right and frankly a startup no matter how much money is a big bet right for people to leave whatever they are doing and they they came uh I think it was very clear that we will train models that part is very clear agents didn't exist as a notion I think people are still figuring out application layer stuff but again we made one more contrarian but we were like uh we said we'll do full stack we application b we let's train models

also. Let's figure out how to run these things on GPU. So right from the early days I remember there was about I think we did maybe a meeting in my living room and there were like a room of 30 people or whatever from different places. We're just chatting about it. Uh we right from the first day it looked like hey we want to do quite a few things right that seems to be a little bit the DNA of the company. Uh we want to train models we want to play around with applications see what we can build >> and also there was energy in the team to look at GPUs and how things run on them.

how do you do inferencing and so on right so in some sense if I were to combine these two points one was the builder's instinct we need to build it ourselves right uh and not everybody wants to think like that right the people want to play at the business model level and so on the second was we want to sort of like look at the entire stack so with that ambition we started of course it's not a straight line you go through various uh left and right turns along the journey and the area has moved like crazy it's continuing to move like crazy uh so there have been thing but I think this these two have kept us going in the same direction and the other thing of course is that we wanted to build for India and I think that was something that was you know uh almost the core uh you know tenate of where we where we came from and what we wanted to do and I think that was something that uh yeah and we always wanted to be full stack we wanted to not we're not theoretical in the sense that that you know hey we're going to build something and then let others figure out how it you know can be used and I see we wanted to be to a place where we actually uh >> uh you impact people right that's yeah >> so yeah I think that is now in the name of the company salam which I think you you guys should talk about what that where it came from and then maybe we'll just then start digging into what is the strategy for the company I mean build for India okay uh it's an approach it's an angle full stack is an angle it's a product angle it's a technological angle but why should this company exist what did you see was happening in the world in the supply chain of AI that made you feel like you know what we can build a large company supplying intelligence to a certain region starting with India with a certain approach starting with you know partial stack then over time full stack I think what you call TRL readiness right no as far as the name serum actually Pratush definitely deserves the credit for that I think uh I think the principles were quite simple we said that you know we wanted to have something which had an indic flavor to it and we wanted to have something which represented the inclusiveness right that that you know But why was that important? Like why was that important?

Was it a business decision or an emotional decision for you guys? >> So yeah. So I mean there were two parts to that right so of course in Sanskrit means all right. So there are two parts to what we are trying to say by doing that. One is this should reach everybody right and in we we believe that even more strongly today the world can easily go towards a K-shaped setup right where some people have a lot more and many people don't have anything can also easily go to a place where everybody is so much more empowered with int basic universal basic intelligence rather than income right >> uh and I think that was one side the other side and that was probably the more uh internal instinct it it seemed like only two companies three companies are building it in fact there's no anthropic at I mean not very well known at that time. So let's say open AI and Google are building it. It can't be that the most you talked about supply chain right the most important component of a post AI world the value creating component of the post AAI world has to be built by only two companies that cannot be right. So I think the fact that more people should be builders was also an instinct right in fact just dubtailing to your point about what this means for the company. We believe that um the model layer is the mo is the lynch pin of the post AI world. This is the model layer that looks forward to the application layer and looks downward

towards the infra layer. Right? So unless you are doing cool things at the model layer really believing and building and understanding and going deeper the other layers will continue to get marginalized. Right? That was a a view at that time. Today I think it is far more clear. Right? The model companies continue to sort of like move upwards and downwards. Right? uh and and given that uh India has not participated in value creating [snorts] platforms.

We've built services, we built apps, we have built usage but we've not built platforms that acrew value in a major way. >> Our bet is that the model layer does that and and you can't restrict yourself to the model layer. That's why you do full stack, right? Uh that is the core thesis and there have to be multiple companies across the world. It can't be many. It can't be like hundreds. It'll be things you'll maybe maybe a dozen companies around the world who would have the mantle to then build at this core layer which is expensive which is R&D heavy but then look forward and backwards towards applications and infrastructure. Right. This whole stack is what needs to be done.

>> Very clear. Love the term universal basic intelligence. This is >> heard heard it for the very first time. Thank you. So, so now let's talk about the tech then the the starting with the model layer going down and going above the model layer. What are you guys doing below at the model layer above the model layer? Can you just first give us the full you know full gamut of the products that you're thinking about? Firstly, the uh the models, let me just start with that because that's where it starts, right? I think the uh we uh as a company are positioned to both train small models and larger models, right? The small models include things like speech models, vision models, document models, uh and uh sort of like even models for things like embeddings and so on, translation. So these are models which are today powering a lot of the work we do in the space on usage of Indian languages and bringing that world to the AI world. Right. So that is lots of these small models which are going back to our previous work on Indian languages going back to our uh instinct in making these models more efficient closing the gap in long tail of Indian language nuances and so on right and I think that is something we have been perfecting and the team has uh depth has has experience and we have been building that that is one class of models the other class of models is these large language models which are capable of reasoning math coding agentic tasks etc which are a completely different beast in terms of what it takes to train them. Many more GPUs, extremely larger data sets required um completely new area algorithmic ideas and so on. But because we are doing both of them, we are also cross- fertilizing ideas. You can take reinforcement learning from the frontier model effort which can then reach the other places like for example vision we have a top class model sort of like borrow ideas from different places. So with models we are interested in both of these right and we can talk a bit about our road map and so on. uh but training these models requires you to be able to manage GPUs at massive scale. For example, the uh frontier class model required us to work with 4,000 GPUs, ensure that we can run them reliably, scale the in the workloads on that and so on. Uh so to the inference of the models at scale, we'll talk a little bit about what we are doing at scale. Uh that requires us to use GPUs very efficiently, manage um when to scale up, scale down things and so on. So we do a lot of work at the GPU layer and we're going to little bit double down a little bit more on that layer going ahead because uh we believe that highquality infrastructure is the bedrock of sovereign AI right and uh we believe that there are some gaps there in the market where the layer of the infrastructure as a service and the management of uh compute uh effectively is a missing piece that we are also building towards. So think of it like bare metal upwards that management is something we're focused on. then the models and finally on the application layer we have a bunch of things that

we do uh very quickly the probably the biggest one is we have voice AI where you can converse with an AI agent in voice in Indian languages multiple Indian languages low latency low cost accurately right uh and we'll talk about volume there really scaling that up month on month uh then we have a product around the agentic agentic space where uh you can design agents that can do complex tasks put together things across connectors, skills and so on. Uh and finally we have uh a product around content creation. So um we have stuff on dubbing. We've been doing some cool stuff there. Uh you can translate books, you can even create audio books, podcasts, etc. So there's stuff around content and we are now moving towards the vision domain as well. So that's the application space, right? Um and we we are also sort of like across these three things there are crosscutting things.

Sometimes an application requires a particular fine-tuned model, requires a particular way of inferencing. We look at that cross cutting part as well. But this is roughly the uh lay of the land across the three layers. >> Very interesting and very complicated cuz uh globally what we are seeing is some of these like for example if you look at uh Hicksfield it's like very videocentric model >> and they just focus on that. So know one of our companies in the US just does audio models or 11 labs just does audio models for consumer enterprises like how much of this is look we should do this because it gets us from TRL3 on a certain product to a TRL7 and then maybe TL9 and we'll see what happens if you if it works it works and this is how a lot of the labs also operate what's your strategy on having this sprawl of launches and then without losing focus and quality >> so I Think even the definition of what focus and quality is changing with the kind of things the kind kinds of things that that that are happening right so therefore I think today uh I think you know various definitions are changing right people talk about small models right what is a small model is is the definition is changing about that right today maybe maybe two years ago somebody say a 7 billion parameter model is a small model today they may say okay maybe a you know a 30 billion parameter or 100 billion parameter is a small model so what I think some of those things are of course also changing and and of course the the speed at which you can do things uh is changing. So therefore it's very uh difficult to say that you know that um you know I'm going to take one way and I'm going to stick to that uh for a long period of time. I think that flexibility on how you can actually take all the pieces and put them together and putting them together has literally become you know you can do it at the speed of thought right literally if you have an idea and you think you want to do put things together in certain ways I think that that combinatorial way of putting them together in a way that is actually kind of uh scalable and usable has become much much easier right so therefore I think the opportunities are going to be many But you have to also move quickly to actually uh leverage those opportunities. I think that's so I think that the answer saying that there is a a master plan that I can follow for a long period. I think that time has has passed.

>> I love it. Uh I think we see a lot of the best companies globally just almost making sense of the mess as a as a scale up and it doesn't get easier. >> How do you guys like maintain speed and you you were you are now how big of a company? >> 200 200 people out there. How is that organized and how do people coordinate with each other? You know, you guys and then maybe just talk about you launched, you know, in a very quick cadence 14 things in the runup to the India AI summit and they were all amazing launches. They're like so well orchestrated uh a rebrand that went viral globally and a bunch of products that went viral globally. So much goodwill both in this market and everywhere else. People in the US were talking about it. You know, head of states were talking about it in the US and in India. I mean it couldn't have been just you wake up one day and let's just launch something and let's see it must have been you know years and years of work. So walk us through what led to

that. How did you orchestrated? I know some people didn't even know this launch was coming tomorrow but you're a team of I don't know a few hundred people. How do they how do they not know like how do you guys organize all this stuff? It's crazy.

>> No no I I so there is a fair amount of chaos that gets injected [laughter] in the company for various reasons. No but see uh I will talk about 14 launches but I think on the broader point of focus I think there's definite value to focus no doubt about it uh but there's also a whole plethora of things which have come and died so quickly in the space right so and on that day it does look like the best thing that is going to happen in AI I think we have seen this again and again so I mean we are builders we think about it outside in from the capability perspective and now the surface area of capability is now very very broad right So um the uh the cost of experimenting the cost of doing things now has come down as we said you can put things together quickly. I think what what the the intangibles that compound in my mind are what you alluded to brand uh talent and capital right uh I think uh what we have started to do now and absolutely with no uh in fact we didn't even have an effective brand marketing strategy and we started this whole thing and what we realized is out of the building that we did the design that we did the messaging that we put out there started creating value to the brand right and I think and the brand is not it's a substantial brand it's it says we will build AI for the country and we will do it from scratch and we are engineers. We're going to make it happen. We're interested in impact that is positive. So that's the brand and I think there is a lot of value in being able to attract the right talent to that brand. Now under the brand there is going to be a focused efforts towards frontier tech right that will continue and that is that is the core of why it exists right but on the application side we are going to be creative and we're going to try things and we're going to ensure that we can reach out to the long tale of use cases that matter in the country. Uh it's it's amazing we are in the market and then the variety of people we interacted with across uh like individual developers digital native small medium businesses enterprises government defense it's a widespread and that widespread of the market reflects a little bit in how serv also sort of builds things out right so so we are fairly comfortable as a company with having breadth in some places and focus in some places but recognizing that some things are compounding right so to your 14-day thing uh sort of we said hey we can't wait until the event which was the India impact summit that was happening in India to do something because hey there's going to a lot of noise on that day now how do we stand out that was the simple question and the idea was let's do 14 days of launches right because there were 14 days to go right so it was just a day before and but it >> we have 14 products to launch [laughter] sort of so so but but but that's also the as you said it's not what we built on that day the team was building and putting together. I think events like this concretize it. They make it clear for everybody. Uh and uh and I I think that will continue to be sort of like the style with which we operate. So we'll keep focusing on some core frontier part. Other sides be creative.

>> And how does that work? Cuz this like you got to have at least a cauldron of ideas that are keep bubbling up and you know reach a certain threshold where it's actually valuable to spend more time like hackathons is one way to do it. At Google X when I was there there was a bunch of these idea marathons they would do for three months and if you can't graduate to a product you get killed. you don't get funded and when Ruth Porat came obviously it becomes even harder to get funding even after six months of work but uh that's for another day how do you guys build this culture of forcing functions in time like you got to experiment you got to come up with interesting ideas and you got to get validation very quickly and then you build it into something that may eventually like Google used to have 20% projects and you know Gmail came out of that Google news came out

of that I'm sure a bunch of other interesting things came out of that what is your culture internally for that >> see by the way just just to uh sort of like uh little bit backtrack on what I said right see you still need to build scalable systems you still need to have things which are actually very good on core system metrics you can't you can't wipe code everything right so you need to have systems which are very good and you're somehow playing at top layers of that right so you have so for example voice AI is something we have scaled out right and that has been a journey I mean that's taken time we have sort of built every layer of it and learned it and across the whole stack of models applications and infra. So once you have core platforms like that you can play at the top layer. So I I I just don't want I don't want to discount sometimes you feel that you can wipe your way to a products. Uh that is perhaps possible but victory in the market is very different from getting something that is working. uh and uh I I think maybe the 2080 ratio should be maybe 50/50 today or maybe even more 80/20 even but that that that packing after the hacking is very important to actually sort of like get systems that are out there >> hacking before the hacking after the hacking.

>> Yeah. >> Awesome. So now let's talk about the core of the company, the heart of the company, the models. What has been your journey to build these models and then what would you do next? A field parameter model. >> Yeah. No, certainly we'll uh we'll be training much larger models. U one trillion you mentioned is is on the cards. >> India should build a one field parameter model. >> I think so. I think so. We need to >> say more. No, I I think see I think uh one of the things that uh you know I think this technology is so fundamental and it impacts so many different things that uh this is not an optionality for us to have right I think that uh uh and the ability of these larger models that certainly there are many things that a smaller model can do effectively but having the larger models actually to even if you're using it just to distill and make smaller models that are good at various specific tasks and things like that and some of these uh kind of things that we look I think having these larger models is I think table stakes for being in the game right I mean I would say that you know in you know some sense I think the one technology that I'd compare AI to is actually nuclear technology right and uh you know the world has countries which are you know have you know there are nuclear halves and there are nuclear have nots and if we look at AI right I think it's very clear that the United States and China are going to be AI powers, right? No other country is in that league yet and I think that I think you know I think we have to make a decision as as a country whether you know we are going to make a a bet for that. I think in the end I think uh AI has you know the ability to impact lives in a in a tremendous way right I think you know I think when you look at something like UPI right for the first time you say hey something's better in India than anywhere else in the world but then there are so many things uh that are not right and I think that AI actually and deployed in the right way can actually make lives of people uh you know significantly better and I think for that reason uh we have to be uh you know a nation where uh where we uh you know control our own AI destiny. I think it doesn't mean that uh you know uh you just have to have uh the you know I think if you're too far behind then you'll become a digital colony right and you know sometimes with pride we say we have the largest number of chat GPT users but you know what are we doing really we are exporting our data and like uh importing intelligence and uh can we actually afford to do that I think that's a question uh that we need to think deeply about >> let's say But we we've got to do it. We don't have a choice is what you're saying.

>> Uh I agree. We don't have a choice. >> The question is is it possible to do or is it a pipe dream that never happens? So this tell us more about like what are the things that we have done so far that allows us to say serbum can do this uh and then tell us let's say path from 100 billion to a trillion parameter model. What is the

gap? How do you go fix it? How do you go fill it? What do you need?

See firstly to on the previous question I would also like to say that uh this is not um um a summit to reach this is um a capability to have it's almost like think about it like uh weapon systems right uh weapon systems are systems and you can see where every country lies on that on every weapon system and you have to keep playing at the game right so I think AI is similar we are very very early in a dramatically different uh technology way we tier three or year four after chat GPT right and um and open AI started in 2015 they put in the hard yards to say we can actually scale anthropic showed that you can do that agentically Chinese companies have shown you can train that in the open source and so on right so I think we are in a narrative which is going to continue for a very long time so the most important decision for us is are we a substantial player in in the world and the answer has to be yes given our size given our civilization given the economy size and so on right so it's not even about today's model. The first point is it is are you on this table right and therefore no doubt we should be training models by the way not everybody needs to train models an IT services company can go use the model and make money out of it that's perfectly fine but there should be some bets that are made and more than one like let's let there be five 10 serums if there's possibility as you said there is need for capital there's need for talent those are the great remaining steps and and seram we have been sort of building that advantage and doing it but we definitely need to be playing at this game right and that's the first part on what we have done right and and you also alluded to the journey we have taken right uh notice that uh the founding team was not from one of these frontier labs that had already trained a model and so on so our job was if I were to look at it what is the metric of this episode that we just concluded which I call serum 1.0 to the metric was what fraction of the knowledge etc required to train a reasonably competitive model do we have in-house that was a fraction we were maximizing right and we took various steps in that journey uh first we went about working with llama which was then the available open model uh we actually uh uh sort of like fine-tuned llama models to show that we can add Indian language skills we did a we realized in that process that you can't bolt on language to an existing model but it was actually a deep collaboration um I sort of then joined the advisory board for meta sort of played a role there and we actually ended up contributing to llama 3.1 and we are the only company from India to contribute there so that's how it started we understood the process then when the first GPUs came to India notice there were not many GPUs in India when the first thousand hopper series GPUs got connected in India we were the first company to go and put our own money and say we'll pre-train a 2 billion parameter model. So we learned what pre-training data looks like. How do we create that? How do you create it for Indian languages? Went and pre-trained it, ran these GPUs and so on. So we learned a lot in the process.

In fact, Infosys also sort of like commercially bought a fine-tuned version of this for IT ops. And so we learned a little bit that step. uh then it was very clear that we have to learn post- trainining and so what we did was to take the mistral small model yeah I think the 30 billion parameter model and we learned how to do uh supervised fine-tuning reinforcement learning in fact at that time did not have a reasoning model and we through our post-raining uh pipeline were able to add Indian language skills post- trainin the model for reasoning actually had a hybrid model thinking non-thinking and scale out reinforcement learning and all these three efforts we didn't open source. We put our model, we put our detailed technical blogs on what we have done. Uh in fact the MR post was read by the Mistl guys and they also commended it. The first blog on OpenI was read by the meta guys and they commended it. So while the world felt hey okay these guys have picked whatever x million dollars and are doing something our internal journey was maximizing uh knowledge in this space right which we didn't

have we didn't inherit in some sense right and then when the first sort of bigger model bigger GPU opportunities came in India we started training from scratch and now what we have done over the last 6 8 months is uh trained from scratch these 30 billion parameter 105 billion parameter in fact we also have in house a 3 billion parameter model which is very good on audio and vision and things like that. But the key part here was we did the entire hog of creating pre-training data at scale. Often you need uh like 15 20 trillion tokens. We created that data set ourselves. Do we have a data vendor? No, we did all this work internally. Uh it is actually quite interesting because these models get better. They enable people like us to do better using open source models doing things on top of that. Then we uh did out the fine-tuning and the reinforcement learning inhouse right and then we released the model. In fact, we couldn't even finish the full amount of the supervised find I mean the the post- training part we'll release a update of the model with the full uh training done but this is an interesting milestone uh and to the best of my knowledge outside North America and China and if you leave aside mal nobody has pre-trained from scratch a 100 billion plus parameter model right so to V's point if you think of it like halves and have nots uh this is this is the this is the map right and now we have we trained this from scratch and it is important for us to now compound on top of this and go to the next larger step for us that's a trillion parameter model. uh even though India might use small models the quality of the best model in your class decides how good a small model is because most small models come from distilling from larger models right so it's super important for us to have a large model and the definition of large also changing from I mean we do know that there are even probably multi- trillion parameter models out there which are actually uh capable of next set of things right so for us uh it is really about uh to your point about is this frontier minus 3 69 I I think we have we I think we are slotting in about that 9 month uh period right now. We have to take that down to 6 months. And notice that we also have an advantage that let's for example the 105 billion parameter model on math uh reasoning stem was as good as the 600 billion parameter deepcar1 model. Right? So there's also the advantage that there is deflationary force or that the intelligence is getting compressed into smaller models. So we have these two forces playing out. models are getting more capable at a given size and we are behind by a few months in terms of what we can do and we keep uh moving right and our intent is that we should be able to have a no regret alternative with using a servum stack of models right at the smallest model the multimodal models and the reasoning large models end to end uh in about 6 to9 months right of course there'll be better models out there which will continue to push the frontier but you have a fully no regret alternative with the stack that we have right >> so so say more about this because a I would love to understand if this going to be so let's say you build out the 1 trillion parameter model I'm curious to sort of know what does it cost a frontier lab to build that today for the you know if they're ahead of everybody else and does it cost the same to somebody who is 6 months 9 months behind and what if not then what compresses the cost because cost also matters access to GPUs also matter if it's all apples to apples and it's just very hard and then the question is like what's the Right?

You're 6 months behind. They took a billion dollars. You're taking you're taking a billion dollars and you're 6 months beh like what's the point? So what's your view of like you when you say no regret alternative? Uh there is a six month gap maybe a three-month gap in a six month time frame but it's a gap. Do people care? Some people care seems like some don't. What's your world view of who cares? And people who don't why don't they care? What are they optimizing for if they don't care about the soda models? See I think that in terms of to your question what brings the cost down uh first of all the the GPU family itself brings the cost down right so every new generation of GPUs has enabled pre-training at larger scales at lower costs uh quantization aware training is

reducing the costs again and so on so in fact that is dramatic it's not it's not moose law scale it's actually happening much much faster right so uh if I were to look at a given benchmark on a particular uh particular test right the cost of training a model for that has been reducing dramatically. So there is there is that effect right? So there are two things that we are not doing right. uh firstly we are not putting out B2C products at zero cost right that is a major burn for these organizations right at least some of the organizations that are positioned in that way right the second thing is that there is a lot of experimentation happening at the utmost frontier uh and these are expensive experiments right they take a lot of compute and a lot of things there again we are not doing that now will we do that in 5 years maybe yes right but we are very focused on building the next capable model at the next size that matters right and I think the the cost structure for that the team structure for that the uh the overall effort towards that is very different right and I think the to a point about who cares in the market I think people will use the best thing that they're getting on that particular day but the ability the fact for us to be playing in the supply chain right that hey we have a model on whatever day that can do this becomes important for us to then play in the market and I'm sure that uh you are recognizing this as well. There are there's a lot of uh uh energy in the system where people would like to have things running in their control. Uh AI is such an important component to what you have and we would like to be able to give people models which is which are theirs. They can fine-tune on their data sets, maybe done reinforcement learning on their environments and then they think it's part of my infrastructure.

I'm running it running it. That's the business we are playing for, right? organizations wanting models to be part of their own security envelope which they can run with no regret. It cannot be that sovereignty which is not just a national concern. It's a concern for every company comes at a cost. uh that should not be the case right and hence the uh companies that matter the use cases that matter would would would play out but that's what that's what we are optimizing >> and from here to a trillion parameter model what do you think the company needs to prove or do like acquire data sets at much larger scale acquire capital where is that going to come from what's your strategy there is capital something you worry about >> no I think uh you know obviously you need capital but I think uh Just to uh as to to reinforce what Prat said I think the fact that we are looking at frontier minus 9 months kind of a thing the experimentation right I think that if you're if you're actually absolutely at the frontier the experimentation is extremely expensive >> and if you're doing it later you're running on better hardware so therefore that's also and I think that is and and but I think of the two I think experimentation is probably the bigger thing once you know that uh you know something can be done and somebody does it it somehow becomes much easier for people to actually follow and then and therefore I think being a fast follower you know does reduce but having said that you know I think these things nevertheless need significant capital and we think that we're also going to take a cost view right in the sense that what is the you know I mean today the the earliest adopters of something right uh do something independent of what the cost is or the thing that they're doing is so high value that it doesn't matter what the cost is but at some stage the cost will begin to matter >> and I think that you know if you actually play play for that game that you know there will be these tasks where the cost will begin to matter uh will also be in a very important place uh for us to play in addition to the you know the sovereignty of the enterprise and this thing and I think as far as capital is concerned I think we've earned the right to play and I think that you know we will you know kind of uh have the capital to play this next round out. I think that's something that uh that we see and and uh we're we're quite confident that uh you know uh that uh we'll be able to do that. I'm sure the capital is of interest to you but uh the u just want to say one thing right so while we are talking about frontier minus x this is still high high-grade R&D right taking and

running reinforcement learning on a large model at scale is not something that everybody's doing or having the ability in a single company to create the data to do the engineering to do the training do the evals and then put them in products right this is high-grade work right uh it's just that we we aren't playing at the hundreds of billions of valuation where things are and you can still have an important game to play but is this something that lots and lots of companies would do that I would not bet on right so I think >> it looks like in the world right now probably five or six companies are playing this game and you guys are another five or six >> so for example Nvidia picked a set of companies globally to be part of its open source effort and uh it's called Neotron coalition and we were there along with the friends from Mistral and other places right but but the bottom line is towards a trillion parameter model. There is a lot of R&D that gets done, right? And that gets created within within Salv. There's a lot of IP around data, how do you train, what is the data mixture? All of these are hard problems. They they're not even things that are being done in research papers, right? So they only you play that game only once you get the compute, the capital and and the talent to do it, right? And so I just wanted to capture that while it is fast following, it's fast following in probably the most bleeding edge effort of humanity, right?

and uh and it's a matter of capital and it's a matter of talent and so on. >> So your question from here to a trillion parameters uh and it's a journey at some point it'll be 10 trillion parameters wouldn't be enough. I get that. What's the capital do you think it'll require? Is it in the billions or tens of billions or hundreds of millions and what's the number of GPUs we require and how much do we have right now?

>> No see I I think uh the uh capital is is a continuous game right? I think the zeros will keep adding. I suppose I think right now in the game that we are playing we by the way we have been as a company pretty uh let's say focused I wouldn't say frugal I think we've been though we've been creative on things we have done we've been very focused on how we uh do this right so I think it's hundreds of millions of dollars that is required to have as I said a no regret alternative stack right that's that's what we are building right uh and uh and from there it is matter of scaling out the usage which is when the uh the scaling out of GPUs have happens even more uh it's very very critical that Saram plays a role in sovereign infrastructure scale out in India. So working with data center providers to have a game where there is a lot of compute available for this stack to be then used at scale in a country >> is it in the millions or hundreds of thousands >> GPUs GPUs has to be I mean let's in the country today we have not that many GPUs right but scaling up uh I think tens of thousands of blackwells is what the country needs uh already if you consider making AI available at scale to people today right and it'll keep increasing as more people uh adopt this right so and and the bottom line is this right and though it sounds like a little bit more on the impact side the bottom line is our job is to maximize token consumption from various products efforts that we are doing for things like for example every student having a tutor every person having a doctor right uh these are the kinds of things that the the demand for this is is insane >> almost like the per capita token utilization is zero right now >> a little bit more uh for sure a major fraction as viv said is outside but it has to become four five orders managed more right and and that is what we are playing for right >> yeah so now let's talk about let's pull it all together right I mean there is a ton of technology product building out uh and I want to maybe just showcase this book that I uh read maybe six seven eight years ago almost it's called crossing the chasm I would recommend it to any and every founder especially as they're entering a wave like this one uh I'll read a small snippet from the from the back which I think summarizes is what the concept that they're trying to cover here is. But the idea of this is like technology alone in these large waves are not is not enough like you got to almost convince the market that there's a need for this in your life.

There's importance of electricity of oil of you know CPUs and GPUs and AI tokens into your life. Uh so what the book essentially covers and I'm going to read it out really uh quickly is that in technology adoption life cycle which begins with innovators and moves to early adopters then early majority and then late majority and then finally legards. So it starts with early adopters, people like you, me and people that are watching this uh podcast perhaps early doctors then early majority which is probably the next 500 million people who are just waiting and I will not take this new airline because it's the first time it's flying but I'll I'll wait for a month and if no crashes I'll go pick that and I' I've done that.

I'm an early majority for airlines and things that could crash. Um then late majority which I think is everybody else and then the lagard which are like you know the lagards. Uh there is a vast chasm between the early adopters people like you and me and the early majority they're right outside the door. And while early adopters are willing to sacrifice for the advantage of being first the early majority waits until they know that the technology actually offers improvements in productivity in their lives. The challenge of innovators for innovators and marketeteers is to narrow this chasm and ultimately accelerate adoption from the early adopters to the early majority. So that this is there's a lot of technology being handed in a nice packaging to the people. The technology does not need to change year or year as much over time but the packaging needs to adjust and then assimilate this into society. How should society in India and globally think of serb? What do we stand for? Are we an open AI alternative? Are we a palanteer? Are we a AI enabled Price Waterhouse Coopers? Are we all of it put together? Like how do you see yourself?

And how should the world see you? And what's your brand going to be 5 years out? What would you want to stand for? >> Okay. Um yeah, multiple questions. It's a very very uh interesting uh way to think about things. I think it's it stands true across the technology generations, right? See, I I think it's interesting question. I feel that there will be dispersion across individuals and across organizations, right? Individuals who have some amount of agency in life will adopt this technology because the proof will be so adjacent to you. It's not like I want to I want to move to NoSQL from SQL, right?

You'll have to figure out whether it works and so this is like you try something, it works and you see the value of it, right? For example, for us, the way to think about it is our voice AI platform going from an enterprise to a small and medium business, right? The small and medium business guy would be like this sees the value right away and they like to adopt it, right? So, I think it's it's a matter of agency. If if individuals have agencies, they will adopt and they will create value for themselves and they'll move ahead, right? And it's going to go down from years to months to weeks in terms of being on the leading side. I think the challenge is more at an organizational level and eventually at the society level because the uh the tax here is human alignment right aligning with your leadership that this is worth doing.

Aligning with people that hey this is not uh about job loss aligning even society that it is important to build uh and not and to innovate. These are obvious things in our mind right like we have strong believers that we should innovate but there is doubt in the society that we should innovate right these are the choices that we have deal with and I think organizations will struggle right uh to adopt individuals will see the proof very very quickly do >> do you feel that people misunderstand what serum stands for I mean there has been misunderstanding in the

past how how do you guys see this now is it changing for you guys >> no I I I think see I think uh you know obviously we are a company which you know is is has a lot of ambition and I think that uh you know I think also things uh you know are evolving right so therefore and it's and people see some aspect of what serum is doing and they're looking at it from that lens you know I think they're not looking at it from a holistic perspective so therefore then then that that may lead to some kind of a a perception about what sur is somebody may think oh you're a you know um a voice AI company or somebody else thinks you're building foundation models or something else. So there are obviously people who have uh uh you know different views and some but but I think in the end uh I think uh you know if you if you want to look at the ambition of serv right I think that uh you know what I'd say is that if a significant percentage of all tokens uh kind of used in the country in the next few years are from serv then then we've been successful at what we want to be able to do. So and that's you know that's a big ambition and I think that it's also uh if that is happening at a large scale then it means that people are using AI at a scale that was you know that that really impacts their life and I think the key point that Pratish made about agency right people who have agency will adopt quickly organizations also you need to figure out the people in the organizations who have agency and those are the people you have to convince to actually make this change and that then you know then then whole organizations can move and shift. So therefore I think that's part of the to be able to influence people who actually have agencies in different kinds of organization actually will be something that and when they individually see what can happen >> I think that's the I think in in the sense it is today if you can imagine what the future will be you can work towards building it but if you don't have that imagination is really our job to to make happen. So on the perception thing right see um I mean different people said hey why are you fine-tuning models and in fact I was in some meeting where uh two days back and people were asking me some seven eight times is this train from scratch right and then they don't believe that we did it and so on I think that is entirely fair right people uh see there aren't many efforts in India which have the luxury of being built to do deep tech work right actually that's a in some sense that's a poverty of the country. We don't have enough bets on that. This is so natural in a country like the US, right? So there is a lot of expectation out of Sarbam, right? And and that expectation comes with sometimes people feeling that they didn't stand up to that expectation. I think that's that's good stuff. I think we are mature enough to not uh worry too much about that and so on. Right? It is it is an entirely different matter that we are playing multiple games. Uh that is not usual like you list a list of companies and so on. Right? uh we are very interested in voice AI scaling that out as an infrastructure not a product not a platform infrastructure for the country we are very interested in having uh sort of like a co-work like experience which is on self-hosted models which you can scale out at lower cost with sovereignty we're very interested in building for the creators because language is such an important thing for us we're interested in building frontier models we're interested in running with data center companies and building the infra so we are interested in all of these things right and depending on who is looking at what slice of it that elephant metaphor is going to show up right but we are we are comfortable with that right I think the main thing is and this is what I think Vic is saying right I think this is a time when the mass the most important failure is not being ambitious and that I think at Som we are carrying that ambition right that we have to think really big and make it happen I think it's very doable the the tech is is good there is value to be created in India and it's good that now we are seeing investors see that opportunity and really scale up from that point of view >> I'm just curious would you ever want to So global what does that look like? Who do you compete with? What's your strategy? What's your right to win globally on >> on the global ambition? See again we we remain clear on our what we are building right and currently that buildout is very specific to uh having the technical capability inside and the market positioning in India right but the technical

capability that we build is not conditioned on specific things about India that is applicable elsewhere but making a market positioning that can actually capture value is a business decision that we'll make as we scale up in fact we are going to be having very important leaders come into the organiz organization who will then sort of like take it to the next stage. In fact, we call it internally serv 2.0 starts now and it's going to be key leaders coming in who have wider vision, wider understanding of this and then we make very clear market decisions on that. We are clear that we the technical capability is generic and is has applicability in in a wide space but what we focus on will become a business decision. I think going global or not I think will also depend on having the right partners and India has some strengths on that which we will we'll explore.

>> Yeah. Amazing. Uh I know we on time and you guys are really busy and uh but it's been a real pleasure. Uh really excited to bring you back in hopefully in a few months and talk about Serban 2.0. >> Yes. >> Thank you again for your time and uh yeah it's been awesome. >> Nice chatting with you. >> Nice chatting with Thanks.