(no title)

66 MIN · YOUTUBE · HTTPS://YOUTU.BE/VFF011KO89O
https://youtu.be/vfF011ko89o

SUMMARY

Maddie, the founder and CEO of 11 Labs, discusses the evolution of audio and voice technology, emphasizing the company's journey from addressing basic audio dubbing issues to creating advanced AI-driven voice solutions. He highlights the importance of community engagement, iterative development, and the need for emotional expressivity in AI-generated speech.

- *11 Labs started with the goal of improving audio dubbing, particularly in languages with poor voice representation.*
- *The company initially focused on text-to-speech technology, leveraging community feedback to refine their models.*
- *Emotional expressivity in AI-generated speech is a key focus, with advancements allowing for more nuanced and context-aware voice outputs.*
- *Maddie emphasizes the importance of collaboration within the industry, viewing competitors as potential partners.*
- *The company has achieved significant revenue growth, reaching over $430 million in just three years, with a focus on both enterprise and self-serve models.*
- *Pricing strategies are based on delivering value rather than costs, aiming to capture a fraction of the value provided to customers.*
- *Future developments include on-device models and enhanced interactive capabilities, while maintaining high-quality outputs.*
- *Maddie stresses the need for robust safety measures and ethical considerations in voice technology, particularly regarding voice replication and authentication.*

Welcome to week two of CS 153, also known as AI Coachella. We are super lucky to be kicking off this week with Maddie. Mattie is the founder and CEO of 11 Labs. How many people here have heard of 11 Labs? All right, so pretty much everybody. Maddie and I go back a ways. About three years ago, I think when I was still running platform at Discord, um a friend said, you know, an there's a a little bot, like a a texttospech bot on Discord that's blowing up. Um you should check it out. Um and you know, we had a lot going on at the time at Discord. And so I I I actually didn't and I should have.

And then a month later, somebody pinged me again and said, "You really should check out this bot." And I I checked it out. It was called 11 Labs. And it was quite an extraordinary bot. It it it was a um a Discord bot that allowed you to generate audio clips with just a text prompt. Uh and within 24 hours, I'd asked one of our mutual friends, Nat Freiedman, to introduce us. Mattie was gracious enough to explain what they were working on. I had you'd let me come on as an angel investor. So, thank you.

Um and since then Maddie has gone on to build one of the most um the fastest growing uh one of the most widely used and I would say trusted brands and services in frontier audio and speech. Um so thank you for joining us Maddie. >> Thank you. >> Thank you so much. Good morning everyone. It was also a crazy thing an uh when we met for the first time it was me and my co-founder P. uh we both came from Google and Palunteer before that.

So we were trying to like redo the company setup from scratch of like what not to do and we tried to like go against some of the lessons from those days. Um so we were allergic to meetings, we were allergic to um to like any email based communication internally but we also want wanted to not do any of the internal communication the standard way. So when we started we actually run a company on Discord. I did not know that. Oh, in that conversation you were you were helping us a on on on on the text to speech and we were trying to like figure out is that the right play for us to base all the company on Discord. We swapped from to Slack >> uh which which was which was uh easier for Freddic but that was a that was an interesting few few first months of trying to build all the bots on Discord to like make it easy and quick for us.

Th this was a bit of a theme we talked about last year too, which is that often gaming ends up being this petri dish for innovation. Some of the hardest infra product design experience problems that are solved in gaming then become sort of uh um leading indicators for the rest of the world. And the stuff you were doing and a bunch of other our friends were doing on Discord at the time have ended up becoming indicative of you know value in AI a few years later. Is that do you feel like that's an a true assessment or uh am I overfitting?

>> Yeah, I think that the the true part there which you know we we were following the journey model at the time of like how Dave built that community piece on this and for us at 11 Labs when we started we knew that we want to fix two things. We want to fix the research and foundational models around audio and voice and then build product around that to to bring that AI into more of an applied AI setting and fix the problems that our customers are facing. We started on a very PLG driven motion. So working on the product led growth with a lot of the the the creators in the space of developers in the space. And we thought that the best way to do it is close the the the loop as close as possible to the people that are using those tools and like Discord at the time and and and and generally keeping access open to a lot of those creators and developers was the best way for us to learn is it good enough? is the quality finally there to to serve the needs they have to what are the use cases we might not predict that people might want to build so we can bring that uh quicker and then free um and that's still a big tissue today of our work across that is um we want to work with the community to find a ways for them to contribute back to the product development and of course whether that's using the models to refine based on the data of of how you use the model all the way through like can you contribute in our case we created voice marketplace where people can contribute their voice to to be used by others. So that community aspect was very important and it's always a true where I feel that technology adopted by the community will show you use cases that might like diffuse to the rest of the world 6 12 18 months later. So being like close is is super valuable but more so more so than not in that early days just uh I think you need to be like extremely problem obsessed. what is the problem that they are having and and the variation of what you think the problem is to what the customer actually thinks is a problem is slightly is slightly different.

>> Okay, let's actually stop take a beat there. Can you go back? Take us back in time. What was the problem you guys were obsessed with when you started 11? What is it today? How might it evolve? Give people a bit of a 11 Labs 101. How do how do you get here? >> Cool. The whole chronology I will I'll I'll give you a zoom in into that first day and then and then and accelerate over last few years. But when we started, so I'm from Poland, my co-founder is from Poland. A very peculiar thing that happens in Poland is that if you watch a foreign movie in Polish, all the voices, whether that's a male voice or a female voice, get narrated with one single character. So you have one voice reading every character. As you can imagine, a pretty pretty terrible experience. And you would think that with the modern technology, this is a problem that would have f been fixed. And no, it's still the case. So most of the content is delivered this way. So that was the >> whose voice was this? Who did they >> They have five characters. There's like five of those voices usually monotone male deep old voices. Uh and the the it's also crazy because the part of the thing is they are kind of encouraged to deliver the the movie in a flat delivery. So the audience can interpret the emotions for themselves.

>> Uh which is like another another level. >> Expecting a lot from the audience. >> They do expect a lot. Um so if you like any Polish person if you ask them they will like account how like not good experience that is and when you learn English you finally get to learn everything in original and that's a an extremely positive one. So that was like the first piece and inspiration for us. We know the future is different. The future will be where you can access all types of content in any language uh with that incredible tonality, incredible emotions. So, so uh so we left Google, we left Palanteer at the time. And >> were you guys both both in the Bay at the time?

>> We are both uh between Warso and London. So at the time when we started, we started in in in London then moved to Warso for a little bit then moved back to London. >> Yep. >> So from Europe uh at the time which actually was uh for that part was pretty useful because the whole language thing is is a big problem there less of a problem in US. the kind of the inspiration might might have not occurred otherwise >> and then we knew that we want to fix that AI dubbing problem and as we started going deeper we knew that we need to understand two parts we need to understand whether there is a a a research potential for us fixing it and two whether there's an actual product need actual problem for the customer so so p my co-founder and amazing researcher started diving in and quickly realized that the current models could potentially create a Frankenstein version of the dubbing. They're not good enough, but they are almost giving you an experience where you can switch from language A to language B, preserve the voice, preserve the inonation, emotions, but still not good enough. So, there will be research, but possible research to fix.

>> Now, in that research, a very important part is as you think about the dubbing problem from speech, there are three models that need to play a a role. You need to transcribe what's happening, who is the speaker, remove the background sounds, background noise. Then you bring it to text. You translate it to the another language. Uh you might need to do additional set of corrections. And then you recast it back on the other side with text to speech to produce audio at the other side borrowing from the original performance. So there are those three key models, transcription, translation, and and the text to speech on the other side. And you need to fix all the components to make it good.

Well, at the same time, I was trying to figure out does anybody need that? So we would call the email of creators

of studios saying like hey if dubbing was a problem if sorry if dubbing was possible automatically and you could take your movie and bring it to all international language with your own voice would you be interested every so often we would give give them like a early version of the samples and um and frequently the reply we got was yes interested but if you can do that could you also do just a simpler voiceover corrections for me because when I record let's say even as we record now after that some of the parts are not recorded properly and I want to fix them or could you replace my voice from the from the script just in my original language so I can narrate that without appearing in the script and screen um and that was like a very clear piece where as he dived into the space on research and we dived into the user problem space it was clear that there's other problems that we can fix first >> and we can focus research on one of the components instead of all of those components and and and actually bring that to the market and bring that to to that.

>> A couple of things I mean the class is called Frontier Systems, right? Um if you notice what Maddie just talked about is the anatomy of what is their intelligence pipeline. Remember last last week we talked about the anatomy of of how um intelligence is manufactured. But he just gave you a breakdown into the system of of that 11 was trying to build at the time which was this cascaded workflow, right? You had the TT, you had an LLM in the middle to do some reasoning essentially, >> right? And then you have a had a speech to text model the other way around. You had transcription, you had an LLM, and then you had >> Exactly.

>> a generative model. >> Exactly. And like it was still 2022. So this was a year where the topics of the day were crypto and metaverse. So that it was still before the GD moment, the good the different days. And LM translation was still relatively poor. So we're like using both attempts. So like the whole pipeline didn't produce the right result which was a great piece for us being close to the users trying to get like what is the actual problem because we then shifted okay let's fix the text to speech which is the most common denominator of of all those problems and let's fix the voice over for the creators and then we've noticed that there's other parts of like just reading the scripts reading articles reading books that we can deliver. So in 2022 we decided that the first set of of the biggest potential will be the more basic version which is just bringing text into audio making it sound human making it sound emotional uh in just English. So you decided not to innovate on the transcription part or the LLM part just the last mile which was generation and you said the mission there was let's try to improve the state-of-the-art in uh like it had to sound natural.

>> It needed to sound natural. So the the main things that were not possible at the time is you couldn't really replicate a voice um with the same characteristics and two you couldn't make it sound and follow the the entire delivery in the way you would expect. So if you have a fragment of text, it's a say a happy happy sentence. If you are reading it, you know that it's happy. You deliver that in a happy way. If it's a dialogue sequence in the book, you know it's a dialogue. So when you read it or a voice actor would read it, you know you need to read it as a dialogue >> because you have context of >> you have the context of the entire thing. And um and at the time at the same time the LM breakers started occurring where you knew that you could predict the next token based on the previous token. So you could bring that context into account in a smarter way.

So that was the first big thing that we knew we could solve. And two, we could recreate the voice characteristics of uh a lot better. So instead of the common approach at the time was effectively hard- coding parameters of a

voice, >> right? >> The the gender, the accent, the age of the voice and trying to predict those. Instead, we can keep them more abstracted and let the model try to define what those parameters are. And and and those two innovations we knew will affect the research side potentially. And then on the product side, we know knew that a lot of those people will want to create their longer forms with books and audiobooks, create uh scripts and turn them into audio, do the corrections on the on the voice over for a video. So we knew that we'll need to build a product for making that easy and and go all in on helping the creators and of course the API to developers to build to build with that.

So at at this it's 2022, you've had this clarity now that okay, the the state-of-the-art the frontier we're going to push is the the this this f flexibility of the voice, the generation part, right? Um but it was just you and Peter. You hadn't raised any money. I don't think so. >> Um yet? Yeah. >> What did you do first? Did you >> Did you go and look for an open model? Did you try to go call somebody up and use an API? What what was step one? Step one, we we of course were drawing from our savings to to to to get the models off the ground and um and and get that first like are we solving the right problem and I you know that that I think the the clear thing for for for anyone here is it's it was so valuable for us to be as close to the users as possible to keep that interactive loop pretty quick. Um but then to your point like the the main thing as we were like actually trying to do the research was of course look what's available on open source what's available in closed source and um and then look for the papers in the space are there other innovations in those papers that might have not been applied to audio and that was the case uh there was uh the closed source was still lagging not nothing really good but at the time the hyperscalers so the Googles of the world were still kind of leading on the best text to speech open source was ahead so you had some uh inklings of incredible voice models. There was a a famous model uh called Tortoise from from an incredible guy, James Becker.

>> We're going to have James come do office hours in the class for everybody. >> Okay. Amazing. But the crazy thing about James, so he was he was at Google at the time. uh he was working I think on infra or something unrelated and in his spare time he was exploring audio and voice and effectively created the best open source model of that time >> on nights and weekends basically on nights and weekends >> as a side project >> uh and uh and it was it was it was finally the model that could produce something that sounded humanlike on short fragments so you could have amazing delivery the right process the right inonation right emotions however two things weren't possible with with that model one it took extremely long time to generate anything and then two it was very unstable below um above that short sentence. So that was a a tricky limitation. Uh and then on the papers there was of course a lot of good innovations coming from the space. The diffusion in 2021 came out um uh the I mean the transformer was like four years prior to that but still some of the ideas from transformer were just getting through to the audio space. So we knew that combining some of those ideas, we can we can potentially take inspiration from open source, take inspiration from those papers and try a slightly different architecture for text to speech and um and for voice creation of how you create those voices.

>> Do you remember by any chance how much compute you spent on the first 11 checkpoint that was inspired by tortoise? >> It was tiny amounts in comparison. It was tiny. So one thing that I recommend uh if you are looking to start a side project or a company is uh is look through all the uh accelerator and quotes programs from big companies that give you free compute and free credits. >> Well that world is gone. There's no free comput anymore.

>> Okay except for maybe in for this class of students actually. >> I should know that before saying this I guess but there was a great maybe they still do something they so Nvidia inception program and a few others like gave us >> I remember that. So we we >> back when GPUs were still available. >> Exactly. So like the first ones for us were in like in tens of thousands of dollars and that felt that felt that felt big. Maybe it was like approaching the $100,000 category but like still tiny. Um I remember like you know like in that days the budgeting and and saying like um like optimizing your budget is so important. It's not a compute piece, but I remember us arguing whether we should do a patent for our work, which we decided against. Uh, but I remember the the the lawyer quoted us uh for the entirety of the patent work.

$6,000 and we said there's no way we are paying that amount of money for for for anything at the stage and decided not to do it. >> But it seems like not having a patent has not held labs back. I think like ultimately like is it is it even valuable that the the you're innovating so quickly that it becomes obsolete too like would you ever want to stop innovation from the other side? So so likely not. Do you do it as a defensive measures?

>> So like then we learn about the whole patent trolls industry that other people that will try to attack you with their own patents that aren't great but just waste your time. Um so like do you do that? Then we decided like no, we'll just fight those cases anyway. So, so, so it didn't stop us. But, um, but bottom line, so small models in comparison, they were like in hundreds of millions uh, parameter models at a time, if not lower. Uh, um, so that was relatively small texttospech models. And you asked earlier just to give you like a quick so that was 2022 and then the kind of the common facet across 11 labs is we continued across research over last years. So the models of course got bigger and that applied across the entirety of audio. So we did texttospech model to start then built a transcription model or speechtoext model to understand what's happening on the audio side. Um the wider set of how you bring those models together in AI dubbing how you bring those models in interaction. So how you bring a conversational tissue around the speech to text the LM the text to speech together and even expanded to music. So entirety of audio and how voice models or audio models work together with other modalities and then uh alongside build a platform that helps businesses and developers um uh uh um or solo creators transform how they interact with their audience or how they interact with their uh their um uh uh their people um through agents and support and sales uh with creative tools and marketing and storytelling. So, how you can create that interactive tissue between between an entity and the audience it serves.

Um, and that's like the common common piece as a >> I I forget when it was cuz these timelines are so blurry now, but there was a moment where I woke up um to like a like six different people texting me a link to a tweet by I think um of of a speech by Javier Mille. >> Mhm. >> Which which had been completely dubbed with 11 Labs. >> Yes. When was was that a year ago?

>> That was uh two years year ago. Exactly. >> I feel like everything changed after some something changed. What what happened? >> Yeah. So it was uh so we to give you a rough timeline year by year of like how audio models develop. So 2022 first breakthrough uh which which my co-founder is like the smartest researcher in the space I I got to know was finally able to bring into the field of how you like get that context get that that tonality out. So that was 2022. 2023 is where you started seeing expansion around the wider voice over, wider narration space. So you could use texttospech model across languages, create more voices. Uh we created the

ability for people to recreate their own voice and a high quality created a marketplace around that and a wider set of creative tooling to help you with the audio books uh for authors uh creative tooling for authors. Then 2024 finally brought good transcription models combined with the LM for translation combined with the speech generation. So you finally had the AI localization version. So that was the Javier Malay speech. So the project was >> there's like few versions of that. He delivered his speech on UN and and we brought it from Argent uh from Argentinian into English so people could could could could listen to that while still having his iconic delivery. And then at the same year we worked with Lex Freedman on on the full conversations that he had with different world leaders. So Javier Mleo of Argentina all the way through to President Zilinski of Ukraine. Um uh to later on with Narendra Modi of India where you could hear both of them speak that language which was which was a new opening. So that was 2024. Um and then 2025 what happened is you could finally like roughly finally have those models act in a real-time basis. So you could create more of the interactive voice agent experiences where you could uh you know you mentioned uh the kind of the frontier system architecture of combining the stack. The same kind of stack applies in the voice agent uh uh uh comp combo where you can have this cascaded architecture of having speechto text transcription. Then you have LM to generate responses back and text to speech to to narrated. So if you are speaking or voice agent then it can predict are you stopping the sentence start generating the response um and all of those components can work in tandem to create that uh uh that experience. So that was 2025 and I think 2026 we'll see an extension of that of how maybe cascaded I'm going too much into detail but cascade go to fused or cascaded continual uh cutting the latency to be great but the back to your question the the first great uh AI dubbing experiences in where in that static context in 2024 and um and we're really good. So, thank you for the the the cheat sheet on on how we got here. About a year ago, you and I started talking about, okay, where does the space go next? And I remember you saying, an I, I think, you know, we've had this cascaded system so far, but now we we we're starting to see what people want to do with it. And a big a big part of what people want from a capabilities perspective is is sort of a more is deeper reasoning from these systems, right? Especially an audio agent that can understand the tone, the voice, the the inflection, the accent of what's coming in because you lose all that context when you transcribe just pure audio. Um I remember you saying um that there were going to be more and more I either some some forms of unification of these modalities um or some new breakthrough where we try to combine these into omni models um where are we right now in terms of how far are we from these like the when I talked to you know James Ber is a good example Matty just talked about James uh shortly after open sourcing tortois TTS James went to OpenAI and worked on ChachiPT advanced voice mode and when I talk to the Chach advanced voice mode, it still doesn't understand if I'm angry or sad. If if if I said the same word, the same sentence in a really angry way or in a sad way, it just transcribes it. And the audio understanding, the semantic understanding of the audio is very limited. Why is that? And what is it going to take to make the systems more capable of understanding what's coming in?

>> Does that make sense? >> Makes sense completely. And and it's you know one of the huge questions and we we kind of changed our own perspective over the last couple of years like what is the what's the right uh approach across the different um uh uh domains across other fields. Um so to to Andre's point uh you know when you have a cascaded architecture frequently what would happen in the past you would transcribe just the text element pass it over to the LLM to generate and then of course you would narrate that on the other side. Uh now of course that's a relatively poor version of that. Uh what um what what you could theoretically do is is still try to capture a lot more of that information. Um and now generally you have those two approaches as you

think about the future. One is you continue on the cascaded side. So you continue with the three different models and in potentially think about how they can be improved to continue bringing the right context, optimize latency, continue reliability. Um or you can think about training a model together where you kind of combine all of them together and then you can think in between where maybe you try to create a a transcription and the L model at the same time but keep the other side separate. So there's like everything in the in the middle. So it's not a truly binary choice, but that's the the big question many people in the audio side will will be um alluding to like do you go the fuse approach where you train them all together and generate effectively when I say something the voice agent automatically generates a speech talk and it doesn't go through text or do you go through text like where we are today as we think about our approach to the space as we think about the enterprise the business use cases where the reliability and the smartness and the intelligence of the model is one of the key things. We think the cascaded approach is the right thing for the next for the next few years. And if you are thinking about places where maybe the reliability isn't as essential but the speed is um we think the fused approach will be and as in so many of those cases probably for for within a given customer you'll probably want to blend and use a little bit of one model and the other depending on what you are solving for.

Now to the actual question we think so there's like kind of those three key parameters the quality or the emotionality of the speech uh and the experience of the interaction two is the reliability does it interact not hallucinate cause the right tools behind the scenes in the right way and then the latency we think emotionality is fixable in both approaches so we hope that later this year already when you are speaking with voice agent you're excited the agent can respond in excited way if someone is calling and stressed, we the agent can respond in a reassuring way.

And we actually just released a new version in our voice agent work which just trying to do that trying to detect the emotions on the transcription side, pass that over to the LM uses that as the context and then generates response accordingly and um and that was one of the big like big breakthroughs on our side to create this expressivity um to be able to have this expressive mode finally work. What happened behind the scenes to make this possible? And what was the hard thing is you didn't have that much data that could tell you is that delivery peppy, sad, stressed. So over last year we effectively investing a lot to create this whole labeling exercise to be able to create a data to train that model and control it. So finally we think you have the expressivity and the second part is you can actually make it controllable. So you can define what type of experiences should happen based on those emotions.

So you think that expressivity or emotionality quality is largely fixable in both uh or like we'll be very close in par. Our friends at Sesame are doing like a great example of how you can >> Yeah. Brendan will be speaking in the class. >> Amazing. So he will probably tell you of like how they are trying to like finally get through um that uh that that cusp as well. So it's almost a race who will be first between us to pass that like emotional voice cheering test across other any of those circumstances. Then the second le level is the reliability level and here in general you want the smartest model. You want the smartest model to work in the agent and and of course we are seeing so much of the innovation happening across LMS our customers the businesses whether it's Deutsche Telecom or Revolute or Clara all of them will have a slightly different models they will want to use um depending on their use case that we can empower them to to to use. Um and then the second thing in that reliability as you think about voice agent let's say you are calling in to customer support to rebook your ticket

um for for flying to to Poland. Um the you will want it to authenticate your account. You will want to pull your own details. If you are to process the payment you want to make sure it's your uh details that are attached. So it needs to be reliable. It needs to follow that flow. And here you are calling so many different tools in the in the in the in in those steps. So you will uh uh likely go through some two-factor authentication to get the code. Uh you'll likely want your your email and the database to be pulled up from the information from that uh from that email. So all of that needs to work reliable. It will take time. Um so how you orchestrate that part is important.

Cascaded models work already really well on that given the intelligence layer will be fixing that. And the fuse you sacrifice that you are kind of you need to then bring all that tooling into the fuse model. um and and and becomes very tricky to track what happened in each of the steps. What happened on the transcription step, what happened step, what happened on text to speech side and then similarly you cannot give the guardrails or the safeguards to bring it >> and then latency the hardest. I think here fuse models are winning where you can make it very quick. You can make it like respond and roughly 300 millisecond response. Uh but you sacrifice on the reliability piece and what we've seen for for for for like our customers which is the businesses you don't want that you want the reliability over overlays.

However, other use cases um let's say like a companion use case that we are not solving in that equation maybe that's slightly different. >> Right. I see. >> Yeah. And that's that's how I would figure about that stack. Um and maybe like last piece maybe in the next few years the way we think this will transform is if you are interacting with let's let's take that booking airline um example maybe if you're just trying to get information about what are the products and services and what where can I travel like where it doesn't have to execute those actions maybe for that part of the interaction there's a version of fused approach but then the moment you need to go into account and authenticated you go to cascad it so on our side what we are exploring internally is trying the both approaches um from research of of of of continually doing doing that and the reason I mentioned we swapped we like almost the good set of of examples in the field there's different companies that will explore both of those approaches so there's very unclear whether the emotionality is fixable with cascaded approaches uh but but a lot of the the recent innovation whether that's great work from suzami or other companies prove that it it is and and hopefully the recent release that we did brings it to at a another level.

>> So, I I want to pause for two seconds to kind of uh sort of overlay something and highlight something Matty may not have realized he did, but did you notice that twice in his last, you know, two three minutes of exposition, he gave a shout out to multiple other teams, right? Including Sesame, Sesame. How many people have heard of Sesame? Okay, it's about half. You know you'll we'll have Brendan who's the CEO of Sesame and he was a former CEO of Oculus. How many people have heard of Oculus? Okay. So almost everybody Yeah. So so this is an important sort of cultural leadership point that I learned slowly over time which is there's two types of leaders in the systems world right there's those who see what they're doing as one part of a greater sort of collective collaborative project as an industry. Right? We're at the frontier of this new space called voice AI and audio AI. Matt's got one point of view on how that system should be developed for the customers he cares about the mission of 11. But at the same time, he's happy to collaborate with other people have different perspectives, right? And one of those people teams was Sesame. And I think it was maybe three years ago um shortly after invested in in um in 11 that I decided you know Brendan Ankit and I Ankit was my former co-founder and CTO of Ubiquity 6. We had been observing at Discord that there was this really urgent need to have better and

better voice models in Discord. You know you spent like 60% of your day in in Discord and voice. And so I remember calling Maddie up for advice and saying, "Hey, we're thinking about building a new kind of product that has AI, sort of a real-time voice companion built into it. What do you think?" Right? And he could have said, "Anre, I don't have time. I've got a thousand other things going on. Not something I think about."

But instead, he took the time to really break down for Brendan, for myself, for, you know, for Ankit, like what what his perspective would be for Sesame's needs. and and he had the maturity to say, you know, even though he was a had to fund raise a bunch of money and there was all lots of competitors and so on to say, you know what, we're all in this together. You're not really a competitor. In fact, maybe at some point we could help each other out. And so, I think actually around that time, Brendan ended up in angel investing in 11 as well. You subsequently became an angel in in Sesame. There's a level of collaboration between now the two teams that is quite rare, I would say, in the ecos. I I wish more teams had that approach that Maddie took and Brendan took. Um actually I think about a year ago based on some of the insights that Maddie had given and Peter had given um 11 uh sorry Sesame they open sourced a speech model called CSM uh a conversation speech model which is just different from any of the models that Lean was working on and uh you know that was awesome because that that model is now on GitHub many of you can use it for your own projects in fact some of the students in last year's project used it and that I think is the through line one of the through lines I want you to take away from this class too. You know, you had you heard Andrew's life scaling lessons number one last time. I would say a scaling law from my experience working with Maddie has been, you know, you can go further together, especially in a new space like this where often what seems like a competitive project just because the VCs or the business ecosystem is trying to create some nice like landscape slide that says here's audio AI and here are the logos and here's like you know visual AI. I mean these these categories and labels are largely artificial constructs, you know, created by nontechnical people to try to make progress legible to the world. But I I I what I find is really it's people that are driving a lot of this progress and collaborations between people is what drives the frontier. And so I know I'm hyping you up a little bit too much maybe, but thank you.

>> No, no, no. This is great. Keep going. You take as much of it as you can. >> No, that's that's that's very true. It's it's very easy to get uh stuck in this loop of like you know like other startups are your competition and ultimately it like does not matter for the mission you are solving. It's a long-term game and and and many of those people will come through different intersections of your path in the future. So so keeping that partnership uh or or working exchanging ideas together is is crucial and if you are to pick competition if you are to start a company you probably want it to be one of the hyperscalers or legacy companies.

It also like will pull you up a little bit. So yeah, I think it's I think it's >> okay. So on that point, let's talk a little bit for a few minutes about the business because proof point that you can actually succeed being a nice guy like Maddie. >> So when we first met, I think the company had less than a million or two million ARR today. Is it public? What what is it public about where you guys are? We crossed 2025 at uh 330 million in revenue and this quarter was our our biggest quarter and we added 100 over 100 million in additional ARR. Um so now we are over over 430. Um >> over 430 million in revenue in 36 months. Can we just like pause for a second and acknowledge that? [applause] >> This is insane. How big is the team? you know the uh it's insane of course and and it's crazy that we get a chance to be building and and all of you are building or or developing and learning at like the frontier of the biggest change uh in the in the world uh uh uh where it's like maybe bigger change than

internet or electricity uh the team is uh although the in perspective of course Antropic as maybe many of you have seen >> oh they were they had a little bit of a head start I would say >> they they added 11 million 11 billion in additional ARR are over last month.

>> Yes. Which is >> crazy. >> Well, they're the outer years, but we'll get there. >> But uh the the the common thread is like the there's such a clear value that you can provide to to your customers across in their case knowledge work in our case how you transform how can any business can interact with their customers and um we are 400 over 450 people now. So relatively onetoone. Um we we have um maybe just to give you like a quick perspective, we have uh a good amount of people between US and Europe. So our biggest bases are in in London, then New York and then Warso and SF are fighting for the third spot. Um we are all the time actively hiring. I need to say this and um we I think that the the the big piece that maybe was slightly different in how we've set up a lot of that that 400 uh people is we keep it in extremely small team. So each team is less than 10 people have a a big uh ownership and mandate to run ahead make their own independent decisions. It's okay to be wrong. the speed in understanding the customer, understanding the problem is much more important than going through this like a carrier process and and that helped us across that journey so often and so much that's why we can do across so many different models and then bring them to customers whether that's in the marketing department or whether that's in the um helping on the customer support side or whether it's growing revenue and helping company figure out how to uh reinvent that with new version of sales or fun engagement. So kind of the common thread across all of the product work we do is how you can really iterate >> right >> in a different way.

>> Well, so uh and this is my last question because I I think um you know the the final project is the oneperson frontier team, one person Frontier Lab, right? The idea is they could now using many of the tools available to them today that weren't around four years ago, including 11, can push the frontier or create a state-of-the-art system that uh would have previously needed many tons of people. And on the business side, what what is it that has allowed you with um such a small group, relatively speaking, to create a repeatable engine where new capability, you know, results in repeatable revenue? I think we talked about anthropic last time where in in that case you know uh revenue scales quite predictably with compute in your case what I've observed is revenue has scaled quite predictably with deployment you know as you built out the deployment team I would love for you to talk about that for a second what what's going on there why why is why have you somehow been able to tell people I'm going to do this much in revenue and just hit it year-over-year with extraordinary growth which is quite rare um and two pricing and let's talk about pricing packaging for a second because these numbers like ARR were again the the accounting of these things was developed in a preAI world for software that look quite different >> now people are going should we do token based pricing annual subscription what is the right way to meter intelligence right and meter capabilities of the kind so can you talk for a second about those two things >> 100% yeah the there's uh and I and I like the ambition of building effectively your final project and And it sounds like a fun one too, especially now you can just do so so so much if you if you have the agency and keenness to just to just go after it. Um I think the two on the two questions. So on the predictability side uh it's still hard.

I it's like actually very I think we you know we we are uh um happily trying to set a target. We are currently overshooting our targets. Um but the main part that is contributing to us to figure out where roughly this will

place is is the deployment side of uh we get a chance to work with some of the iconic and the biggest companies in the in the world. We combined a lot of the four deployed engineering to work alongside them, which is the team that effectively figure out how to take the AI and and kind of bring it into the applied AI side, be the lab version of a lot of those companies and transform their work together. And um and in those cases, you know, it's it's um we roughly know how much value we can deliver each year. And then that really the main bottleneck is can you bring incredible people that are passionate that have that um that level of IQ and EQ are striving for excellence but stay humble like how do you bring those people that are really keen to innovate and and and and be incredible at their craft. Um so if you if you are able to predict the growth of of of the of the people of the value of how much you can deliver then you can start getting back to the predictable part of the business. That's more on our sales side, enterprise side, which is more than 50% of the revenue. And then the 50% is still uh PLG and and continue growing on the self-s serve side. And this one is a little bit harder to predict because it's effectively a big part of can we continue our stride and innovation across the models. And we took the approach um of all the product all the research we launch is available to the to the biggest companies in the world, but we try to make it available for everyone. So if you are building your one uh person project in the future, you have access to a lot of the superpower that the biggest companies will. There is some concurrency limits and of course the the the the closer compliance elements, but ultimately you get the same uh capacity as the as the top lab. Um but what this means for uh frequently for revenue it's that part is less predictable. Um however we know where our initi initiatives are lying and roughly can can can estimate like what's the value of the world and happily frequently we see the value for the world is bigger than than we expected. Um that's the the first one on pricing always think about the value you deliver to the customer and work backwards from there. Never from the cost of how much it runs to do something. If you deliver the value, you roughly want to capture onetenth of what you delivered as a value in your packaging and any pricing and packaging that accomplishes that which is the hard piece of like how you calculate the value, how you drive that good uh metric for that is is the problem frequently more than anything else. So never start from the cost start from the value and work backwards from there. Thank you for the question. The big question is uh voices uh with a lot of the technological advancement you can replicate voices relatively easily. What this means for the future of the of the security and safety in the space. So I think there are two parts. The one we as a company uh given we we develop our own models. We can actually build and and bake in a lot of the safety inside of the models and that's something that we've done from from from the start.

Whether that's being able to like trace back the content that's generated back to who generated it and take action as needed. Two, moderate whether you are abusing and doing a fraud or a scam and stop it before it's generated or or flag it internally. And then three, contribute to the wider space which is where we think it's headed. It's like can you have a publicly available system where you can uh drop an audio sample and get information is it AI generated or not have a watermark piece and that applies as much to the negative version of the use case but also on the on the positive side if people will want to license their voice how do you make sure that this was licensed in the right way which we you know we work with people like sir Michael Kane or Mafi Maki and you want to have that information in there. two to your second part of the question like a lot of systems today will rely on voice authentication in the banking systems and other parts. We think this is not the future and you should step away from this and not use that as an authentication side. We think that's uh from the security perspective it's the it's the it's the it's the wrong approach. Um and every so often we see an interesting way of using voice agents against abusers too. We had this amazing charity we worked with who

based on IP of a caller could detect whether there can be a likely scammer and if it was a scammer so kind of the opposite of the usual they would serve it a voice agent and the voice agent would then speak with the scammer and the whole intention was to waste their time and some of the most fun conversations happened from that from that part. So there are some ways of use that against >> counter offensive counter offensive love it troll the trollers. So the question biggest bottlenecks for 11 laps and maybe for wider audio space you know beyond the obvious ones which is which is um incredible people incredible research so they kind of continue innovating on the architecture side um and and apart from compute like the main thing from the research we would love to fix this year is how you combine so we spoke a little bit about this cascaded architecture of how you create the speech to text and text to speech how you can tr make it truly interactive um where where where it understands your emotion and understands can pull up any of the knowledge from your systems and become that personalized extension of yourself. So maybe the nonobvious one it's um you know every every time you interact with different service you interact with different experience you will have your own preference of of what's uh what's good for you and what's good for everybody else. Uh and we are trying now to figure out how to make that interaction like really custom, really personalized and really good for all those cases. Uh and just to recap the obvious one so hiring is just the architecture side compute if we had more compute you can run more expense more models there's a version where maybe too much compute is also harmful like the necessity is the matter of invention is also helpful in some cases >> the optimal amount of comput >> the optimal amount of compute um and the preference piece of what really works is is going to be helpful and like maybe a a good example of that in healthcare space if you are deploying an agent that is taking an appointment but then follows up with the person asks about how they are feeling. Um maybe recommends uh additional points. You need to transcribe very different nomenclature and make it perfect. Um there's also different people on how they will communicate. Some people want a a slower delivery. Some people want a faster speed. So how you cut there and make that experience unique and different and deliver the most value is is uh is not so much of a bottleneck but more like thing that we want to solve and we know we need to more collect more data, collect what works and then apply to the industry. about the com the difference um and and the other complications um on the training side between the cascaded approach uh and the fuse approach. So you know when you work on the cascaded side of course you uh you get to train the models independently to a large extent um uh but then you need to figure out how they work in tandem together after that. So you spend uh kind of quite a bit of time on on trying to figure out what will the interaction look like when I combine them together and when we actually combine them together you might need to fine-tune and tweak some of those experiences. So we spoke about how you finally got the controllability and emotionality of that work. The pipeline for passing off on like transcribing the sentiment is it a a stressed speech and then passing that as a parameter and making that delivery emotional.

You need to bake that in as you're training the model before you combine that in the in the in the in the pipeline. So you need to bake that in in the training step before before you bring that across and maybe there is a version of how you do the language control or how you specify the um pronunciation in different setups which you all want to make sure it's it's clear. Um on the fused approach there's more of the emergent behavior of course.

So, so you get that for free to some extent. But the main thing you need to do in the training side, which is the usual complexity, you need a good open-source intelligent um model for intelligence and LM that you can bake

into that model. Um, and then there are two complexities. One is how you fuse the tokens from the text space to the audio tokens. Super hard. Uh, and most people cannot figure that step out. And then the second part even if you do figure that out you do rely on the open open models that are out there in the field and today at least the the open source will lag behind the closed source and the intelligence space. The question was around uh the future of 11 labs and what what's what's going how is it going to look like in the five years between the models between the platform between the the the application side of our work. Um at the core five years is an interesting timeline because there's so much research that we know will need to happen in two three years and we think there will be still still improvements you can make beyond that but at the core we know we want to continue leading the space in all foundational research around audio. Um so whether that's the conversational models in uh in the future, whether that's how you fuse it with other modalities uh is going to be uh one of the big parlings that we want to be the best at. So like truly be able to pass that voice during test and any conversation um and potentially extend it to any interaction. So on the research side, what this means for the tech space already to potentially visual avatar space, we want to be that leading frontier uh uh uh continuously specifically figuring out the conversational or in the interactive side. So step beyond voice in that side.

Um the platform it's effectively where we see the the biggest value we can provide in that fiveyear time frame. So as the models start becoming more incremental, you need to really understand in depth what the business what the what the what the creator what the developer is trying to solve and give them the tools to to do that. And for us it's going to be the the effectively the go-to platform. And in some ways we imagine in the future the same time the same way you have three or four clouds that serve all the compute needs on the on the cloud stack. We imagine there would be likely three to five platforms that's that help on that conversational setup between any business and their audience. And we want to be one of those those platforms whether that's in the in the in the conversational setup and the support and sales and in in in in that um in that part of the business or whether that's the marketing and how you engage your customers on on on on being able to deliver the better story all the way through to the internal of how you hire people to scale and train them to then uh to then let them uh continue get that expertise. So if we can be that for the for for the businesses and in some way everybody will be that maybe one person business we we we we will do that and maybe last part on your question because that's definitely something that's back in our mind. It's where does the platform and applications start and end.

We think this will blur in the future of AI where you will be able to create those applications on the platform uh much easier than you were in the past and and we help to provide all the different modules for the builders of the future to be able to build to build that seamlessly. One of the proudest work we do at 11 Labs is actually working with people that lost their voice and we can bring it back. So people with ALS or fraud cancer uh and and so far we've been able to work with almost 10,000 people that lost it and we could synthesize it back so they could communicate uh um naturally uh which is uh which is we hope like something that can be available for anyone anyone in in need. Um and similarly on the question around around Ukraine. So maybe to give you a quick context what happened there.

So of course in that when the war started the government need to figure out how to provide a lot of the usual services to people everywhere around the country and as you can imagine all of the people will just not have the usual access points. They will not be able to go to to a local administrative office to get help on um on you know

where can you travel, what benefits can I have because something got displaced. how can I get additional support for uh for food or benefits? So, it's like a lot of different problems that you need to face and that stretches across the entire economy. In education, the same thing.

People cannot go to a school in the usual way. How do you deliver that education to the people that that need it? How does the government send messages around the country efficiently if you don't have that access in the same way? How does it send it outside of the country so people know what's happening in Ukraine? So the way one of the many initiatives they've done uh was creating a central citizen app called DIA where every person can access a lot of those services going directly through their mobile device and and and gets access to information what's happening uh some of the educational courses or or some of the guidance and um and one thing that that was missing in that equation is how can you open up even further where people uh that might not have that technical acument to to to be able to get that.

How can people u that might not have the internet call a number and get that information um in an easier in an easier way. Um so we worked with them effectively on adding that voice side of um uh inside of the app and outside of the app so people can make it a lot easier. So he traveled to to to to KF at the time to to work and understand those problems and and and work with different set of their ministries on on that work.

And you know one very clear thing that came out was an an incredible model of how they were running it. So every ministry had their own technical resources to try to innovate and bring that across not go through the red tape go independently go quickly and bring that across. And so far we've seen an incredible way of people being able to engage with the government and something we think might be a version of the future of government services. How incredible if everybody here could could could um could open an app and have access to your to your passport to your driving license all the way through to accessing uh the best educational courses uh uh maybe like the Coachella course uh directly from the app too. Um now to your second part of the question which is important on like how we think about approaching our work with um with governments in the in the war zone.

Ultimately as a company we are choosing to be western allied and and um western uh uh um uh countries plus their allies and support their work in in in a way that's uh of course following the the the legal the legal uh guidances and something that that we'll continue continue to do. Can we can we talk for a second about since we're in the zone of sort of sovereign scale deployments? Um how are you thinking about China?

>> Are there distillation attacks you're seeing? Um you know there's a bunch of news this week about how uh a number of nation states have been trying to run distillation attacks on western companies that produce models like 11. Um is that a concern? How do you reason about the ecosystem in China which historically has been very collaborative with the western ecosystem but there's tensions on various fronts. How how should they think about that topic? >> Yeah, I I think in general there's a clear race in AI of you know uh the western world and the and the developments on the on the China front.

Uh from our side we we try to stop all types of distillation attacks uh period but of course uh with with any any

any of the kind of IP coming from that region we we are taking that as additional strong signal of how we can build that into the the protection. Uh the truth is there's and I'm thinking about some of the other questions that were asked that uh that there's a lot of great models coming from that region too and and in audio voice especially as you need to optimize for that language layer too. So you need to you need to really get that nuance of the dialect of the accent of the different voices. Uh there will be models in the region that will be better than 11 laps models um for their their use case in the remmit. um uh and and to a large extent um we try to out compete them on that on that level and provide a better service to to their companies on the other side.

>> And you know as it relates to the open ecosystem because one of the things as we've talked about is having a good open base model allows you to then customize it for various enterprises for various specific regions deployments. What we're starting to see at least in some of the other modalities coming out of China like video models a year ago lots and lots of innovation lots of open video models now not so much as they it seems like as several labs start to catch up to the frontier you know seed dance is a great example of a soda video model that came out it's not open source uh or it's not open weight is that a trend that you think is going to continue what does that imply for a team like 11 and what do you wish more participants in the ecosystem like the students here other labs in the space would be thinking more about or doing that would collectively help?

>> Yeah, great question. I I think so there are two parts uh like as we look at the models in in that space frequently we take slightly different approach in general because there's the the research element we spoke about the product element and there's the ecosystem of how you become a trusted brand that people can can work with. Um, we briefly skimmed across one of those examples of like how you can create and give ability for people to create their voice, share that voice and earn possibly on that voice and figure out how people can participate in that model innovation too.

>> And that's like a stark different approach between like some of the work we do on that front versus some of the the the the wider models that the uh players in China will take. And the same will apply to video of course you probably seen the the work on like how people approach the IP from Disney or Netflix on that side and how they approach it on on the on the western side. So I think that's a big dichotomy and I wish they didn't and I think um on the security level question. And it's going to be a like almost a big combination of the work that everybody needs to do where you uh you you can contribute your work but then you need to figure out the watermarking system and you need to figure out how to have the the models coming from China follow that paring right or at least if not then you don't serve the content or don't give the same content um uh permissioning on some of the platforms.

So there's a I think the two different parts one is how you think about the wider ecosystem participating in the model. to how you bring um uh bring the safety parameters around >> ah I see >> around around around around the models in in tandem and free uh in general of course very helpful that open source ecosystem continues and I think you're doing phenomenal work to help nurture that across the the companies out there where um where ultimately you will have always a incredible set of builders that need the access to the weights whether it's to fuse the models whether it's to fine-tune the models for different cases. Um, and I hope our open source our meaning western opensource models are at least at power better than the the ones coming from there. Um,

which which I think we have a chance to do.

>> I I hope so. >> Some great speakers coming through here as well from Australian back labs that will speak about this effectively. Why the studios not adopt um still uh like why they're hesitant of of switching to a lot of the AI voiceovers and is it is it because of the fear of AI slop or the backlash? itself. In general, as we think about a lot of the creative tooling that we provide for for studios or creators, uh we heavily believe in this concept where um you want to use the tools effectively middle to middle rather than end to end. Meaning you will have the story you want to tell. Then you will you lose use the tool to create a narration in this case. Then you'll want to refine it then maybe recreate and then you can get a great output at the end. So there's this this u big iterative step that needs to happen and why we think about it is like this kind of m middle to middle AI part that it will replace. uh as I think about AI slop version of that I think about the kind of the end to end version like can I type a prompt and get a voice over or video or any any any of that work and that doesn't carry it doesn't have either the initial input of the story or doesn't have this great it inter iterative spirit of that and I think studios realize that so so I think studios realize that but at the same time you'll have very different set of studios and I think you know the the high end [snorts] of the version of Hollywood version is just getting there where the quality became good enough where you could go through those iterations and finally get what you intended. So to make it more specific um until very recently you were in the speech side at least you would give a model a text you would rely on the model to read out the text in the way the model uh thought is best and you could regenerate it. It would understand the context but ultimately it would be down to the model to decide how to narrate it.

until six months ago roughly we finally figured out how to control it like the director would do in those studios where you tell it redeliver this in a slightly more dramatic way while slowing down a little bit. That was a big breakthrough and bottleneck for that to happen and and in the last six months as that started happening we started seeing more studios finally adopt that technology in their work to to bring that. Of course too, there's definitely understanding from the studios that if you are to bring that technology, you need to figure out how the wider economic model will will work. Like you want to respect the IP of the people you work with. Um, and the economics of that haven't been figured out like how much do you charge for AI voice over a person that would not otherwise go to the studio. It's um so I think studios from our conversations and and ourselves are trying to figure out what's the right balance in those economics. Um but frequently what we see like actually happening is you will have AI replace parts of the work that uh that uh that you wouldn't want to do in the first place. So frequently in a studio you have a scratch work that you read and want to listen to. Post-production almost where it started where you repair the lines and then ultimately the actual delivery you want the the additional art coming coming out in the in the top end.

Um and then of course there will be like other things that they will start with which is AI localization or interactive experience of how you can bring experience and and and and and interact with the with the movie. So that's where we see like most of the the use case and as we think about the future that the two pieces that will be stopping it is figuring out the economics model that's backlash related if you don't figure it out in the right way and then two um those step changes that were impossible to make that really high quality versions are just becoming possible uh but I think there still needs more needs to happen as you think like high end of that of that content >> so will the models be on device and and and how we think about 11 apps future in that context

By the way, when does the uh when do we put it on a stream?

>> Uh I think tomorrow. >> Okay, tomorrow. So quick. Um I guess I can still say it, but we finally figured out how to bring our models on device. So we will we will um so we we we found a way to constrain it to a given language and and potentially bring it to any device out there which will be selectively working with with bringing that and opening it up to to wider side of the audiences which which is the big innovation there. Um but it's still the quality there's the definitely quality difference between the on device version and of course on cloud version and the wider set of things you want to do. So the ondevice version will do text to speech but you still won't have the wider transcription interactivity how you transfer the emotions from one side to the other um how you make it uh uh with additional kind of reliability elements built built in. So there's a gap definitely that will will will exist between those for for a long while and you still haven't fixed them on that side. Um and maybe as a quick side note here our approach in general as we think about on device was we need to fix quality first. We want to make it as good as possible only when we do that we'll consider bringing that on device or on prem. So like instead of trying to go on device with lower quality we want to deliver the best experience to everybody there in that context and I think that's like starting to happen but still not happening in everywhere. We think it happens on the texttospech narration less so on the interaction. uh the role of 11 labs in that future will continue being the the platform of how you really go deeper in the problem that the customer is facing or the enterprise is facing and delivering not only the models but delivering the wider tooling that's required for you to bring that technology in the field. So how you bring that applied AI to your local sorry to your to your exactly your custom context work. So and you know if you are deploying that in a um in a in a any any interactive support or sales context you will need the entire knowledge base of how you want the the interaction to work with your customer.

You want the piping of how it calls the phone number or it goes through chat or WhatsApp or email. Um you'll want additional tool calling to get and pull data from the database um from Salesforce or service now to to deliver that experience. And then the whole framework of how you evaluate, monitor, test that is essential. you it's like is it working? Can I self-improve based on that? Um what is the domain tests that I bring? So uh we spoke about the the the the route scheduling for airlines. How do I know that there is additional speed seats available in that route? You need the the the test for always checking for that parameter being true and if you left customer happy. So in that future where models become more available, we know there's a still a huge gap of what the models can do and what you need to bring that in that um enterprise context or even creator context for question earlier earlier on. We are out of time, but thank you Maddie. Hey.