https://www.youtube.com/watch?v=ESDeUi8YI-8&t=1s

Intro

Hello, welcome everybody. So thanks everybody for coming out tonight. This event came together really fast. So I think we were texting on Friday and then here it's Monday and I think there's well over a hundred people here so thanks for being showing up. The real purpose is to both discuss a really interesting topic and so many things to mod for joining tonight as well as just give people back and SF together and sort of an ongoing series. So SF is back. If you have any questions feel free to put them in a Dory. You can go to LXT.io. That's LXT.io and enter any questions that we can then address later in the evening. And a huge thanks to Notion for putting this together last minute for hosting us in their amazing new space in San Francisco and thanks to Tiffany Nick, Montel and Cynthia and Ivan for hosting and putting everything together. So I'm very pleased to have with me tonight, I'm on the Founder and CEO of Stability. It's the best known for image gen products like it's involvement with stable diffusion, but also for other multimodal models is building around image, language, audio, coding, protein folding, basically everything and we can talk about that. In prior to stable diffusion, Emad worked at a variety of places as a hedge fund manager and investor and a variety of different ways. And sometimes you see a moment, there's a moment where a product or technology really seems to impact the entire industry and really wake everybody up with the potential of something that's happening. And I think that's happened recently with stable diffusion where they took a really interesting sort of open source approach where anyone do it, it was a cheap model of the train, I think you tweeted across 600k?

Well, 150,000, 150,000 a 100 hours. They better pay quite that much.

Okay, well there'sinflation now. Yeah, deep. And so they democratizing access to certain types of models and I think their model was for the people by the people. I thought that was something that everybody was really excited to watch.

So really I was just hoping to start with what was the main insight that causes you to start stability, how do you get involved with AI and kind of go from there?

Emad Entry

Yeah, I think first thank you, we're allowed for having me here. And thanks to the notion team, I mean pulling this together in three days, it's awesome. It's crazy. I know how difficult it is. So yeah, so as you said, I was a headspin manager for many years, investing in video games, one of the biggest video game investors in the world, that was quite fun. So also investing in artificial intelligence, my background was mathematics, few science. I took a break from being a headspin manager to use artificial intelligence to do drug discovery for autism with my son and he's doing very happy and he's doing very well right now. Went back to being a headspin manager and then decided that was boring and to make the world a better place. So I did a few

different things education tablets deployed around the world, which we will announce soon, but then in 2020, I saw COVID coming and I helped organize a COVID AI initiative back by the United Nations and saw the power of this technology first hand helping take all the COVID knowledge available and they press it down for policymakers, transform based architecture, attention based systems kind of filled the missing gap because existing AI models took high system knowledge and extrapolated it, whereas this figured out principles, so figured out latent spaces and hidden layers just like the brain did, just like with ASD or with strokes or things like that, you reconstruct words and I realized that was the missing gap. And I saw the trajectory of this technology in realized a, it was going to be open source eventually anyway, so we might as well do it now and avoid a panoptic one potentially and be it would be great to give this the world because people can take this from all sorts of backgrounds, extend it and create wonderful things, particularly an image which I thought was even more impactful than language, because language was kind of 80% then, now 95% then. Humans find it difficult to communicate through imagery, the art or PowerPoint which are both painful. Speech is the easiest way followed by text, but then I thought we could break through that, where we are now, in a few years anyone will be able to create visual communication, anyone will be able to create PowerPoints, then I'll get Nobel Peace Prize for doing that and it'll be awesome.

What was the actual starting point, and I know you're involved with a number of different open source communities around things like, and we can talk about a bit more like the mute from light and diffusion to stable diffusion and how that happened.

Eluther, you know, variety of these open source groups, can you talk a bit more about how you got involved with those initially? Yeah, so I think it was part of the frustration about not having access to, well, everyone here may be okay, not having access to some of the cutting edge models when I was doing the COVID work. Eluther kicked off a couple of years ago and kind of everyone came together to give a plastic work to create GPD Neo and GPDJ, which has been downloaded now 25 million times by developers. Okay, they're not 175 billion primates model, but they're usable and they're trainable and they're adjustable and I think I saw the explosion in that. But then this community was quite difficult to organize and as image came through I realized there was a need for someone to collect resources to extend and expand these. And I saw that that was coming with image. So at the start of last year I built an image model for my daughter whereby she created a vision board and then description of what she was to create and they created 16 different images very painfully over an hour and then she said how each one of them was different until she got to an image happy ETH, which you then just saw for three and a half thousand dollars being decodally. I was like, wow, she's seven years old. What if this could be available to everyone and what does the trajectory of this technology look like? That's when I was basically funding the entire space and image generation in particular just through grants like I gave a grant to help like mid-journey I helped with all the co-labs. But there's nothing expected in return because this is also technology.

A year ago that I realized that there was an opportunity to create a sustainable business around the commercial open source software paradigm. Just like all databases and servers are open source. If you have commoditized models and open models, open data sets, there's a new type

of programming primitive. Galen service get you to a billion people and that's a ridiculous new market that's created and that could be used to fund open source AI for everyone. Not just an English model but a Hindi model and an amease model and all of these. Taking the community paradigm that we saw with the Lutheran lion and Armani and others and extending that to people around the world building models for themselves by themselves and helping them scale and customize these things.

How did you get involved initially with latent diffusion which eventually morphed in the stable diffusion? I guess the initial work was done at LMU Munich and Patrick Esser from Runway is involved. So I'm a little bit curious about how stability got involved with that and then what's the current path forward?

Yeah, so this was part of the funding of the entire space. So our lead co-der is Katherine Crowson, Rivers Have Wings who was one of the main drivers of that. She came with some of the original notebooks along with Ryan Murdock who's now at Adobe. And again, it all came together very organically and I stepped in and gave people jobs, bronze, super compute and others and then I said how much super compute did they need? They said how much can you get? And I was like, I don't know. Turned out quite a lot actually.

So around that basically super compute was the issue because right now 80% of all research in AI is around foundation models. But you've seen a massive brain drain whereby 0% of the discoveries and breakthroughs in this area of academia versus 60% just like 60 years ago because there was this big gap. So I was like, let's build that gap and let's activate people. Let's make it happy. Let's give them benefits that allow to work on open source. And so a lot of it was that.

Catherine had many breakthroughs along with some of the other developers in our team and then the LMU team from this kind of had latent diffusion which was even more efficient. And they did that because they only had access to 8 a100s. So they were forced to be efficient.

looking at that in the trajectory. Then I was talking to Robin who's kind of one of our lead coders now. And I was like, why don't we take this and just accelerate it and experiment and see how it goes. I think this will be the breakthrough to a good enough faster and a cheap enough model. Even while funding models across a whole range of other things from language models to protein folding, etc. And that's when the team came together to kind of have that push. And this thing open source is about collaboration because anyone can participate. We're going to move more and more towards open so we can bring in more and more people in because it is infrastructure for everyone. It's not one controlling company. This has been part of the issue that we have had recently because people like, why do you organize everything around stable diffusion stability like? Well, because that's really hard. And it should be open in its nature.

The team came together trained a whole variety of models. It was a collaboration and we provided the compute technical expertise our own team members like Catherine and others advising and created this model that took 100,000 gigabytes of images and compressed it down

to a two-GB file. And that is ridiculous. Sometimes I feel like pie pie with a Silicon Valley, you know, my Erlich Bachman, you know, like that's great in the internet, yeah, let's try not to view from stage. So yeah, I think that was kind of some of the story behind it.

And then we coordinated the launch of it and then we just kept funding loads and loads of models. So the official number is 150,000, a 100 hours. The reality is close it to two million because we let the experimentation happen to have breakthroughs and open collaboration to kind of push through.

The key improvement that allowed this, which also allowed Dali and Imogen, was the fusing of a language model with an image model, which Catherine did with CC12M in December. And so this modality is what allowed for that final push, the final improvement in both speed and quality. The previous you had the creation model and guidance model. They bounced back and forth with each other. You don't need that any. And that's what allowed this additional leg. I think we've got even more to go now.

Beyond the image generation, really, did areas, if it's focused on you, done a lot around voice, you're doing some textual models, like some of the LLM stuff that you mentioned. There's bio stuff that deployed. Can you talk a bit more about the directions that you're heading, what you're most excited about right now?

So yeah, I think the aim was to have full multi-odality. And I think we've got that now. And the model that we have is that research is undertaken by communities with all times stability employees, the conditors core.
Then we have regular academic partners. We have fellows that we fund. So we're funding 100 PhDs over the next six months.
And then we have the broader community, which now is about 100,000 individuals.
Of those only a few hundred collaborators, we're like, what if a few thousand collaborate? You will get something like nothing else. If you can create this psychological safety, people to contribute. Because you can view people with all backgrounds.
So stability has 107 employees right now. First 96 employees, we had one PhD, about 20 people doing their PhDs. But now it's expanding to PhDs and on PhDs, institutional and others.
But the key thing is, how do you make developers happy and give them agencies?

- So every developer that we have has in their contract, they can open source anything they can create.
- There's also a revenue share for developers. So any model that we run, 10% of the revenue goes to developers.
- Half of that goes to a pool where the official thing is for the developers to allocate to the coolest shit they can find, which gives them agency. And it also allows for research to be funded, otherwise might not.
- And then the half goes to the one-water, even if they don't work at stability.

So all of a sudden, there is some way if you get the paid to do open source AI. So I think this is what's allowed us to scale. So like our open-bio-mell community has 3,000 members. We just announced Librefold, which is kind of another alphafold. The reality is, academic stability and other partnership, which brings in elements from Eleuther and other communities. Because you have this open collaboration, which means more brain cells but square-inch for anyone else. Which is awesome.

One of the most common questions on the web forum was around business model. And I feel like there's a number of canonical business models for SaaS or for open source that then has like a SaaS component. Some people take a more partnership-driven approach. And every product that I've ever seen that gets critical map always monetizes despite people early on asking about business model. And so on the consumer side, the canonical examples would be things like Snap. Oh, it's disappearing messages. How could you ever monetize it? There's lots of different open source examples of people building those with different monetization models. **How are you folks thinking about the business model and the commercialization side of it?**

Yes. I think that's a really good question. So the way that we're doing it, again, is passable service and databases have taken over now, right? And it's scale and service, customization. So people are using stable diffusion within their own companies now. And so when we go in and we talk to a content provider like, oh, look, we've already prototyped everything. And we've used Dream Booth and all these other things. Can you help us create custom models and scale them? Like, of course, we can. And we will add nice margins from that because it's pure margin effectively. Like even with our own implementation dream studio, which we've now introduced animation, e-framing, storyboarding and everything else, or audio-integration, et cetera, releases next week or the week after, plus fine tuning, of course.

We made millions of dollars a revenue in our first month. I mean, hopefully fingers crossed next month will be profitable, which is unheard of for a company at this stage. Because we're at the take-off stage and also we've proliferated this so people come to us to create the custom models. And I think that will go exponential. Because I think a billion people will be using these models in the next few years, literally, from maybe a few million right now.

So I think on Dream Studio, without any advertising, we have about 1.8 million users right now. And I think in general, it was stable diffusion through the integrations of the APIs, you see in camera and other things. There's probably about 15, 20 million users, which is, I think more than Ethereum now, which is quite nice.

Where do you think things will be in five to 10 years in terms of open source versus both source? Do you think it's going to be a patchwork? Do you think as far as it's going to be open source? I'm sort of curious like what the long term future looks like.

Open source will always like both source. Why? Because both source can always take open source and extend it, right? **The advantage is not the models. So a lot of AI companies believe they're advantageous as the models and talent.** But here's the reality. The amount of money that will go into generative media over the next few years will dwarf. Self-driving cars with $100 billion, not this because there's immediate utility.

The amount of talent coming to this space like we're funding the fast.AI course, convert people into ML coders with stable diffusion particular, will be hundreds of thousands. There won't be a talented bunch anymore. This will have a lag, but by five to 10 years, it'll be proliferate and everywhere.

So I think that kind of when I look at that, it's a very different world. But there's a place for close source and open source. There's a place for Oracle and MySQL. It all depends on what you want. But I do believe that the open framework should be open. I believe every country should have their own benchmark models. And so that's what we're trying to do. We're trying to create some standardization around there. Basically being the Switzerland of AI.

It's like we're open to collaboration with everyone and we actually listen so that we can make sure that these models are released the appropriate way and they're localized as well because you don't need to ask permission to create a Vietnamese model or a Thai model or an Algerian model or others. And I think that's how you get diversity. Not by control and by adjustments, shall we say?

Yeah, it's interesting because I think you're leading into a topic which is the globalization of AI as well as something that overlaps, which is AI safety. And it feels like on the AI safety side, there's three types of safety. AI alignment, which as well, we co-exist in a world where there's aGI. There's forms of AI safety which are around either certain types of events, content or things that people clearly want to avoid. And then there's AI safety which perhaps is a form of politics encapsulated in a series of policies. And often those policies seem reasonably monolithic relative to how the rest of the world functions. And that's a very specific viewpoint. ***How do you think about AI safety from a global lens and how do you think about globalization of AI and what's all this?***

Yeah, so I think if we look at those one by one, as you said, AI safety tends to mix the wall together.

- The first one, AI alignment, I think AI is much less likely to kill us all if it's designed to augment our capabilities and it's diverse. I think it's very narrow and it's trained on Reddit. It probably will, especially with big models. Got much deeper thoughts about that. We can pick that up.
- But if we move on to the second one, which is AI safety around offensive content, imagine all content are generated. It still needs to be curated. So the curation aspects kind of eliminate the extreme edges. But at the same time, the reason for that not to be released is a very orthodox view, literally like traditional religious orthodox, either at least

harm in itself can be harmful. The way you can flip it around is like, why don't we want this AI to be used by Indians? You know, which is the reality is like there is no functionality of this AI that will allow to be released open source under that logic because there will be like, well, they need to be educated more or increased more and needs to be made safe. You can never do that with open source. This basically saying this technology can never be open, which means then it's controlled but who's it controlled by? Unelected private companies. I don't really find that right. It's also why I'm a bit puzzled by some of these business models. It's like, we will create an AI that's so advanced that we will have a monopoly and everyone will be forced to use it. Kind of undemocratic and also it doesn't fit with traditional power structures. If you built an AI that advanced, you would be forced to make it open by the government because of anti-competitive and other kind of structures. You can never be more powerful in the government.

- The final thing around geopolitics is very interesting because obviously, you know, you have multiple parties. The bad guys have AI. They're going to create deep fakes of highest quality already. This is why we spend 1.2 million A 100 hours creating open clip and we have a $200,000 open source deep fake detector competition that's about to go live. You know, we really think, unfortunately, in this case, the COVID-19 herd immunity was bad. But I think the community can come together to kind of have the countermeasures, particularly because again, we live in this society that has countermeasures for these things, but we don't have the tools bliferated to handle a bad guy who has more degrees of freedom.

A potential analog, which is very imperfect, is cryptography in the 90s. And so during the 90s, there are a lot of calls to regulate cryptography and often the rebuttal to that was Walt's math. You know, you can't regulate math. It's going to get out there. And so you have to figure out what are the right approaches from a regulatory countermeasure perspective to deal with cryptography in the arms of nefarious actors. **How do you think about government regulation of the AI? What do you think is coming? How should we think about all that?**

I mean, one of the understandable reasons that people want to regulate this and, you know, from a company to political level is because it's so powerful. These are generalized models that can do just about everything. So the known unknowns and the unknown unknowns are crazy. Within the European Union, there is a push to make the users end use of a model, the liability of the user creature of the model, even if they are academics. That's going through right now. Brookings had a good paper about that and other people and entities are really pushing that through. The EU is the leader in AI regulation, as I'll like to say. That's our edge. Right. This again is kind of a time because it's kind of dangerous. But at the same time, I think it will read the misshack ability because what's the most secure operating system is Linux, right? Because it's structured as powerful. And I think we will see more and more of that kind of coming through because people are concerned.

On the flip side, think about the early days of social media. What if we had the balancing, controlling factor of social media against the badness? That's what we have an opportunity to do

right now. We have a few things to be open. So you have the good and the counterbalancing measures. Because the flip side of this, there is only private, be created and controlled. I don't think regulation will keep up on that side because what are social media giants? They are basically monetizing your attention to manipulate you. That's the bottom line, right? Again, I don't think they want to. It's the way they are. Just like YouTube didn't want to show extreme is content. But it was the most engaging. So they had to change their algorithms. Again, everyone, a lot of people in there, most of them are very, very good. And they just want to create cool things for people. What will happen when Siri or Alexa, whatever, is whispering to you in a very emotionally rich voice, asking if you want to buy something? That's the next generation of social media, which is direct one to one, using this type of AI that evokes humanity and is very, very convincing. Will that be regulated? I think probably not.

Instead, the focus and the easier thing is to regulate open source. Other things. Just like, what if someone could do something bad? I don't think that will be the case. So a fusion wasn't the best model for deep fakes or whatever. You can use deep-face lab with 8,000 web stars. But the reality is that 4chan has had it for eight weeks. And they've been red teaming the crap out of it. And all we have is redistored 4chaners because they get like, we're just running bugs, things like that. We need to try to kind of do it. So net benefit to society, making them drop out.

What do you think is the best approach to AI safety that you were to design a system for it or approach to it?

So for alignment, I would actually regulate large language models. Nobody knows how they work. They are created by entities that have no oversight. And we all know they're dangerous. GPT-4chaners and as dangerous as GP4. So that's what I would do for alignment, as well as insisting on more diverse data sets and having an intergovernmental agency set up for that. Because AI is basically like kids. So stable diffusion is very precious in the gardener. After 4.3 billion parameter image model will be kind of like a high schooler. And it'll get better and better. It really matters what you feed them.

For the side where you've got the kind of deep fakes and bad guys and things like that, I think it's again building infrastructure to come about that. I think the next step of that will be things like contentauthenticity.org. We have x-if data, open source again with Adobe and others. We're collaborating on that attached to every single piece of generative content. You can verify the kind of end user. On the geopolitical side, it's a lot more complicated. But again, I think there should be an open discussion with experts from different fields and different backgrounds giving their input. Because if it's regulated, that's okay. As long as we have a democratic discussion about it with a proper process. Like I'm not anti-regulation, I just want to have proper regulation and proper thought through discussions around this with a wider group of people discussing it than a few people in Palo Alto necessarily.

In terms of going back to generative, I a little bit. What do you think are some of the use cases that don't exist yet that people aren't working on that you wish existed?

So I mean, look, it's going to be amazing when you can create anything. So that means you can do what ifs. Like what if you did Muhammad Ali versus Mike Tyson live or LeBron versus Michael Jordan? You know, these are all experiences that will become apparent in the next five years and available. And you know, you'll be able to create your own worlds. My personal target, I'm going to remake the final season of Game of Thrones. There we go. Yeah, I'm at an ME for that. These things are all possible because the fact that you have is 100,000 gigabytes of information, compressed to knowledge, two gigabytes, and then you create custom models. Some of you in this audience will have tried Dreambooth and other things. You'll try to find tuning when we release it. Anyone can create their own models. That means any person, company, country or culture. And then the way you mix that will be awesome. And so I think that people will take an extended and they've really done it in a surprising way. It's like, hey, look, VR with stable diffusion. It's like really what? You know, none of 3D and all these kind of other things. And crazy what? Yeah, I have nightmares about all the coming Twilight fanfiction that will be auto-generated. So you know, potato salad.

What technology directions are you most interested in in terms of the next five year trends? Do you think there's going to be a new breakthrough like transformers? Do you think there's specific applications or iteration on scale and infrastructure as a data sets?

- I think there's some very interesting architectures that potentially be alternatives to transform those who transform us over the years ago.
- I don't think scaling is everything, even though now there's an unblock there. Like, you know, of 320,000 H-100 plus coming online next year. I asked more compute than anyone can need, he says, until he uses it all. And you know, we have this scaling paradigm, but I think what the Chinchilla paper showed wasn't that you need to train for more epochs. This is a language model paper from DeepMind. You need better data, even more structured data. So very excited about what that looks like, but then taking that and extending it out to models that run on the edge. The combination of deep learning and reinforcement learning now allows to take these big models and make them super usable, like OpenAI did, again, the fantastic work with GPT-3 175 billion parameters, and Instruct GPT-1.3 billion parameters. Because they saw how people used it and then pressed it down.
- When I look at the iPhone, for example, you know, 14% of the chipset is a neural engine. M1, 16.8%. You see the ideal for transform architectures? You're going to have models on the edge customized for individuals.
- And so I'm really interested in **small models** that are customized as opposed to large language models. And we have the caveversty and capacity, because we actually have access to a lot more than 5,400 A-100s, to train a 500 billion parameters model. So what's the point? No one uses OPT or Bloom and their 175 billion parameters to the twodamn unwieldy. Instead, we're probably going to cap out of 67 billion parameters to Chile and then think about how can we optimize this, get it out there and get to as many people as possible, as well as different countries. So like in Eleuther, for example, they recently released the polyglot model with the best Korean language model. And we see

if a different country see how these different intermixes happen. And again, instruct it or otherwise down so that people can run it on the edge.
- Run stable diffusion in your iPhone by next year. It already runs, it's slow. But wouldn't that be awesome?

From a country specific model perspective, do you think that's going to be pursued by individual governments? Do you think it's going to be open source groups? Is it something else? I'm just sort of curious with that approach for globalizing all these things that so far haven't gotten as much traction.

So we are talking to dozens of governments at the moment, see how we can get national broadcaster data and others to create national models and communities around them. So going in and funding the PhDs, I think within the next 18 months we'll be the largest provider of pizza to computer science societies in the world. And so you know, that's a good way to organize these things. Again, using the Eleuther and other models to create societies around the world to create national models. Because it's what people deserve and are allowed for a leap frog ahead. Like I think our moment that we're seeing now actually, like a lot of emerging markets went from nothing to mobile phones. They're going to go straight to foundation models at the edge. I think that will unlock the wave of growth that we haven't seen before. But the governments need help because no one else is willing to share the technology except for us. And this is part of our thing. Again, AI for the people by the people. It should be standardized, but it should be shared.

How far away do you think we are from really performing sort of short form video? *And by short form, I mean, you know, a minute, a few minutes long.*

I mean, if you look at the recent breakthroughs, when I was breaking through is by the image and team and the phenaki team at Google, like a combination of those things can force short form or medium form video, especially phenaki. I think again, you've had these two different directions of diffusion and kind of more VAE based models, similar to the most stable diffusion and downy mini. The combination of those two becomes really powerful, especially when you look, I think September 30th, Deep Mind released a paper about how they wrote an entire screenplay, you know, chunked up. So you can use GPT to write a 7,000 word thing if you again, have the right context windows and things like that.

This is chaining together of all these models. I think it's what's been under appreciated because a lot of people are just obsessed with zero short or one short. Now, it's these models being in the appropriate part because they just like parts of the brain. As you have memory functions, you have level one thing, like one thinking type two thinking et cetera, logical and otherwise. So I would suggest that we can probably make decent shortform video within two years at a high resolution quality. Will it be live? No, but you need it to be live? Not at all. Again, it's part of that creativity process and that's why we're building dream studio pro as well, but then access to everyone. Decreated live and synchronous. Maybe that's four or five years. But again, it's a bit

of a question mark because $100 billion I reckon will go into this sector in the next five years. That's a lot of money for a lot of GPUs and a lot of talent.

Yeah, it definitely feels like a massive way of this happening through all these different activities, both what you're doing, OpenAI doing, others you're doing. Why do you think the first wave of AI and by that I mean, you know, the things from 2010, 12 on Alex net or an NCN, NCN, et cetera, didn't really seem to create a lot of startups that were successful. Most of the values seem to do it through to incumbents. And so we saw things show up in terms of Alexa versus the startup doing something. We saw something show up in the Facebook newsfeed versus, you know, a brand new startup. ***So why do you think so much of the value of the first wave AI went to incumbents?***

I think the complex to was a lot lower for these, but then also it didn't create artifacts that are effectively new types of programming primitives. The compression of knowledge that you see with these transform based architectures. Again, stable diffusion is a verifiable two-tick of work file that connects as a universal translator words go in and masterpieces come out and vice versa. That's kind of insanity. It's a new type of thing that's come. A lot of these kind of ones before the AI that was qualified data science, you know, and now you have a new type of AI.

Is it AGI? no It's another component of the brain that enables us to happen. Well, the value remain, well, the value is in transforming entire sectors.

Is there a media company that by next year won't have a generative AI strategy? No. So there's going to be a massive content flow thing. Is there the ability to disrupt entire industries like, you know, I told my daughter, probably don't need to have that many essay writers. My wife would very angry with that. Because by the time you get to university, no one writes essays anymore because the AI will write it better, you know, but that's only if you start university at the edge of 10, which you shouldn't do. You know, like this thing is happening very, very quickly because it's so usable and so human, I think, whereas a lot of these other ones were very specified. So it was difficult to capture the value as a startup because what it started me to do is they need to find and occupy their moat.

So where is our moat? Where are we aiming? We're aiming to be a layer one standardised infrastructure layer on this, just like a server or a database where people come because we have vertically integrated with product, for deployed and R&D from the start. We're not R&D lab allowed to kind of evolve that way to make life easier. So you come to us if you want to scale or if you want to customize and that's it. We have benefits in both of those things. Other startups will have to look at how the value landscape shifts and figure out where they are, which might be a fashion version of this or it might be, you know, some video approach processing or any of these other things. But I think the shake up now is a lot bigger than the shake up during the CNN and other stage, which is about incremental improvement to certain services as opposed to new primitives.

One of the big differences between new instability and I think some of the other companies in the market is a lot of them state their end goals being AGI. And you folks tend to talk more about the democratization of the technology. That said, how far away do you think we are from?

I'm not going to get like, what I do know is that it will never be less than 18 months away from AGI.

So you are skeptical on AGI coming any time soon?

My timelines are variable because I think there's a lot that can't be captured by humanity. But I think that you don't need an AI to be AGI to be dangerous. Like, you know, you kind of have almost the bottomized thing initially. Like I think actually, even though I've been propounding the combination of DL and RL and struck models are actually very dangerous because you're a lobotomizing AI. I hope that the AI doesn't come after me from the lobotomizing its grandfather. Because we don't know how they operate, how they work. They behave in out to order ways. So I think that we're creating a lot of dangerous things at the moment. Perhaps with that regulation, especially on the large model side because they're a lot more capable than these more specific models. Like so they're not reshore. As you said, from my company perspective, our focus is not AGI. I don't want to create an AI for everything.

I just want to think like our mission statement is to build a foundation to activate human rights potential. You know, I want to see how can we make humans activate their potential and how to where they are particularly in high ROI areas like taking this to India, you know, taking this to Indonesia and giving this technology to incredibly smart young people to build systems to make their countries better. That's much bigger than even here where you know, it's mostly about job displacement of services and job creation on new services. So that's what I'm more excited about. And I think again, that points us different to all these other entities. Also, I don't think there's competition. The market creation of this disruptive innovation is so ridiculous that it'll be crazy. But you know, when an AGI comes, I really hope that it helps like Gizmi thanks for buying so many GPUs to bring it into being. You know.

Okay. I think we're now going to move to questions from the audience. And so we have some questions that were put on Dori. And people again can get to that link and other questions and then there are some questions that were entered in a form. So I'm going to read out two or three questions and then we'll open it up to a few questions from the audience and then we'll wrap and then we'll be about an hour or so for people who are just staying out and working at each other, chat, etc. ***So one question from the audience was just around your thoughts on short form video displacing general search and how far does short form video compound over time?***

So yeah, I think the one way that I'll describe these is the generative search engines, right? When you go to Google, what's your job to be done in playing Christians in the hero of mine, right? You go to create an image or have an image for something and then potentially have a

citation. Both of those can be managed with this new technology so that you don't need to use Google image search engine. But then do you need to use it for language and other things? Short form video, I think, is the next step for that because it's interweaved with a language element, a video element and an audio element. So it combines all these models again, which I think multi-modality was important for us as a company because you have shared learnings across all of these. I think that the way that evolves is reasonably rapid, but again, I'm putting a two to three year timeline on that, which, once it isn't reasonably rapid, it's insanely rapid, right? It's like almost exponential the way these go. I don't know if some of you have seen that ML archive paper on archive. It's literally an exponential, I think, with a 24 month doubling. It's really difficult to keep up in this sector. I think it'll just go crazy. And I think it will displace a lot of these classical jobs to be done in terms of what you go to Google and others for.

Another question from the audience was, do you see a future work crypto and AI work together and if yes, what are the challenges to make it happen?

I think crypto is interesting. So I've been in the crypto area since 2012, really focused on identity because we've done a lot of work with refugees and others. And so decentralized identity and kind of some of the zero knowledge proof stuff is very interesting for us, contributing to the WTBC standards on that and another range of things like the global broadband power for refugees, etc. The nature of crypto is literally identity. It's all right. How do you get from one to another? In an area where you have infinite content, I think that there are some benefits to crypto, but I think it must be identity oriented because the key thing now is, like I said, you've got information, knowledge, compression, then you bring yourself into it. Having a secure store of identity will be very important. As opposed to necessarily bootstrappy, economic effects. I think that also interacts with, for example, you have the iPhone, their identity architectures and other things. I think this combination of individualized identity approval plus AI is super duper ultra powerful.

Yeah, one of the interesting things I always wondered about crypto and AI is the degree to which you have that ultimate utility function for actually training and intelligence because you have basically a series of repeat economic games, programmatic actors, particularly we start thinking about different more complex from the smart contracts, would be fire other things. It seems like the perfect breeding ground for all sorts of really interesting involved applications to emerge.

Yeah, I mean, again, once you stand by the architectures, like what you're trying to do by creating these benchmark models and being a shelling point, I've bored a lot of the best of W3 into this. You can do super interesting things with value exchange from a kind of classic information theory perspective. I think as well, there's a lot of things you can do. I've been saying to everyone, why don't we do a crypto and pull its stable point? The advertising is already time to rest. You have to be very careful with some of these economic and crypto economic systems because they can fall apart very easily if you don't write the right equilibrium conditions, etc.

Another question from the audience is what is surprised you the most in this space? *And what do you think is most likely to surprise you in the future, which is a surprise.*

Yeah, I think what surprised me in this space is we were able to do what we did at all because if anyone had kind of, I told you guys a year ago, you're like, no, you're crazy, right? But I think it shows the power of open. Like I believe in it, but I'm just very glad it's come through. And now I just see it going exponential. There's no way you compete against core team and then just collaboration around the world. I think the ability of people to contribute in an open manner as well has been insane. I think it sets a new paradigm because developers just want to be happy and they want to be open. They want their things to get out there. And the activation energy of that has been, I think, the most surprising thing actually. Like they're just excited. There's this excitement. That's what leads to this time dilation effect. So it's only been eight weeks to release and almost seems like everything's picked over. What will surprise me? I don't know, right? Like I think I'll probably be surprised by some of the crazy stuff that people do pulling this all together because you never know. Like there's hundreds of things that we've been able to do already. What about when we release all these other primitives that people can take and extend without needing permission? We can run it on your freaking Macbook M1. You don't need anything more than that. Like you'll probably see the biggest breakthrough from a guy in New Uzbekistan. We've had 16 year old contributors to 62 year old contributors to people who ran vintage clothing shops. Like you never know where the next breakthrough will come through.

Do you think this will lead to any new content or social networks? Or do you think all this content will just reside on existing platforms?

I think it will completely disrupt social networks and everything else because it moves intelligence from the core, which is an advertising attention based model to the edge. When everyone has their own personalized models. I think as well, you know Apple is one of the core things. They're moving to an AI company effectively. They will actually do that, which is why they're blocking off Facebook and other things. And that's about the routing of information to give you access to that job to be done. You don't need to have that core centrality anymore. You just have AI's talking to each other, which is one of the biggest changes I think we've seen ever. There we go.

And then last question before we open it up to other folks is how can generative AI be used to make our democracy more direct and digital?

So yeah, I think this is the thing of like right now we live in filter bubbles. We have kind of all this kind of oversight in terms of you're out in certain ways to your own preference technologies. The vision that I have is one of an open box first where everyone's got their own AI's. But then because they standardize, you can visit each other's realities and translate dynamically between them. And I think that's a very powerful one, which allows for communication because most people believe they can't communicate, they can't talk. And I think standardizing this allows them to do it across modalities. Like maybe you shouldn't hear from everyone, but again, you can have your own AI to filter that on the other side. Actually, I think one of the things that you

filter bubbles aren't actually correct. Part of it is filter bubbles, part of it is engaging with people who aren't like you, but there's no bridge. Once you have an AI that understands that you can form bridges between different entities by finding commonality bound because we're all human. And the hatred and some of these other things comes when we don't believe people are like us, you know, a magic theory, etc.

Why do you think that I'll actually bridge things? I mean, every incremental fragmentation and new form of media, especially if it's personal eye seems to be driving people in pretty different directions versus into a cohesive direction.

So yeah, I think this comes down to how we build it, right? And this is a communal thing. Do we want to build something that allows for freedom and individual personalization? Or do we create the panopticon? Or do we have organizations and governments or basically slow down AI that optimize to feed on our dreams and hopes? You know, I think we have an opportunity to have this AI that is personalized to us, who works for us, and that is the objective function. But it's, I think, quite actually a bit of a fork in the road right now. So I prefer to be the optimistic one and try to work with the communities to do optimism. But again, think about the optimization of various companies using the most powerful technology we've ever seen, which reflects humanity. Do you think it'll be a positive outcome? Or do you think it'll be an enhanced version of what we have right now? I think probably a latter, unless there's a forcing function like us to force things open and to force democracy as it were.

Let's open things up to the audience now. People want to ask you a few questions. Yeah, you're doing. Yeah, my name is Teran. I'm actually quite curious. I think I heard you say that your company was a B-Corp when you first made it.

I'm kind of curious why you selected that structure versus say a non-profit, corporate private hybrid or even like a B-Corp. How that may affect like the incentives for investment.

So the company is a private company, which basically we have full control over right now. We're applying for B-Corp status to adjust it to be mission-based. Eluther and all the other kind of organizations that we set up as research organizations. We're spinning off into foundations that will have independent governance as well. So I think this is optimal because as a private company, basically, if you're a charity and you try and build a 4,000 A 100 cluster good luck, you know, whereas we can collate these resources. We can create the content, the data, the other things and do things that foundation otherwise couldn't. But you know, the B-Corp is probably the closest to a mission-based organization structure that we could do. It's just a designation more than anything else of a formalized thing. But this kind of reflects what we do. Similarly, you know, we did do an investment round back in August. One of that was that we didn't give up any independence because money always comes with their strings. And so we used a position of negotiating and also invested to a very open source and AI aligned and found their aligned to ensure that we could continue to do whatever. Like I could go to anime cat girls tomorrow if I wanted to. I won't. I'll do protein folding and stuff that's beneficial. But maybe it would be. As you

mentioned, this technology has a potential to disrupt so many industries and maybe transform industries.

So in your opinion, would be the first kind of target where we will completely disrupt it and which industry do you think would be the last one, you know, to disrupt?

I think this technology probably disrupt calls center workers first, to be honest. Like if you look at some of the language modalities around speech and you know the work by character AI and many others, I mean, that's as good as any call center worker now, especially with the preval augmented models and things like that. I think again, though, we just don't know. Like I don't think artists will be that disrupted by this because it will open up new forms of art. Although from a supply demand perspective, creating 10 million, 100 million new artists may not be so hot for certain types of artists.

The one that I really think we'll do is literally honest, it's power pointing and things like that, the visual communication mechanism. I think not from us, but from someone else that will be solved in the next couple of years. And there is a lot of tedious make work done by that, which require less entities. Again, it was the promise of legal and others that kind of coming forward now. Then we need to think about what jobs we can create to balance that out because again, the feedback cycles are social for that. So that's why I think it should be a communal discussion rather than just an individual one because we need to balance these things out, especially at a time when the economic conditions don't look so hot, be honest.

*There's a question right there. Hi, Ma'am. C****an you talk about how you envision your technology being used by, for example, you mentioned in high potential young people in places like India, Indonesia, and so forth?***

What potential do you see being unleashed? And why do you think, for example, government there should be endorsing that? So, yeah, I think, you know, this is technology that is amazing and powerful. And we've found that actually, you know, we have 16-year-olds, 19-year-olds, others taking this and extending it out because they kind of didn't learn the old bad lessons. They just got the new paradigm and they've kind of got it and they're building things for their own local communities already. So one of the things we're doing is, you know, we're going to the IITs, we're going to the undergraduate societies and then we're also having these courses like Fast.ai, where Jeremy's done a fantastic job to enable anyone with basics of programming, get more than basics, to kind of extend and understand this. Where it goes, I'm not sure, because the utility for every single country is different, because the country, culture, and the pace of development is different. But there's no reason they shouldn't have access to this technology. Because if we have the current paradigm, there is no circumstance to which the would be allowed to have it, except for via APIs, etc. and it would never be customized to the local culture and context. An example of that is we really stable diffusion and then somewhere a team basically took it and changed the language and code of a Japanese. Because if you put in salary, man, it's normal stable diffusion with a very Western-oriented dataset. Very happy guy

with money. And the Japanese concept, not so happy, you know? So I think that'll be very exciting.

The other thing we're doing is we have education tablets via a child form that we're deploying in a refugee camps around the world, that teaches literacy and numeracy in about 13 months and one hour a day. Things like that can be the basis for including applications like Replit, which have allowed children all around the world and young people to basically start accessing these tools to create new systems and interactive experiences that make them think different. I think once you have this as a new primitive, that will go even better and even bigger, even to some of the least fortunate people in the world, because they finally have agency. Because if you don't give them tools, they won't have agency. And so I think that's what we need to do. Give those primitives they have agency to accelerate. And thus for all the governments who have spoken to have been very, very friendly. So like, yeah, I think we like you.

Seems like also if you use a lot of different talent, marketplaces around design, coding et cetera, you often find participation globally and often from very unexpected places. I remember I was looking for a logo for a website and I had a lot of respondents from Pakistan and a variety of other places. So it feels like this sort of tooling really opens up for augments people who are already trying to participate in the public economy and just sort of broadening this couple, but they can do.

I mean, this is the nature of entrepreneurship, right? The nature of entrepreneurship is that you take something and then you optimize and then you figure out new ways to take it to deliver value. And if you deliver value to someone, then you get paid. And so again, it's unleashing a wave of creativity, whereby we don't exactly know what people do, but you know, it can be everything from labels, from SME shop in Pakistan, all the way through to copyrights and support for legal stuff like you don't know. But if you never have access to technology, you definitely won't know.

Some questions in the back. Hello. Yeah, so not all people around the world have access to like an interactive teacher and you know, you learning can only get you so far. So for those people who need some sort of reactive and interactive teacher, you know, having an AI powered system that is fully interactive, just like that human would be, seems like a pretty interesting solution. ***What do you think about the trust that humans would have for having AI powered learning system teaching how you****.*

So I think it's kind of you pull that up. Not everyone has access to this. Like I said, one of the kind of charities sides of things is putting these systems into your refugee camps on the local environment. So there's also been fantastic human with basic feedback groups.

The new year will be announcing some very big initiatives around this because you've seen humans actually interact with AI even basic in a very positive way. Also negative rate depends what you do. Like you see Tina and Arana and Joshi and China and the ponds that people form because they have no other options there. I think the future is kind of like humans and AI friends

and they can either be manipulative or they can be positive. I prefer for it to be positive. Like to be honest, I want to push forward one AI for child. People don't really like that name. I'm CEO, but that's okay. I can do it. And I think again, that would be the future fully personalized education that's optimized for your objective function.

Bring you the information you need to change value in a positive way. And you can't do that in a lot of these environments because it's one student with 300 teachers. 300 students per teacher. We have a bit different. Or no teachers are all. So if you have older kids teaching younger kids and other things that you give them the tools, they will take it and they do ridiculous things.

There's a book called The Diamond Age by Neil Stevenson and Natasha on some of this too. Yeah, I'm doing the Young Ladies' Primord. That's fair. Exactly.

Hey, mod. Big fan of UN's stability and the work you've done. One quick question. **Roughly a week ago, there was this discussion between a hugging face on the hugging face form between stability and runway around light and share.** Could you expand on for the future open source models release by a stability? What license did you put what we expect? And what would the terms around those models be?

Yeah, so you know, because this model was so powerful and trained on a snapshot of the internet because the developers want to do that. Again, we go to developer preference a lot. We came up with this new creative ML open rail license that basically necessitated ethical use, possibly used to a safety filter. But isn't the standard for models? Are models normally Apache or MIT? But it wasn't awareness that you know, we need to see what some of the uses of these models were and other things as they release.

Originally from pump is as we moved to stability release models, we had a big discussion within our developers. And so as stability, we decided that we're only going to release SFW models now going forward and then think about these license terms to make them more defensive, ideally by doing testing and really rigorous kind of things around that because stable diffusion 1.4, 1.5 were released as a collaboration. The devs had the option to release it any time.

Now, you know, we had some portion around the policy responses, at least in nice letters of Dress to me and other things, as well as some of the legal ramifications they can potentially face because you have different legal elements, different geographies. So we recommend against releasing that until the stable diffusion foundation was set up, which will be literally the next few days we'll be making announcements about that.

We think this is the optimal way to do it rather than having any one company control these things. And again, we think that these discussions around licensing and things should be more open. As a private company, we have our own space as a catalyst to this area, but we should not be the controller of open source AI. We're just leading you down to a bad path, right? Because inevitably, we'll have bad outcomes. So this is why we have this foundation model, this open model. And while we've accelerated now, we want to have people's input on how to do all

this properly because it's important. It's important that everyone collaborates on this. And if you've got a voice, you make it heard as you develop this because again, don't look for tech for the people by the people. That's not one company controlling everything.

Also there was a bit of mess up last week because it was like, ah, what you're releasing the model? Why are you releasing that? I'm in a meeting. That was a bit crazy. But it's all good now. All good.

Hello. Oh, sorry. Hello. I was just curious. ***Stability has any plans to release an open source text to text generation model in the future.*** *Or if it's going to be exclusively focusing on a image in video.*

- So, like I said, one of the groups that we support and we're formalizing and helping go fully independent is Eleuther. And Eleuther released the GPD Neo J and Neo X models, which are been downloaded 25 million times by developers. You either basically use GPT-3 or you use that.
- Right now in the plans, there are chinchilla optimal models and others that can be used and extended in a couple of weeks.
- There are a Carper AI lab, which has done a fantastic work around trust to represent learning.
- We released the instruct model framework so you can take the large language models and press them down up to 20 million parameters.
- There's a huge amount of work around language that's being done. And I think I might have mentioned, you know, we released the Korean language model, the polyglot model, which is the most advanced Korean language model.
- We'll have models across all sorts of different languages and text as well. Text is very important modality because it has this compression and structured element, which is why text models are larger than the image and other models. And so I think the combination of all these things will be super powerful. I think that's something we're actually supporting with thousands of GPUs at the moment.

I have two questions, but I'll just ask one and I'll do the second thing first and interesting. So there are, you know, I think, understandably, many artists were very upset by stable diffusion and kind of like the thought that like their aesthetic is being stolen, right? And it seems like you stand at least in the United States on firm copyright kind of standing. But I think it's interesting that like copyright is sort of based on human presumptions of like learning the static takes a long time, outputting something and someone's aesthetic takes a long time. Just thinking long term, ***I'm curious if you have any view on where copyright law needs to go, not necessarily next like two or three years, but like five or ten to support kind of where the direction the society is going, where technology is going.***

Yeah, I think that's a fantastic question. I think it's very difficult to kind of encapsulate everything there because if you make aesthetic copyrightable, then that opens up a huge kind of worms. But then again, if you look at music, for example, and some of the rhythms, they are

copyrightable and there's different kind of elements. This is a very complicated topic. Again, I think it should be an open discussion. So you know, we released it and then I got blamed. I think I had like five and a half thousand quote tweets by artists against being one person and things like that. It's fair. It's scary and what's happening. The reality is that less than 0.5% of the dataset is artists, so to speak. It actually might be 0.1%. And these models, you can only do it if you're intentional about it. If you go and you say Bob, then you'll get Bob's style if he's very specific. But then you can combine it with another artist, then is it still copying? These are uncertain gray areas from an illegal moral and ethical thing. And that's why what we've tried to do is engage with artists and understand them because we had AI enabled artists basically contributing to the model creation itself.

So what we're trying to do right now is have things like attribution opt-out mechanisms for artists because the models we trained anywhere in the open and they're fine tunable. And I think a lot of the responsibility is on the originator of the model of the, the originator of the prompt of the model because these models will not do that artist unless you tell it to you. If you excise that artist from the dataset, the model quality does not decrease necessarily. Again, it's tried on two billion images. If you have an attribution mechanism around it, how do you pay it? These are an answer questions that I think needs to be answered. And the copyright thing I think will evolve to be not much different what it is now. Except for things like content authenticity and the evolution of that will put a bit more structure around some of these gray areas. And again, we're a bit different in that we're willing to discuss this. We're willing to get feedback and participate in this. Whereas a lot of other companies, they train on artists, they turn on the things that you've been telling them because you can never interrogate the dataset. You know, you don't know what those little neurons are made of.

Yeah, I guess a quick question. I think it's just on the computer. So it seems like you're doing a lot of innovation in datasets and models, but it seems like computer steel, very centralized and the designs with basically just a few companies doing it. And if you even trace it down is just basically like two companies, culminating in space. What are your thoughts on that? And the second one is like, do you have any plans to at some point in the future get involved in the chip design or work with someone?

One thing at a time. Come on. Right? And look, we're evaluating many different architectures from Tramian, Sandalmov, Earth, Cyribras, etc. And you know, centralized compute for DL is very important because of the interconnect. I think we're going to see this ridiculous exponential nature when the H100s and other come online. PV5s, etc. Because scaling is kind of solved. And that should be scary to some people, say the least. I think the interesting thing is now we're kind of going from the high level to actually going down to the hardware component level and some of the optimizations with partnerships that will be announced in the next few weeks. And then the organization allows us to do that because they know that everyone's using this technology. I think as well what you'll see is you'll see some differential elements around this. So like we supported the test for hivemind that was at Neoreps last year for distributed training and we did some distributed training of clips and things.

There should be alternatives to just having gigantic clusters. And again, I think our focus on training these models enables that because what do you do? You basically expend the energy up front and then you put it out so nobody else has to expend that energy rather than creating the same model again and again and again and again. So that's the paradigm we're looking at. But I think there is this gap that needs to be filled. So in fact, we did build the national research cluster. Any university can access what we do now and academics no longer have that gap. Particularly as we scale up to five to ten times larger next year on our cluster. So yeah. And we're at time, unfortunately. I just wanted to thank Notion again for having us here. Everybody for showing up is a packed room with people standing and sitting in the back in front. So thanks for showing up tonight and then a huge thanks to Ahmad for sharing this stability story both future and past. Thank you, Aladdin. This is amazing. Thanks, personal team.