Innovating at the Frontier of AI Building for Tomorrow

Words	7,660
Duration	00:42:48
Recorded on	Unknown date
Uploaded on	2024-09-18 00:30:38 UTC
At	Unknown location
Using	Uploaded to aliceapp.ai

Speakers:

Speaker A - 20.47%

Speaker B - 23.37%

Speaker C - 34.75%

Speaker D - 21.24%

Speaker E - 0.17%

Notes:

- Panelists include Richard Socher, co founder of U.com and former chief scientist of Salesforce. Jeff Zhang is the CTO of together AI, a leading AI acceleration cloud for building and running generative AI workloads. Today we're going to talk about innovating at the frontier of AI.

- U.S.com has over 100,000 organizations today that have been created on the hub. For enterprises that need to know what's going on, we have this enterprise hub subscription. As a startup founder, you have to remind yourself always, every crisis is an opportunity.

- There are hundreds of thousands of models. How do you pick which ones you're going to recommend to your customers? The goal is to become the single one stop shop for people's acceleration requirement.

- Richard is now a CTO of a foundation, fast growing startup. He was a professor for seven years in Zurich. Tell us a little bit about how you've adapted personally to this sort of new role.

- Richard Quest: One of the key questions companies want to know more about is differentiation. Quest: People can replace ten Google searches by asking you. com and an agent a question. He says agents can also ask you questions back and you can program them for any workflow. Quest: What are some things that companies can do to differentiate themselves?

- Dreamforce attendees can receive a complimentary special edition of this year's time 100 AI issue. With our thank you all for listening. We hope to see amazing things from all three companies.

Speaker A **00:00:00**

Um, I have an amazing set of panelists to join us for, uh, talking about innovation at the frontier of AI. Uh, let me start by, uh, calling them, uh, one each. Um, Richard Socher, CEO and co founder of U.com, an AI productivity engine that offers multiplayer and AI agents for knowledge work. And also the. The former chief scientist of Salesforce. Please welcome Richard, followed by Jeff.

Speaker B **00:00:40**

Hi.

Speaker A **00:00:40**

Hey, Jeff.

Speaker B **00:00:41**

Hi, everyone.

Speaker A **00:00:41**

Welcome the head of product at Hugging Face, the number one platform for AI builders in the world. And also please welcome Sir Zhang. Sir is the CTO of together AI, a leading AI acceleration cloud for building and running generative AI workloads. He's also the associate professor of department of computer science at UChicago. Thank you all for joining.

Thanks for having me.

Speaker A
00:01:08

So today we're going to talk a little bit about innovating at the frontier of AI, what it means for being the builders that each of you are. Uh, thank you very much for your time. Uh, let me start by setting the stage a little bit. Uh, the speed of change in genaie has not slowed down in two years. Uh, and given the quickly evolving nature of building this technology, companies are building products and it's a constant change. We're here to discuss challenges, we're here to discuss strategies, and we are here to discuss solutions. Uh, before we dive in, opening it up to all the panelists, you have each being leaders at various levels in the stack. Richard, you've been a researcher. You, uh, worked on several companies. Uh, you guys are doing fantastic work in hugging face and together. AI is powering a lot of, uh, generative AI applications on top as leaders. What have you seen in the last two years, this sort of tectonic shift in technologies? What have you seen? What has surprised you the most? And what are you most excited by? Start with Richard and everyone else, jump in as well.

Speaker C
00:02:15

Sure.

Speaker B
00:02:16

Yeah.

Speaker C
00:02:16

Uh, hi, everyone. Great to be here. Uh, back at Dreamforce, um, I think in the last two years there's sort of a big shift. Some, uh, people call it BC not before Christ anymore, but before chat, GPT. Uh, and that transition, uh, in terms of helping companies use Aih, has moved the conversations from why do I really need AI to. I definitely need it. But not every company yet knows that over the next few years, they're likely going to hit some kind of innovator's dilemma, because AI is going to massively disrupt almost every industry, starting most urgently right now with knowledge industries. So it's still surprising to see that people have been able to build some prototypes and know, they should be building some prototypes, but to get from those proofs of concept into actual production is actually still very hard. And that's what we're focused on@u.com. to help different kinds of companies from biotech firms, media companies, hedge funds, VC's and so on, to actually, next year, really get this into their productive workflows. And it's a little bit like self driving cars. You know, you can, even 20 years ago, you could build a self driving car that drives on the highway very easily, and, like, weather is good, and for a couple minutes, it's fully self driving. But it took another decade plus to actually get the car to work every day at scale in San Francisco. And so we're seeing, uh, similar things with agents.

Yes, in San Francisco, not yet widely available to the rest of the world. Jeff Builders know you, your company. Tell us a little bit about what hugging face is, and tell us a little bit more about what has been exciting for you in the last two years and what has surprised you the most.

Speaker B

00:03:57

Well, hugging face is the leading open platform for AI builders. What we're trying to do is that, uh, all of you can build your own AI. And over the past couple of years, we've seen sort of, AI go from, uh, research and scientific driven, uh, topic to an engineering topic. And now it's very much a product topic. And I think we see this all over Dreamforce. Ah, uh, today. And what worried me, though, in this process is that, um, we could find ourselves in a world where only one or two or three, a handful of companies in the world, decide what user experiences we all are experiencing in our daily life. For everything that we do, every single app on our smartphone, and the world that we want to live in, the world that we want to enable, is one in which every single company can control their own AI, can maybe build their own models, can be responsible for their AI features. And so that's what we're trying to do. So I'm really excited about this new phase, as you say, Richard. Like all companies last year were wondering, okay, what can I do with this thing? Now they know what they can do with this thing, but the next step is like, how do I build it in a way that's responsible, in the way that I can own it, in the way that it's reproducible, and all these great stuff that open source does for us.

Speaker A

00:05:24

Thank you. Same question for you as well, sir. What is exciting for you? What does together AI do?

Speaker D

00:05:30

Yeah. So, together, we are trying to build up an, uh, infrastructure tailored for acceleration of all the technologies that, uh, Jack is talking about. So I think what's really exciting and surprising to me personally in the last two years is how a single technology is able to enable so many different things, right? So I think that's really the paradigm shift that we are seeing today. It's not like for one application, you have one technology and one solution. It's just like single layer of technology, it's able to power so many things, right? So the technology being generic and general means that, uh, we are able to build infrastructure for it, right? If you want to enable every single company and every single person in society being able to use the technology, how can you scale it up? How can we make it fast? How can we make that cheap? How can we add safety and transparency layer over that? That requires us to actually optimize the single workload across different stack of the infrastructure. And that is why, uh, we are very excited about together and that's what we are trying to achieve.

That is awesome. Um, I want to change gears a little bit and talk about one attribute that every company today is really keenly focused on, which is adaptability. Uh, change often comes slowly first, and then very, very fast. I think we are sort of in the very fast phase in our domain. We each adapted, uh, quite well to this sort of slew of technologies that are coming. Uh, let's talk about adaptability a little bit. Richard, you and I have known each other for a long time. Uh, you are the founder of Meta mind. You're now the founder of you.com dot. Uh, how have you adapted as an entrepreneur? What has been the change in how you approach building companies, building technologies? Tell us a little bit more about your journey.

Speaker C

00:07:17

Yeah, so I think at meta mind days, like I mentioned, we have to actually convince people that AI could be relevant and tell them that there are use cases for it. That could be exciting. I think the big shift now is that thanks to generative AI and large language models and APIs, but also open source, everyone can build a quick prototype. Uh, and now we're seeing currently agents being built. The problem is, with every step, if every step of an agent is like 95% accurate, you end up after 15 steps, failing half the time, just mathematically. Uh, and so while it's exciting for companies to be able to build these prototypes quickly, um, it turns out it's still a lot of work. So what we're often seeing now, and what surprising is that companies actually have tried to build what we built first, uh, as a prototype then realize it doesn't get the adoption. People don't quite know how to use it, uh, and it's not quite accurate enough. And, uh, so it is still just like in the past, it's still, uh, very hard to make it accurate at scale. Uh, and that's kind of what we're excited about helping companies with@u.com. especially when it comes to knowledge work, taking external data and internal data and combining all of that. Uh, I think other than that, um, what's been really exciting to see, of course, is that the technology is now flourishing. Everyone actually understands the impact it can have, though a lot of times we think about it like there's sort of two buckets for companies. You can, in the periphery, improve all your processes a little bit with AI, and everyone understands that, right? You have a CRM, you have sales, service, marketing. You can make that all a little bit more efficient. What a lot of companies haven't fully understood yet is that it can also massively disrupt the entirety of their business model. Like, if you're a car company and you have self driving cars, that fundamentally changes your entire industry. Uh, if you're helping people with programming and you're an agency that helps find programmers, and you can have more and more programmers be automated, especially if they're doing relatively simple kinds of programming, your entire business is going to get disrupted. And not every company has fully internalized that innovator's dilemma that's coming, uh, towards them.

Speaker A

00:09:32

That's right. I think embracing the innovator's dilemma is probably a key prerequisite to adapting in the first place. Um, Jeff, similarly hugging face has been around since 2016.

Recently, uh, you guys announced some amazing work you're doing with enterprise companies. Uh, a lot of companies are using what you guys are building and then building systems on top. What are you seeing with your customers in terms of how they are adapting to the technologies? What are some things that you are doing to help them along in this journey?

Speaker B

00:10:04

Well, first, um, in terms of adaptability, we're in such a better place today, uh, for users of open source to build any kind of AI application and adapt over time. If I think back to, uh, just a few years ago, transformers, uh, was used for natural language processing, right? So you would complete the sentence or say that this email is urgent and not that email. And then the computer vision folks were working on different kind of algorithms, like convolutional neural network, and the audio guys were using something else. And all these different AI teams didn't have the same language. They didn't have the same tools. Um, and today, Transformers is what powers any kind of machine learning use case. Like, we even see time series use cases now being solved with, uh, transformers. And that means that you can adopt a single set of tools to build any kind of AI feature, uh, within your product. And if you build those features on top of open source, when the Lama 3.1 comes around, we like better, uh, performance than the Lamma 3.0 and 2.0 before then. It's really easy for you to adapt to it. Another thing that is new is that today, in order to adapt some of those Bayes models to your use case and your data, it's much cheaper and easier. And you don't need like a big cluster of GPU's to retrain all the model weights to your data. You can actually do it very efficiently and then apply that recipe to any new model that comes along. And we're very excited at hugging face to be in this position where we can provide a, uh, central, uh, set of tools, from our open source libraries to the hugging face hub where you can find, I think it's, what is it, like a million models today, uh, that are contributed to by the community for any kind of, uh, use case. It's all there, all the same set of tools. And you can apply that today. And you were mentioning enterprise. So of course, uh, the hugging face hub is sort of a single player mode, uh, kind of, um, a portal. And when teams want to come together and work and do things privately, they create organizations. And we have over 100,000 organizations today that have been created on the hub. Um, and so for enterprises that need to know, like, what's going on and at some security onto it, we have this enterprise hub subscription, uh, that's really, really helpful for companies to build AI with open source.

Speaker A

00:12:38

Well, I had no idea. Go ahead, Jeff.

Speaker C

00:12:40

Uh, brings up a really good point, namely, uh, and related to your question, too. There's

almost a crisis. And as a startup founder, you have to remind yourself always, every crisis is an opportunity. And the crisis here is that there's a new model, a new large language model that comes out every month, right? And so then you have companies, they just spent a lot of money getting a thousand enterprise seat licenses or a big API deal with OpenAI, and then two months later, there's a better model out there. And they're like, ah, now I'm stuck with that other LLM and I can't easily switch. And so for us@u.com, we basically said well, we're going to have all of these models. We had strawberry zero one, like the new OpenAI model, the same day it came out on U.com. and you can just know that you're kind of future proof of all the LLMs, having a partner that does all that change management for you, that updates your vector databases on your own internal data and all of this. So every, every crisis is an opportunity, and for startups, and, uh, it's really opportunity to help all the other companies deal with that, because we're, this is our, like, lifeblood. But if it's not your core expertise, it's helpful, just like using a CRM, it's not helpful to rebuild all of that internally.

Speaker A **00:13:46**

Yeah, exactly. And I think it's so strange to be in a world where you're getting better, cheaper and faster. Usually it used to be two out of the three. Right? You're getting all three right now. And I wanted to follow up with you, sir, as well. Uh, you are in the hosting, training, inference, um, efficiency play. Uh, what do you see? To Richard's point, there are hundreds of thousands of models. How do you pick which ones you're going to recommend to your customers, which ones you're going to help, um, customers build on top of?

Speaker D **00:14:19**

Yeah. So I think there are multiple factors in that. So, like, what we see when customers actually try to navigate technology is, there's the tension, actually are two different dimensions. Right? So there's one dimension that, uh, there's technology coming every day. Not only new models, but every single day. There are probably like tens of, like hundreds of papers that have been published. Some of them make your things faster, some of the things make your sales better. So all the new technology being produced essentially every single day, that's like one of the tension. And on the other hand, there's another dimension that your workload, your data, what you want, what you have to achieve, also keep changing every day. And one of the struggle we are seeing that, uh, our customers are trying to navigate is how can we understand the latest technology, how can we understand research, and how fast can we bring research to production? So I think that's really something we are really passionate about. So, like, about this concrete question about how can we pick which model to use. Right. So there's the technical solution to that.

Speaker E **00:15:18**

Right.

Speaker D **00:15:18**
For example, we have this, uh, kind of like, interesting paper called mixture of agents. We show that you can actually have a whole bunch of models. They kind of divide and conquer, and they actually achieve something that's better than a, uh, bigger model can do.

Speaker E **00:15:32**
Right?

Speaker D **00:15:32**
So that actually gets some of this kind of out of the picture.

Speaker E **00:15:35**
Right.

Speaker D **00:15:35**
You can actually have this single abstraction that, oh, model can draw that model, have good at different things, model have a different profile and the performance and the cost. Right. So this is a single abstraction of acceleration to actually help you to actually achieve the optimal part.

Speaker E **00:15:49**
Right.

Speaker D **00:15:49**
So that's actually something that we are trying to build. So the goal of, um, together is really try to make sure our customers do now think about performance, uh, and the cost and all of those. Ah, essentially, we become the single one stop shop for people's acceleration requirement. So that's what we are very, very excited about.

Speaker A **00:16:08**
That's amazing. I want to close out the adaptation topic, uh, pun intended, uh, with a personal story from you. You heard from Richard about sort of his journey, adapting. As an entrepreneur, uh, you're from academia, and you're now a CTO of a foundation, fast growing startup. Tell us a little bit about how you've adapted personally to this sort of new role.

Yeah, so it's actually pretty big change. Yeah. So I was a professor for seven years. I did Zurich.

Speaker E **00:16:40**
Right?

Speaker D **00:16:41**
So I think there are things that change. I think that actually, like, kind of different.

Speaker E **00:16:46**
Right.

Speaker D **00:16:46**
So what was similar? I think, uh, what's really important in building any organization is try to empower the talent, right? So we were pretty kind, uh, of, uh, like, lucky before as a professor to work with really great students, right. And right now we are building up this amazing team that actually, uh, bring all the acceleration technology to our customer. I think that's something that kind of the same. On the other hand.

Speaker E **00:17:12**
Right.

Speaker D **00:17:12**
So, uh, now we work with more customers compared with, uh, when I was a professor.

Speaker E **00:17:16**
Right.

Speaker D **00:17:16**
Uh, and, uh, work closely with the community to bring many of the amazing ideas actually to the real world.

Speaker E **00:17:22**
Right.

Speaker D **00:17:23**
So, yeah, I think for me personally, it's kind of a lot of fun, uh, to see the transition. And,

uh, yeah, personally, I'm very excited.

Speaker A **00:17:31**

Thank you. Thank you very much for sharing. Um, I want to switch gears now to talk about a different topic. In addition to adapting, I think one of the key, one of the key questions companies want to know more about is differentiation. Everybody seems to be jumping on the bandwagon. How does one differentiate themselves? How does one focus on the focus? Um, starting with Richard. Richard search is a super competitive space. Uh, there are big companies, well, capitalized companies doing amazing research, some of which is original, in fact. Uh, and then there are some amazing incumbents, like u.com, like perplexity. Um, how do you think about differentiation? And what are some things that you can share with companies that are looking to, to sort of get into the space, uh, and want to differentiate themselves? How does one stand up to the big guy?

Speaker C **00:18:23**

Uh, yeah, so we used to be, when we started in 2020, a, uh, search engine. Uh, and then we realized over the years that a lot of people ask very simple questions to a search engine. How old is Obama? What's the weather tomorrow? Things like that. And it turns out there's not much you can do to be ten x better in telling someone, this is Obama's age within 1 second or less. But what we can do, and the reason we're now calling ourselves a productivity engine instead of a search engine is that what we realize is people actually can replace ten Google searches by asking you.com and an agent a, uh, question. That agent will then go out on the web, search for you, find new information, put it into a program, write code for you, do some mathematic, realize it needs some more information, go on the web again, and then actually give you a final work product. So if you're an account executive and you want to be prepared for a meeting, you can just say, oh, here's, uh, the company I'm about to meet. What should I know about them? And it'll go on the web, make sure it knows their stock price, or maybe things are going up. Maybe they just had to lay off a bunch of people. Right. You might want to go into that meeting differently. Sure. Uh, and then it finds all these information and then all this information, and then the big difference, once you realize you can actually be ten x better by replacing ten Google searches and productive workflows, is the agents can also ask you questions back and you can program them for any workflow. You're in marketing and you say, oh, I want to, I'm getting a new document, a new technical description of a very complex product. And now I want you to describe that to people in manufacturing, in the industry of manufacturing or in finance. Uh, and you basically get a description of that very technical product. And then out comes a bunch of tweets, a bunch of LinkedIn messages, and a bunch of like, email campaigns that come out and you all of a sudden had to not look at all the competition on the web and your internal documents and so on. And we're replacing more and more steps for you to just become more productive. And so that's the first differentiator. Even in that space, there are some folks who are saying they're doing

this, they also have citations and bringing facts together. And there we're focusing on accuracy. Uh, we've been at this for a very long time, for multiple years, doing retrieval, augmented generation, before the term was reinvented. Uh, and it turns out, again, very quick to build a prototype. But then you realize some of these prototypes, their citations are just random links behind sentences, and they have nothing to do with one another. And so the companies that really care about accuracy and fintech and biotech and media and so on, they start to, uh, end up appreciating that additional accuracy.

Speaker A **00:21:04**

And working with us, that's so important. So you're saying go deep, focus on the Personas, find what's important, and find what is different, differentiated between each of these Personas, their workflows.

Speaker C **00:21:17**

That's exactly.

Speaker A **00:21:17**

And then that's how one finds differentiated. Thank you, uh, Richard. So, on similar lines, part of together, AI's mission is to foster an open source ecosystem. Um, how does one differentiate in that context, when you have proprietary things that you're doing as a company, but you're doing it on top of some open source things that are coming, how does one draw the line? This is sort of the million dollar question for a lot of companies in this space, because there's so much being done in open source. So how does one think about doing it in a fair, consistent manner that can help customers?

Speaker D **00:21:54**

Yeah. So I think one thing that's kind of important is to realize we are actually seeing now the end of the field is kind of beginning of the industrial revolution, right? So, because it's beginning, I think the tension is actually twofold. The first one is we have to come together and facilitate the market altogether. The market need to grow, right? So we need to produce technology ideas and to make sure other people can actually build on it. And together, we are going to have something, have a much larger pie to actually work with. The second, uh, dimension is, of course, we need to differentiate ourselves. But what we found, uh, in my experience, is, so, first one, there's a difference between idea and a product that actually our mission critical applications. So there's a difference between running something on one gpu or running something on 10,000 gpu's and just squeeze the last 1% of the performance. So I think when you go to that region, there's actually a lot of things that, uh, you can actually build to actually differentiate, uh, uh, comparison ideas. And another thing that kind of interesting is because we are building infrastructure, so we start to optimize, uh, the workload across different layers of the infrastructure, the more

deeper you are going into the infrastructure, the more closely you are growing together with the customer. Uh, we found that pretty natural to actually draw the line. So there is something that going to facilitate the market, going to facilitate, um, all of our friends in the openstart community. So we should definitely work together with people, really make the market larger. Uh, and there's something that, uh, is probably only useful when you have 10,000 gpu's want to run that. And that's probably a very natural, uh, line to draw between these two things.

Speaker A **00:23:45**

Yeah, love to hear from you on that one. How do you think about building proprietary differentiation when so much of what hugging face does is about open source models?

Speaker B **00:23:57**

Yeah, I guess, um, we think about differentiation very differently between, um, if we think about hugging face or if we think about our users and our customers. Right. If we think about hugging face, um, our goal is to sort of expand the universe of AI. Like we don't see it as a zero sum, um, sort of game or market. And our goal is that through our platform, platform, we can accelerate and democratize and give to as many people and companies as possible access, uh, to the latest, uh, AI technologies. And the way we go about this isn't to think about differentiation, but rather to work with everybody. Because as I say, we want all of you to be able to build your own AI. And maybe some of you are working on AWS, maybe some of you are working on, on Azure, some on Google Cloud, etcetera. So we work with all of these companies so that if you're working in Sagemaker, Azure, machine learning, vertex, AI, Cloudflare workers, etcetera, like we have great product experiences for you to do that. If you want to use an Nvidia GPU, if you're working on an AMD instinct GPU, if you're working on an accelerator, inferential on AWS, TPU, on Google Gaudi, with Intel, like we, we have built solutions for you to do that work. And I think on our slack channel we have like ten or 100 x more people who are outside who are not hugging face employees than hugging face employees. Um, so that's how we think about that in the context of hugging face. Now, for our users and our customers, it's very different. Because if you think about um, how they are using hugging face, they are using hugging to customize their AI and build differentiation for their own business. And they will use our open source libraries to do fine tuning, uh, to create their own models, uh, to create adapters that work specifically well, for their companies, so that they can control their costs, get better, more specialized, smaller models to, uh, save money and take control of their own AI. So yes, through enterprise hub, uh, through our expert support, we help, we teach, we accompany, uh, companies in building differentiation for their own business using open source AI.

Speaker A **00:26:22**

That's awesome. Um, when we look at companies, we, at least in the AI domain, we look at companies as is this an AI native company or is this a company trying to build AI and infuse it within the products that it's building? Um, in your opinions, anyway, uh, I'm just going to open this up to all of you. In your opinion, like how should each of these companies think about their differentiation? And if you're an AI native company, what does differentiation look like for you? What should your focus be if you're a non native company? Very interesting moats exists, but how should one think about that?

Speaker B

00:27:02

I think, uh, one thought maybe just kick it off is that, um, I think we're going from, uh, a, uh, world where we think of AI feature being powered by a model, to a world where we think about AI features being powered by systems. And in systems, there's a lot more than the model, and that helps build a lot of differentiation. So if you take as an example, hugging chat, which is sort of like our version of chat, GPT is like to use open models. The app itself is open source, but it's about the same experience. When you're using hugging chat, you're not using just the uh, model that you've selected. You're actually using a whole bunch of models because you're giving tools to the agents in order to do things. We're looking at your inputs to filter things. We can add some guardrails on top. And so all of these things are different models. And by building systems, not just models, you can build differentiation by customizing models to your own data. So that's the fine tuning the adapters I was talking about. You can create differentiation.

Speaker C

00:28:04

And uh, I love hugging chat because it also has u.com integration and our APIs. What we realized is there's some use cases where companies can just use an off the shelf product as is like u.com for the knowledge work, or the analysts or the researchers. But then you may also want to incorporate AI deeper into your own product, and then you can go one level below and just use APIs for that in your business and infuse that into your own workflows. Excited to partner also with Slack and other folks at Salesforce, uh, to bring up to date LLM outputs that incorporate news and web knowledge into and merge it with company internal facts. So I think that's one. And then that lower level below that is even you build your own AI, right? You need to look for every business, whether an application is part of your core or part of the things you have to do around it. Like it doesn't make sense to rebuild in CRM. And then it also doesn't make sense to rebuild all the AI for sales, service and marketing automation. But if you're an insurance company and your core is understanding risk, then you probably should be building your own AI models, either starting from open source or going really deep into identifying and classifying, uh, and quantifying risk as an insurance company. Right. If you're a car company, you should really work on your cars, being self driving, driver assist and so on. And that is very core to you. And I think that's kind of two way. Those are two buckets that you fall in. And usually it's a little bit of both, right? Even a car company needs to have sales, service and marketing,

and can rely and build versus partner with others. I think maybe another interesting aspect of this whole thing since tsur, uh, brought up, um, the industrial revolution. During the industrial revolution, there are a bunch of companies that worked on making steam engines more efficient. And initially people thought, oh, if we have more efficient steam engines, we will be able to use less coal, uh, and less energy. But really what happened is people just used more steam engines and use them for even more use cases. And the equivalent in the age of AI that we're now in is that the marginal cost of intelligence is going down. And all of a sudden you can use it in a lot more areas. Like everyone can have a personal assistant, everyone can have a personal doctor, everyone, uh, can have more and more automation that used to require intelligence be done for them. Mhm. And I think that is something that not everyone has yet fully internalized. Like these workflow tools to build your own custom agents are incredibly powerful for a lot of different use cases.

Speaker A **00:30:43**
That's right. They are multitask learners.

Speaker C **00:30:45**
Right?

Speaker A **00:30:46**
Yeah.

Speaker D **00:30:47**
So agree with all of this.

Speaker E **00:30:48**
Right.

Speaker D **00:30:49**
So I think fundamentally, if you look at what the AI really is, right. So it's really the reflection of two things. So it's a reflection of the data that you have. It's really that you have very unique data set. The model is a summarization of that. It kind of reflects the distribution of the data that you see. And second, it reflects part of the workflow that you are trying to essentially try to replace and try to optimize. And if you take this wheel, the differentiation is actually pretty natural. Every single company have very unique data, and every single company have the workflow that I've been relying on that probably only one of the few companies have in the world.

Right?

Speaker D **00:31:31**

So I think you take this view, there's going to be this very natural but unique place that AI going to kind, uh, of like, improve different components for different company. I think for that part, it's pretty unique.

Speaker E **00:31:42**

Right.

Speaker D **00:31:43**

So what we are very excited about is all of those kind of very diverse applications from different companies magically is kind of empowered by this one technology. So that what makes us really, really excited, because that's one generic thing that you can actually build an infrastructure for. You can optimize them in a generic way, and people can customize and also adapt to their own applications. So that what makes, uh, us really excited. And, uh, that is why I think pretty much all the company going to find their own diffusion factor by looking deep into what they are doing today and looking at the unique data that they have.

Speaker A **00:32:19**

Awesome. Um, I'm going to shift to this favorite part of the conversation here, which is you guys are just world class researchers, amazing practitioners in your own regard, and there's so much happening in the genai space. Um, let's talk a little bit about what you're foreseeing will come in the next one year, uh, or two. Help the audience and me understand a little bit more about what it is that, uh, people should keep an eye out for, uh, what is the future going to look like when it's going to unfold? Of course, it's always a risk to, um, predict something that's exponential the way it is. Uh, but you guys are the closest to the curve as anyone else. What's your take? What are you excited about? Multimodal models, agentic architectures. What are some things that we should be looking at thinking about the next year or two?

Speaker C **00:33:16**

Well, uh, I guess I'm really excited of moving out of the proof of concept stage, uh, into actual production. Like everyone, every company. When you ask, maybe I can actually do it for the audience. Who here has worked on a proof of concept, uh, with something, with AI? All right, so, like 80 plus percent. Who here got that proof of concept to actually run fully in production for everyone at the company? Very similar in most audience now five to 10% of the audience. And so I'm really excited to get that 5%, uh, to be up at 80%. Also next year by helping people build custom agents, in some cases marketing, combining

external information, company internal information. You know, Google Gemini talked about having a 2 million context token window for the LLM. Wow, that sounds amazing. Until you realize that's a five megabyte file. That's not that much like if it's all text, right? It's a simple text file. We are going to launch soon, like a 50 million context window assistant that can just reason over so much more of your company internal data. Uh, go on the public web, ask you follow up questions. It's going to be mind blowing. But it's also, again, very easy for companies to be like, oh, I mean, like how hard could it be? You can hack up quick prototypes, but I think next year we're going to see sort of, okay, we hacked up the prototype, we know where it fails. We can compare it to, like other folks that are focusing only on accuracy and things like that. Now we can actually partner with them and move into production. I think that's very exciting. I think in terms of predictions, um, out beyond that, you know, we're going to see these agents be able to reason over longer time horizons. Right now you can kind of think about some of the different agents as replacing maybe 510 minutes of work. Some are maybe five, 6 hours of work. I think over the next two years we're going to see it replace weeks or months of work, like, and just completely automate an entire, not just single task, but an entire workflow at a much larger scale. We're going to see, uh, these. You know, a lot of people think about large language models, chat, GBT. What that ultimately is on the technology side is a neural sequence model. And these neural sequence models don't really care about what kinds of sequences they're being trained on. It can be sequences of proteins, the sort of foundational blocks of life and biology. It can be sequences of tones. We have music sequence of words and code. Code is particularly exciting because in code you can actually simulate the outputs, test it, and then iterate and eventually maybe even get better than the human training data similar to chess or go. So, uh, coding will happen much, much more. Uh, basically, anyone that is below average at or below average in their knowledge work, including programming, is first going to get amplified a lot. But if they're not using the tools, they're also going to have a harder and harder time in their work. And so, uh, it's pretty existential for a lot of people to go in, um, and maybe to think of other sequence models right. Obviously, voice, uh, and speech is incredibly important. It's going to be more and more seamless to be able to talk to these agents. I don't think it's going to be all voice. Whenever companies think, oh, everything, I'm going to, uh, all my interactions are going to be voiced. They forget that people want to see graphs and images, and even when I ask for restaurants, I want to see images of the food and all of that. But they're going to be multimodal in a much more seamless way. You can have a conversation, you get some input and outputs on your phone, uh, as well as on your laptop. You can basically interact more seamlessly across different modalities, like speech, text, visual, with these models, and they're going to incorporate all of it. Uh, so those are just for the next two years.

Speaker A **00:37:07**

Amazing. That is such an amazing hit list. Uh, personally, I think I'm excited about the

potential of what I call solopreneurs, single engineer, starting a company with all the tools. I'm just amazed with the quality of coding that's possible with these systems and, of course, all the other things that you said as well. So you want to take this? And then, Jeff.

Speaker D **00:37:28**
Oh, yeah.

Speaker A **00:37:29**
What are you excited about?

Speaker D **00:37:31**
Oh, it's very hard to predict the future, because if you think about the lama thumbnail model, it's only a little bit more than one year old, right. So it's very hard to look into the next one or two years. Who knows? Right? But I think there are multiple things that I'm very excited about. So one thing is, uh, I definitely agree. I say, I think, I think next one or two years is where this very big wave of people actually push those ideas into production. So I think that will be the year that people understand, oh, this thing is not only a model, but a very small component of a very big ecosystem that I have running in my enterprise. And, uh, to do that, people need to understand, oh, this is what this thing can do, and this is what cannot do. And this is the way that I guard rail the behavior of the model such that I can see that into my existing workflow and work well with all the other component that I'm relying on. So I think this will be the year that people kind of figure out for some of the critical applications. What are those answers is for? Yeah, for those questions. Uh, and of course, there's going to be this boost of capacity of those models, multimodality, how do they work together? How do different models divide and conquer, and how can we optimize the infrastructure. So, and it will probably also be the year that we see the cost of uh, doing many of the operations that we are doing today start to decrease, but total volume of the market start to increase.

Speaker E **00:39:01**
Right.

Speaker D **00:39:01**
So I think that's what we are excited about.

Speaker A **00:39:04**
Uh, it's awesome. I can't imagine what's going to happen when a cost of compute becomes lower than the cost of storage. Some interesting things are likely to happen that happens. Jeff?

Well, first off, disclaimer, I'm not a researcher, uh, but uh, I'm lucky enough to, uh, work with some amazing researchers at hiking face. And by proximity I see a little bit of where the puck is going on. Uh, uh, some areas, um, one area of research you mentioned, code, uh, and starcoder, uh, family of models sort of show us, um, what is going to be possible. Another direction is a smaller model with a small Lm family, uh, of uh, model releases. What that enables is private, uh, AI, AI that runs uh, in the browser, AI that runs in your phone, on your laptop, etcetera. Soon we'll all sit, maybe with apple intelligence. Um, but uh, I'm really excited about that. Um, I'm also excited about the trend that we've seen over the past year of open models, ah, catching up with the best closed models in terms of what you can do with them. In terms of performance. That's been sort of a steady trend and even accelerating in, uh, the last few, uh, months. And I think that's promising because what excites me actually the most is not so much, um, on the science and research side, actually, it's more akin to what Richard was talking about today. Um, the reality of it is that all the companies that have built use cases, uh, for AI and some initial features were built on top of OpenAI. And I think that's a huge opportunity, uh, for companies, uh, to go from that to something that they can truly own and control, like they would any sort of technology that they put out in the hand of their customers. Um, like an analogy, like imagine that everybody in the US would, would be uh, eating burgers from like one cattle, uh, ranch in the United States, you know, and, well, you don't know exactly, you know, it's like only one breed and you know, it's only one way of doing it. You don't know exactly how they're doing it, but you know, it's the, uh, best beef around, um, that creates a whole lot of issues if, um, something changes along the way, um, if uh, you want to have something that's different from your competitor if you want to own your own burger recipe, uh, et cetera. Uh, so we're trying to make this world where everybody can cultivate their own garden of AI, uh, using the ingredients that are contributed to by the community. The pace, the accelerating pace of, uh, open source models in terms of how they compare, uh, with closed models is a testament to the, uh, open science community working together. Like, if you think one lab by itself in a university, maybe it's, uh, a, uh, few handful of people, it's hard to compete with thousands of AI researchers, uh, in closed labs, but if you put all of the open source community together, we're actually 100,000 times larger. Um, so that's, uh, what excites me the most.

Speaker A

00:42:18

Well, thank you very much for coming, all of you, and sharing all your amazing insights. Uh, welcome to Dreamforce and hope to see amazing things from all three companies, especially. A lot of you work with our teams as well. Uh, and you are, uh, our coveted portfolio companies. Thank you for being a partner to Salesforce and thank you for coming, uh, here to Dreamforce, sharing all your insights.

Speaker C

00:42:38

With our thank you all for listening. Dreamforce attendees can receive a complimentary

special edition of this year's time 100 AI issue.