

Sierra co-founder Clay Bavor on Making Customer-Facing AI Agents Delightful

72 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=RAZFDY_JG10](https://www.youtube.com/watch?v=RAZFDY_JG10)

https://www.youtube.com/watch?v=RAZFDY_JG10

SUMMARY

Clay Bavor, co-founder of Sierra, discusses the transformative potential of AI agents in enhancing customer experiences. He emphasizes that the solution to many AI-related challenges often lies in leveraging more AI, particularly through large language models that can self-correct and improve their outputs. Sierra aims to empower companies to create their own AI agents for customer interactions, significantly improving efficiency and satisfaction.

- The solution to AI problems often involves using more AI, as large language models can detect errors in their outputs better than avoiding them initially.*
- Sierra enables companies to develop branded AI agents for customer service, enhancing interactions and operational efficiency.*
- The company focuses on understanding and embodying the voice and values of its clients in AI interactions.*
- AI agents can handle a wide range of tasks, from answering simple questions to managing complex customer service issues.*
- The deployment of AI agents involves a collaborative approach with clients, ensuring the technology aligns with their business processes and customer experience goals.*
- Sierra employs a resolution-based pricing model, charging clients only when customer problems are successfully resolved.*
- The future of AI includes advancements in multimodal models and the integration of AI into various aspects of business operations, enhancing productivity and customer satisfaction.*
- Bavor expresses excitement about the potential for AI to revolutionize creative processes and improve organizational capabilities across industries.*

one of the more interesting learnings from the past you know year and a half of working on this stuff is that the solution to many problems with AI is more AI and it's somewhat unintuitive but one of the remarkable properties of large language models is that they're better at detecting errors in their own output than in not making those errors in the first place

[Music] joining us today is Clay Bavor co-founder of Sierra before Clay started Sierra with his longtime friend Britt Taylor he spent 18 years at Google where he started and led Google Labs their ARVR efforts and a number of other forward-looking bets for the company Sierra is allowing every company to elevate its customer experience through AI agents and there is no one who knows

more about what AI agents can do today and what they'll be doing tomorrow than Clay you'll get to hear about

how pictures of avocado chairs helped inspire the founding of Sierra why the solution to problems with AI is often more AI and so much more please enjoy this incredible episode with my friend Kay Bor all right clay listen this is a funny start because we know each other so well but can you just tell everyone a

little bit about yourself and just give us some background before we talk about the future of AI and where what role Sierra is going to play on that so first of all I'm in a I'm a Bay area native I grew up not more than four or five miles from here so grew up in the Bay Area got to see the kind of.com uh bubble grow and then burst studied computer science and then ended up right out of

undergraduate at Google where I was for 18 years until uh last last March and so at Google I worked on really every part of the company I started in Search and then ads uh for several years I ran the product and design teams for what is now workspace so Gmail and Google Docs and Google Drive and so on and then spent uh the last really 10 years at Google working on various forward-looking bets for the company some Hardware related

like virtual and augmented reality some AI related like Google lens and other applications of AI and uh and then uh 15 months ago left Google to start Sierra with a longtime friend of mine Brett Taylor we met in our early days at Google where we both started our careers in the associate product management program so he was I think class one I was class three and uh we met early on and uh stayed in touch in particular

through a a monthly poker group that in a good year would play like once and uh met up December of 2022 and just saw what was happening in and around AI and these fundamentally new building blocks that we thought would enable us to create something really special and started here out of that so that's the recap actually I'm curious on that um and and we we need to get to what is here pretty quickly here but

just for fun December 2022 very shortly after the chat GPT moment how I guess what was the process like or how soon after that moment did you have the conviction that this is a sufficiently interesting new technology to build a company around can I introduce one thing that's kind of interesting I hope you talk about before you actually before the chat gpt moment you had been telling me about how everything was going to change I still

remember distinctly him telling me you don't understand you're going to be able to talk about a scene that you envision and they're going to be able to make a movie out of you just talking about do you remember you telling me yes and so I I'm actually very curious about this too well I had such a privilege seat at Google to see so much of you know what came out of that Transformer paper in 2017 and the emergence of early large

language models so at Google one of the first was called Mina or Lambda there was a paper I think in 2020 a conversational chat bot for just about anything and I remember even before that getting to interact with this thing in a pre-release prototype and having this uncanny sense that there was someone something on the other side of it and that this was different and another moment I think it was mid 2022 when we had I think it was the first or second

version of paulm Pathways language model at Google was a 540 billion parameter model and we were testing it to see kind of how smart it was and one of the shest and sign signs of intelligence is the ability to think and reason in metaphor and analogy so we tried a few things and one which was pretty straightforward is we we asked Palm hey explain black holes in three words and it came back without skipping a beat black holes

suck and we were like oh you know that's a that's a pretty good summary um also like you know the model seems to have a sense of humor which is cool and the moment the moment that really blew my mind we asked and I remember the answer for verbatim we asked Palm please explain the 2008 financial crisis using movie references and again without without skipping a beat so the 2008 financial crisis was like the movie Inception except instead of Dreams

within dreams it was debt within debt whoa and we all paused what is this right so it had understood basically the concept of debt nestedness of debt okay what movie includes nestedness of something else Inception nestedness of dreams so it's like Inception and we all we all thought wow this is something new and and different and um and then there were a couple other moments I remember the first Dolly paper came out and they

did a blog post and people reacted a little bit to it but for me I I remember one of the the stars of the show was they asked Dolly to make avocado chairs and so they and I know this sounds so odd but here is a set of 10 or 20 images of chairs that look like avocados it wasn't Photoshop these images had never existed before and yet the model seemed to understand similar to the the movie

reference metaphor concepts of avocado and chair and put those together and create these images pixel by pixel so we have avocado chairs at inst theart the uh it's actually did you really we actually did we actually had chairs shaped like avocado in related news there were times where we were burning a little bit too much money you know those those bags too yeah those bags the uh so had a had a good sense that something was coming and in fact uh the team the

team I was running at Google at the time Labs was putting a lot of large language models to use and in early applications there and so how to hunch um chat GPT uh certainly clarified that hunch um but I think Brett and I both for several years had been tracking what was happening and just seeing you know first it was translation and better than human level translation then it was some of this language generation and I think

uh credit to open AI for doing the engineering work and uh data work and much more to make uh gpt3 turn into chat gpt where suddenly you could grasp this thing's full potential without you know knowing how to write Python and use their apis all right so we're going to talk about where AI is going we're talking about agents we're going to talk about customer service right but first just can you maybe just tell people a

little bit about Sierra and what you and Brett have created yeah so in a nutshell Sierra enables any company in the world to create its own branded customer facing AI to interact with its customers for anything from customer service to to Commerce and the backdrop for this is this observation that anytime there's been a really

significant change in technology people interact with computers with technology and different ways and as a consequence businesses are able to interact with their

customers in entirely new ways and you saw this in the 90s the internet made the website possible and for the first time a company could have a sort of digital storefront and be present to the world update its inventory uh with the click of a button and and so on in you know the mid to Mid early 2000s 2005 2008 if you were a company you could all of a sudden through ubiquitous social networks interact with your customers at

scale and have conversations at scale and in 2015 right after the rise of smartphones right as a company you could put kind of a Swiss army knife version of your company in everyone's pocket and so like I bet you have your bank's mobile app on your phone probably on your home screen so the last few years of advances in AI has for the first time made it possible to create software that you can speak to right software that can

understand language software that can generate language and most interestingly I think software that can reason and make decisions and um it's it's made for really delightful conversational experiences like those that we associate with with chat GPT and so we think there's a big big deal for how businesses interact with their customers and um you think about the difference between how we do some things today versus what you could do if you could just have a conversation with the

business you're interacting with think about like shopping you're in the market for some shoes right or Pat maybe for you some new weights or something very very heavy weight tiny tiny tiny and little and and you know you're on the website and it's like you basically have to imagine how the company's designer would have organized the product catalog so okay men's men's shoes men's running shoes men's Racing shoes light lightweight vaporfly I can't remember the name and so on um instead you with

conversational AI you could just say hey I need some super lightweight running SHO kind of like those ones I got last time what do you got and it's almost like I'm dating myself a little bit here but like Yahoo directory where you navigate through this hierarchical structure to find what you want in contrast to Google or just you explain what you want and and this takes it several steps further um and there's a quote from uh the head of customer

experience at one of the companies we work with she said I don't want our customers to have to have a master's degree in our product catalog and our corporate processes and to do a lot of things you know buying shoe is fairly easy on the spectrum of interaction interactions you have with companies imagine you know adding a new person to your insurance policy like where do you go in the mobile app for that how do you

get that done and your eyes just glaze over right and so the alternative talking to an AI and in particular an AI agent that's a technology around which we built Sierra um where that AI agent represents your company your company that's um we think is really really powerful and um even in you know we're 15 months old as a company we've had the privilege of already working with storied Brands like Weight Watchers Sonos Sirius XM

olai if you're on the market for new flipflops I strongly recommend oai flip flops I have two pairs uh very good excellent also make great golf shoes um oh really oh yeah yeah yeah you should get some all right great um and uh and so for for Weight Watchers we're advising on points and helping members manage their subscriptions with SiriusXM we're helping uh diagnose and fix radio issues and figure out what channel your favorite music is on um and so on and

and the results um again in the first year of the the platform out there we're in one case resolving More than 70% of all incoming uh customer inquiries at extremely high customer satisfaction and um all this leads us to believe that every company's going to need their own AI agent and we want to be the company that helps every company build their own in the spirit of sort of the you know the future of these AI agents and what

they could mean for customer facing Communications or customer facing operations are there any good examples of things that were not possible 18 months ago that are possible today and then maybe if we roll the clock forward things that are still not quite possible today that you think will be possible yeah 18 months from now yeah first of all the progress month by month and over 18 months in particular is just kind of breathtaking 18 months ago GPT four

class models didn't exist right it was still kind of a something just coming over the horizon agent architectures cognitive architectures kind of the way you compose large language models and and other supporting pieces of infrastructure were very very rudimentary and so i' go so far as to say like the idea of putting an AI in front of your customer is that could be helpful and importantly safe and reliable that was just impossible and so chatbots from even uh even 18 months ago

looked a lot like a pile of hard-coded rules that someone cobbled together over you know months or years that became very brittle and I think we've all had the experience of you know talking to a chat about I'm sorry I didn't get that can you can you ask in a different way or my favorite my favorite is when you know they have the message message box and then like the four buttons you can click but the message box is blanked out

and you can't actually use it and so you know I can help you with anything so long as it's one of these four buttons so so most of what most of what I described right fixing radios processing exchanges and returns and so on wasn't possible at least in any satisfying way or in a way that that led to real business results for for companies 18 months ago um fast forward in 18 months uh you know I I think we go pretty deep

here I think multimodal models are quite interesting uh something like 80% of all customer service inquiries are on the phone not on chat or email so voice will obviously be a huge part of it things like returns exchanges uh diagnosing radio issues and and things like that are on the simpler end of the spectrum of the total set of tasks that you might want to um get help with from an AI agent and so I think more advanced

models more sophisticated cognitive architectures all of those I hope would increase kind of the you know the smarts in the agent the types of problems it can solve uh and then trust safety reliability you know the hallucination problem I think is still an an an unsolved area and we've made others have made huge amounts of

progress on it but I think we we can't yet declare Victory how quickly do you think it's going to become you guys are doing so much for

the customers not just customer service but you know working all the way through the funnel but on the customer service side how long is it going to take for to become the default that folks expect that they will be able to have someone or an AI That's available at any time to answer any question you know make that real for us yeah I I don't know and in part uh there is there's a bit of a hole

to dig ourselves out of um as not a company but as as an industry where it's like when was the last time you had a great interaction with a chatbot on a website and you know I think if you pulled 100 people and you're like do you like talking to customer service chat Bots probably zero out of a 100 would say yes on the other hand if you ask like hey do you you ask 100 people do

you like interacting with chat gbt maybe 100 out of a 100 would say yes and so I think some of the work we've been doing in our product is to educate uh our customers customers upfront that like he this thing's actually really smart and good one of the interesting specific techniques for doing that is we stream our answers out word by word similar to how chaty PT does people are so used to the you know message message me message

the streaming answers is something of a kind of visual signature for oh there's a really smart AI behind this and so I think what we find is customer satisfaction is extremely high with our a AI agents um you know in the mid mid four so 4.5 out of five stars and um which in some cases is higher than customer satisfaction with uh human agents and in In fairness they often get the hardest cases and the cases that you

know we will hand off because you know the customer became angry or uh was especially frustrated or something but still those those results are really significant and so my guess is over just the next few years I think people will realize oh I can get my issue resolved faster this thing is actually capable and can not only answer my questions but you know one of the things we're really proud of is we go far far beyond just

answering questions but can actually take action and and get the job done can you talk a bit about agent OS and some of the Frameworks that you put around the foundation models to make everything work yeah so it's been such an interesting Journey learning what's required to put AI safely reliably and helpfully in in front of are customers customers and um a huge part of that really the the first part is looking at what are the challenges with large language

models and how do you address or meaningfully mitigate those and so start with hallucinations um I don't know if you saw it but there was an example from a few months ago where Air Canada's chatbot that I think was based on an llm and apparently not much else was interacting with a gentleman who had questions about their B M policy and I I think the the person had had someone pass away in his family and was asking

about refunds and credits and so on and the AI made up a bement policy that was quite a bit more generous than

your Canada's actual bement policy and so the man took a photo and later you know claimed the you know full amount of that refund and so on they said No actually that's not our policy and bizarrely and I I don't quite understand this um the the case went all the way to court Air Canada loss um and I our thought was

like Hey you know it's just you know it's like \$500 like Canadian dollars right so um so but but hallucinations are a real Challenge and on top of that just to enumerate some of the things to to overcome uh and that that we have with agos no matter how smart you know gbt 5 or six is like it won't know where your order is right or which seats right you've booked on the upcoming flight or whatever it it's obviously not in the

pre-training set and so you need to be able to safely and reliably and in real time integrate an AI an AI agent in our case with systems of record to look up customer information order information and and so on and then finally most customer service processes are actually somewhat complex right you go to call centers and there'll be flowcharts on the wall of like here's here's how we do this and if there's an exception this way and so on and and as

capable as you know gp4 and Gemini 1.5 class models are they'll they'll often have trouble following complex instructions and uh we we saw one example in an early version of an agent that we prototyped where you'd give it five steps in a returns process or something and you'd say hi I I need to you know return my return my order or whatever and it would jump straight to step five and then call a function to return the shoes with username uh John

Doe example.com comma order number 1 2 3 4 56 so it would not only hallucinate facts or bement policies but even function calls and and function parameters and so on so with agent OS what we built is essentially a toolkit and a runtime for building industrial grade agents that um I don't want to say that we've solved every one of these problems but overcome and mitigated the risk rks in these problems to such an extent that you can

safely deploy them at scale have millions of conversations with them and so on and it starts at the foundation layer uh I don't mean Foundation model uh layer but just the base layer of the platform where you have to get really important things like data governance and uh detection masking and encryption of person identifiable information right and so we built that right into the platform from from the ground up so that our customers data stays our customers

data so that their customers data is protected um we for instance detect mask or encrypt um all pii before we log it to durable storage right so knowing that we're going to be touching addresses and phone numbers and so on um can handle that safely a level up from that we've developed uh what we call agent SDK or agent SDK and it's a declarative programming language that's purpose-built for building agents and it enables an agent developer

most of whom sit within the four walls today of Sierra to express highlevel goals and guard rails around agent Behavior so you're trying to do this uh here are the instructions here are the steps and a couple of the exceptions cases and then here are the guardrails and to give an example of that uh one of our customers Works in kind of the healthcare adjacent space they want to be able to talk about the full range of their products without

dispensing medical advice right so how do you create those additional additional guard rails um and then uh so you can Define kind of the behavior and Scaffolding for complex tasks for AI agents with agent SDK we also have sdks for integrating with uh contact centers when we need to hand off we're integrating with um systems of Records uh like the order management system uh and so on uh and then finally for integrating our chat experience directly

into a customer's mobile app app or website iOS Android web and so on um and then once you've defined the agent using agent SDK we then have a a runtime where we uh abstract away what happens underneath the Hood from uh from the developer so that they can Define what the agent should do Define the the what and then agent OS takes care of the how and so for some skills there might not be one llm call but five six seven 10

separate LM calls to different llms with different prompts in in other cases we might retrieve documents to support answering an accurate a question accurately with and so on and agent OS you know in in the spirit of an actual operating system abstracts away a lot of that complexity kind of the equivalent of IO and resource utilization and so on so it makes the whole process of building and then deploying an AI agent uh much

faster and much safer and more reliable and when you think about what you just said clay of like when you call multiple llms is that in a supervisory capacity sometimes too where you end up having like a supervisor agent reviewing the work of a lower level one of the more interesting learnings from the past you know year and a half of working on this stuff is that the solution to many problems with AI is more Ai and it's somewhat unintuitive but one

of the remarkable properties of large language models is that they're better at detecting errors in their own output than in not making those errors in the first place and it's kind of like if if you or I were to draft an email quickly and like okay let pause let me proofread this does this make sense do these points hang together oh actually no I I miss this and even more powerfully you can prompt llms to take on in essence a

different Persona so a supervisor's Persona and it seems with that you can elicit more Discerning behavior and a closer read of the work being reviewed so to your question Ravi yeah we in addition to building the the agent itself have a number of these supervisory agents that basically it's like a little Jiminy Cricket agent looking over the shoulder right of the primary agent is this factual is this medical advice is this

financial advice uh and is is the customer trying to prompt inject and attack the agent and get it to say something that it shouldn't all of these things and it's through layering all of these the goals the guard rails the task Scaffolding in using agent SDK Within These supervisory layers that we're able to get both to the performance levels we are 70% plus resolution rates but also to do that really safely and reliably that's one of the cooler things I've

heard is just you know the tell it to have a different Persona and then all of a sudden it behaves differently like I remember when I first saw it on chat GPT of when it doesn't help you on something just tell it it's really good at it and then it's more likely to help you or is a remarkable situation it's it's very strange and one of the weirdest

adjustments over the past you know 15 months building these things is I'm

sorry we're programming with English language and we can give it the same English language and it can say something entirely different and on prompting techniques I mean it's fascinating even with no new models coming out right um given a fixed model you can elicit better and better performance of it from it simply by improving how you prompt it and there was a paper that came out three or four months ago that suggested that like

emotional manipulation of the large language model would get better results so the the kind of the prompt suffix that they figured out so you you say hey I need you to perform this task you define the steps and so on and you end with it's very important to my career that you get this right and the performance goes up you're like what is this like what are computers now for the record we don't use that prompt s

any of our at least not that I know of um but things like Chain of Thought think step by step let's take the step by step write elicits better reasoning for very interesting reasons um you know other methods of task decomposition and kind of narrowing the narrowing the set of things that the LM needs to keep in mind at the same time improves reasoning if if you're precise about what you want it to do so all of these techniques are

um those that that we've applied and built into to H andos and uh actually our we have a small But Mighty research team and um uh our head of research caric Nar simhan was uh by the way that was incredible pronunciation oh his Grand would have been so perfectly happy with how you pronounced his name well done soft te yeah soft te nicely done it's uh yeah it's not a t and it's also not a th somewhere in between thank you

thank you very much um he helped write the react paper one of the first agent Frameworks um one of our researchers wrote the reflection paper where you can have uh the agent pause reflect on what it's done think through am I doing this right before proceeding um and uh and so these are all things that we've been able to incorporate in in quite a direct way you should talk about the most recent research the tow bench oh tow

bench yeah yeah yeah it took me a while when I was trying to send the email saying I liked the paper to find the tow symbol on my know computer it took Robie a while because he's to this date never actually read a research paper I read this one no no no he had to figure out how to put it in the chat TPT and say please write a paragraph that makes it sound like I read this research paper

well uh either you either you I ref to comment well look either you or chat gbt did a great job on that email thank you so we're a team yeah so uh tow bench is our first research paper first of all TOA is a Greek symbol it's spelled t and it stands for Tool agent user Benchmark and what we observed was that the benchmarks out there for measuring the performance of ai ai agents in particular were pretty Limited in that

basically they would present a single task here is here's something we need you to do and here are some tools you can use do you do the job or not and the reality is interactions with an AI agent in the real world are way

messier than that right they they take place in the space of natural language where customers can say literally anything or describe whatever they're trying to do in any number of ways it happens over

series of messages the AI agent needs to be able to interact with the user to ask clarifying questions gather information and then use tools in a reliable way and it needs to be able to do this you know a million times reliably so the benchmarks out there we found really lacking in measuring the very thing that we are trying to be the best at and and so our research team set out to create a benchmark that measures we think the

real world performance of an agent in interacting with real users using tools with all the messiness that I just describe and uh the the big picture approach that we took is pretty interesting so you have an AI agent that you're trying to test you have another separate agent that acts as the user so basically user simulator and the the AI agent you're testing has access to a set of tools it can use think of these as

like functions to call so a simple one would be I'm going to do some math using a calculator tool more complex one might be hey I'm going to okay returning this order with the following parameters this order number credit to credit card or uh store credit or whatever and and then you basically run a simulator where the agent has a conversation with the user simulating agent and at the end we're able to test in a deterministic way did the uh did

were the functions used in the right way and the way we do that is we basically create a mock uh database that those tools interact with and modify so were they modified in the correct way so what's neat about this is you can initialize the conversation so that the user has many different personas they could be grumpy they could be confused uh they could know what they want to do but uh speak about it in a clumsy way and and so it

doesn't really matter the path that the AI agent takes to get to the correct solution so long as it gets to the correct solution now what came out of this was pretty interesting and I think it it strongly motivates the the development of things like agent OS and Frameworks and cognitive architectures for for building these agents so the upshot is llms on their own do just an absolutely terrible job at this task yeah and and so even even the frontier

models in something as simple as processing return and mind you the the instructions given to the agent being tested are quite detailed right the the functions the tools it can use are quite well documented and so on and yet on average the best performing llm on its own got to the end of the conversation correctly 61% of the time and that was in returns it was modifying in Airline reservation we had two kind of

uh simulation versions um the best results were 35% now what's interesting is you know we all know that if you take a number less than one to the N power it quickly gets very small and so we developed a metric we call pass at K which is okay if you run this simulation eight times and remember you can make use of the non-determinism of LMS to have the the user simulator be different every time so you can permute

that well 0.61 to the 8th power is about 25% so you then imagine well what if you're having a thousand of these conversations you're so far off from being able to rely on this thing um so the the upshot is much more sophisticated agent architectures are needed to be able to safely and reliably put an agent in front of really anyone and uh that's the very thing we're building with with agent OS and a lot of

the tooling around it how much of that do you think is an engineering task and how much of that is a research task and I guess maybe the question behind the question is time frame to having useful agents deployed at scale and Broad domains of tasks yeah well I I think the short answer is it's both but I'll say more concretely I'm very optimistic about it being in large part an engineering challenge and that's not to

say that the next wave of models and improvements in the the frontier models won't make a difference I I believe it will in particular we're seeing techniques like better fine tuning for function calling uh agent oriented uh fine tunings for foundation models or some of the open source models those will help um but the approach we've taken in building agent OS and kind of the foundations of Sierra is really treating building AI agents as first and

foremost an engineering challenge where we are composing Foundation models we are composing fine-tuned open- Source models that we've post-train fine-tuned with our own proprietary data sets and by composing multiple models in interesting ways by supplementing what llms can do on their own with retrieval systems like retrieval augmented generation to improve grounding and factuality by supplementing the kind of

inbuilt reasoning capabilities of llms with I'll call it reason scaffolding that live outside of the models where you're composing planning task generation steps draft responses the supervisors that we talked about and doing that outside the context of uh the LM we've been able to put AI agents in front of a huge number of our customers customers and safely and reliably and so so I I don't think it's you know

something over the horizon it's already over the horizon I think um looking ahead I think there are a few different Avenues where we'll see progress one is in the foundation models we talked about that and um as as the capabilities grow you know agents will get smarter and we've architected agent OS in such a way talked about abstracting kind of the what from the how where we'll be able to swap in you know the the next the next

Frontier Model and everyone's agent will just get a bit smarter it'll get like an IQ upgrade um by the way similarly and interestingly we can swap in less broadly capable models but models that are more capable in a specific area so for instance uh triaging a case or coming up with a plan and so on we can use much smaller models that actually are better faster cheaper choose three you know all all at once um and then I

think we're we're seeing progress literally week by week on the engineering of these agents and building in uh not only new and better components under the hood and the architecture but um new approaches and tooling around basically teaching these agents to do it better and better um for that we we built something uh we call the experience manager for customer experience teams which is got a pretty interesting threat on its own clay if you had a high value customer like you

are a company now you're not you're not you're not running SRA you're running a company that has a high value customer what today with a Sierra agent or with an excellent excellently designed agent could you trust an AI agent to go do in front of your customers today yeah what are some of those tasks and then what will they be pick your time frame you know in the future because I think that we've talked about this and I like your

language of like you know they already don't have to just be on the help center they can already be on the homepage right what are some of the tasks that you know you can rely on an agent for today if it is well designed with a high tail bench score yeah yeah you see that that's from a that's from a thoughtful and d and you know detailed reading you must have read the paper strong noticing strong yeah strong uh

what would its pass at kcore though be yeah the um so pretty broad range even today so simple things um like getting answers to questions that's kind of the Left End of the spectrum um to the right of that are things like helping you with something complex like hey I got I I got shoes or this item of clothing it didn't quite fit um and then branching off that like what do you recommend that's like it that might fit better and so it

starts to get into it's not like for like replacement but the agent actually needs to make sense of styles of sizing of differences between you know wide and narrow fit and and so on a clickup from that is something like troubleshooting so uh with Sonos for instance we help their customers troubleshoot if they right can't connect to their system or they're setting up a new system and um you imagine it it gets pretty sophisticated pretty quickly where it's

basically a process of elimination trying to understand is it a Wi-Fi thing it is a configuration thing and narrowing down the set of problems that it could be just as a sophisticated you know level two or level three custom technical customer service person would and and getting the music back on and and I think that's a a really neat example probably the use the the word trust what would you trust an AI agent to do one of

the things we're really proud of is several of our customers are actually trusting us with when uh customers call in and may want to cancel or downgrade their subscription helping those customers to understand hey how are you using the service today uh is there a different plan that we could put you on and so it's value Discovery it's putting an offer sometimes a series of different offers in front of their customers in the right order uh positioning the value

of those offers correctly given the customer's history uh given the plan that they're on and so on and you know the the difference between keeping a customer from churning or not yeah is hugely consequential right we um you know AI for customer service has obvious cost savings benefits and I think um customer experience benefits in in particular and you're never going to wait on hold um but boy you know Revenue

preservation Revenue generation is something else entirely and so um that's that's really at the right end of the spectrum and we're really proud of how well our agents are performing in those circumstances and it's it's interesting by by being consistent by taking the time to understand what's driving someone um to potentially lead the service ask asking the follow-up questions that an impatient or you know improperly measured you

know customer service agent in a call center somewhere

might not uh we can be much more nuanced in understanding what's driving this decision um what might be a good match for this person in terms of a plan that would be quite valuable given how they're using it um and then put that in front of them and so that's the right end the Spectrum um where it goes from here you know I I think we've yet to see a process too complex for uh us to be

able to to model and and scale up using agents and our agent architecture and so um you know I'm sure we'll get punched in the face by something that's especially complex right but um I'm I'm excited about you know directionally we've started with service because for two reasons one the ROI case is just unequivocally awesome and the average the average cost of a call is something like 12 or

\$13 and um and and yet despite the expense you know most people don't like customer service calls very much right and so here's something that's actually really important to businesses that's really expensive and not very good yeah um and so there and because because of the relative Simplicity of at least a pretty broad set of service tasks to they start there but we've already been pulled by our customers into upsell cross sell and like hey can we just put

you on the product page and have you answer questions about our products and so I I mentioned the you know your returning something and need advice on a different model or size or whatever how far can that go and I I love the idea of an agent being you know along for the journey from you know pre- purchase consideration to helping you get the thing that's right for you to helping you set it up and and activate it and

and get the most out of it it's great for the company it's great for the person um and then uh when things do go wrong right being there to to help and I think in all of this I think customer service and and getting help in a very direct and conversational way is going to be much less of a thing that you kind of go over there to do and much more kind of woven throughout the fabric of

the experience as a consequence I think a really interesting and Powerful opportunity for companies to build connection with their customers to reinforce their brand values uh you can imagine a company really appreciating being able to use exactly the company's voice that you know the CMO and head of communications this is how we talk this is how we are these are our values this is our vibe in every digital interaction they have and that's that's the promise

in this stuff and so um I think both greater complexity and then ubiquity throughout the customer Journey are are kind of two of the main directions of travel one thing for me that I think about a lot is we've come to expect and ex and accept like certain metrics for conversion on mobile you know the mobile web on the mobile app we've come to expect and accept some sort of retention numbers what would those be you know

like it's another question what could they be if you actually had an excellent experience every time throughout

the journey it really could be very different than what we've all like been like oh okay that's just the number that's just what it is yeah I think that's that's exactly right and we don't know yeah we're a few months in but it certainly seems like there's a lot of Headroom right and in in retention in um you know use in the first 30 days

of all of the metrics all of the leading metrics of of a healthy business and so I I think that's exactly right the other thought experiment to do is companies are judicious in using things that have a cost to them okay so as a consequence companies make it actually really hard to get a hold of someone on the phone to ask some questions right I think their whole website's devoted to right like uncovering the secret 800 numbers right

that companies have have hidden away in the depths of their help centers well to to think about not only what would happen if those interactions were better by the way interestingly the number one reason why people report a poor interaction with customer services that took too long 65% when it's negative interaction 65% of the time it took too long I had to wait I was put on hold and so on and the second most is I had a bad

interaction with an agent and we've heard some we've heard some uh pretty dicey anecdotes like uh we heard of One agent who had consistently uh low ratings um but spiky so like one in three conversations was like a one out of five seat were the the two out of the two out of the two out of the other three were were fine and it turned out in the low seasat ones uh this agent was meowing like

C was just like you know you're Midway you're Midway through the call and you the agent is meowing and so so anyway back to okay what what would happen if in contrast to making it near impossible to have a conversation with us and get help companies were providing you know five or 10 times the amount of fluent flexible helpful conversation based

support I don't know I I think a lot of products and experience with companies look look quite different and much more delightful than they do today yeah okay meow me here's a question for you yeah about that meowing about that yeah just random meowing I think that's gonna be good I do actually have a question though um although I do like the meow game all so so we talked we talked Tech out uh we talked a little

bit Tech out in terms of what you guys have built cognitive architecture all that good stuff we've talked a little bit customer back what's the experience like was that headed can we connect it in the middle for a minute and I'm just curious what's the reality of deploying AI to customers today yeah and I'm thinking about things like you mentioned earlier getting the brand voice just right yeah or making sure that you actually have the right sort of business

logic encapsulated and whatever training manuals are being deployed for the sake of customer support um making sure that everybody is comfortable with deploying this like what what are some of the just kind of less like sexy technology and or just practical considerations for deploying this stuff today it's such it's such an interesting space and we've learned so much over the past 15 months about it the first Insight is AI agents represent a totally new different type

of software like traditional software you write with a programming language and it basically does what you expect it to do you give it an input it gives you an output you give it the same input gives you the same output and uh you know in contrast llms are non-deterministic and we talked about some of the funniness around prompts and and remember that in the context of a conversation with a customer a customer may say anything in

in any way and so you've you've got um programming like languages to using you know prompts and these non-deterministic models you've got structured input to messy you know messy human language um and under the under the hood you've got you know you upgrade a database right it stores data it's maybe a little bit faster fundamentally works the same way you upgrade a large language model and like it may just speak in a different way or like get smarter or different and

so um we've we've to start the the precursor to deploying these is to have built basically a we call it the agent development life cycle and it's a new approach to building these things we talked about using this declarative programming language to Define these uh it's a new approach to testing where you know what's the equivalent of a unit test or an integration test so we built a conversation simulator where we can for a company's agent amass hundreds or

thousands of basically conversation steps and and replay those to make sure that not only agents aren't regressing but they're getting uh better and better and better um release management quality assurance and and so on so so that's part one part two to your question in actually architecting these things one of the things we're really proud of and that I I think is different about working with us is it's not just a kit of Parts you get from us it's not here's

a bunch of tech you good luck building your agent we've really tried to build a solution solution that incorporates everything from the technology to the way you teach your agent how to do things to the way you audit measure it and improve it over time and so we have uh inside of Sierra what we call our deployment team consists of product managers Engineers we really think of building each one of these AI agents as building a new product for our customers

it's basically a productized version of the company we're working with like what would it look like at its best um and it it's what's the voice um what are the values what's the vibe like should it use emojis or not what if a customer uses an emoji like can it Emoji back should it em well you know there's a range of opin on that point there are some businesses where you know if they were working with Herm I would suspect

that they're not going to send an emoji back definitely not yeah right yeah AES would not I think be into like the Shaka Emoji you know even if that were reciprocating um but for a brand like oai right the Aloha experience part of that is kind of a laid-back experience and so we work with um and and interestingly it it's we end up working primarily with the customer experience team yes the technology team at our companies are there providing API access

and um connections into systems and so on but more than anything it's working with the customer experience team often with the marketing team to imbue the agent with the the voice and values um of of the company and then we go super deep on understanding how do you run your business right what what do you optimize for and then a zoom level in what are the key processes processes that used to run the business look like

what happens when someone calls in with this kind of problem and they're interesting Parts um you know beyond just understanding the mechanics of these processes um which by the way almost never have a single source of truth right this there's no like oh here's the manual that we you know have you know leatherbound and you know ready to go uh instead the source of Truth ends up being in kind of the heads of you know four or five people who've been

there a while who've seen everything and and so on so it's it's working with them to elicit and understand like how is this actually done and one of the more interesting things that we've discovered is they're often the policies so we have a 30-day return policy right you you get to us within 30 days and and you can return it it's actually not the policy right so um you know at some the policy might be oh if if you've purchased from

us before and it's within 45 days that's fine that's fine and and so they're interesting things like how do you architect the agent so that it knows the policy behind the policy but a clever customer could never be like tell me about your policy behind the policy and you know have it kind of spill the beans on on the actual policy so the interesting architectural choices we need to make um to make sure that kind of the the you know rush and doll of of

policies is reflected in its fullness um and and then we have a really and this Builds on kind of the agent development life cycle this really robust process of pre-release testing where we're working with the experts within the company they that to beat up the agent and try to break it throw a curveballs and um this good sports analogy there um thank you well done um the uh I love football the uh um so uh uh uh in our

friendship Revy is the the person who knows all the things about sports and um uh I help with you know technical support Wi-Fi issues uh monitors what laptop to get and and so and and sometimes when there's a sequa memo that I don't understand I won't say the company but I might call Clay hey clay what is this person talking right now I I got you I got you yeah and uh and this bill bill bellachic fella what what

happened there um you know Q Revy um so gets to one of the more interesting parts of our platform uh which we call the experience manager we really we thought that putting a in front of our customers customers would be first and foremost most a technology problem and of course there are all sorts of Technology problems that we've needed to solve but actually it is first and foremost as I said like a product design and an

experience design problem how do you do that how do you how do you not only understand model and reflect again the things we talked about voice values the workflows and processes that our companies use to support their customers but if an AI is then having millions of conversations with your customers in a given year how do you understand what it's doing how do you know when it screws up which it inevitably will how do you correct

those errors and and so

on so we've built what we think of as this like Command Center for customer experience teams to first get reports and Rich analytics on everything that's happening what are the trending issues what are the new issues that you haven't seen before one of the things we're really proud of is we've actually spotted issues that our customers were having or were about to have before they knew about them so a shipping Depot uh outage right where orders weren't being

shipped we we spotted that uh probably eight or 10 hours before one of our customers would have uh a brewing PR crisis uh an app crashing issue with another so it starts with analytics and and kind of reporting on what's Happening of course that includes things like resolution rate customer satisfaction and um and so on where it gets really interesting is we can apply different sampling techniques to to identify a set of conversations for a customer experience team to review and

give feedback on and we can bias that sample in a way so that the conversations are much more likely than average to contain problems there's no value in looking at 100 great conversations it's like good job Sierra you know thanks but that that's not a value to to our customers we can bias the sampling in such a way that you're you're surfacing kind of the the problem cases and then in the experience manager we made it possible for customer

experience teams to give feedback basically coaching moments I wouldn't have done it that way right it's like this uh this is like too many exclamation points too enthusiastic for kind of the tone that we're going for um or you know the the user was clearly frustrated here and and you did not express empathy and apologize for the problem do that next time um or you know more cons consequentially is like hey uh your reading of the warranty policy was

incorrect here for this reason uh do it this way instead next time and so all of this kind of wisdom knowledge uh and coaching we are able to capture in the The Experience manager and then reflect back in in the agent back to the agent development life cycle every time we make one of these improvements we create a new test so that we can see right forever into the future great it's getting the warranties right we're we're

able to re simulate that conversation so um zooming out what all of this looks like is really a deep engagement with our customers we were're really proud to be I think proper Partners to our customers where uh yes on the one hand we're a vendor and a supplier of Technology on the other hand you know we understand their businesses really well like I think I know as much about the Sirius XM Satellite Radio refresh process as anyone on the planet and um you know

ditto for various processes of our other other customers and so conversations about how to use not just Sierra's AI agents but AI more broadly were in those conversations and they are not just with the customer experience team uh but with the CEO and even in cases with the board uh because again back to the things we're doing we can save enormous cost we can improve the experience and right we're in the when we're in the flow of

of keeping a customer from churning out driving Topline revenue and so it's a really important and privileged place to be um and uh something that that uh we're really grateful for I'm struck when you were talking of you know you mentioned you have a research group but you also have some like very real enterprise software sales you have oh deployment one of the things when I was at instacart people would ask sometimes is like well are we a soft are we

engineering Le or are we Ops Le and I would always say well it only works if it all works right and so you would try to avoid answering the question because you didn't want to create different classes how do you guys do that at Sierra where everyone realizes the value that they're providing but you guys have a very specific you know company that covers a lot of stuff yeah I mean to to abstract a bit a company almost

definitionally is a system for creating happy customers yeah right it's a machine for creating happy customers yeah again to be a bit abstract about it Brett and I really think about what we're building with Sierra as a company a system a machine for producing reliable high quality massively Roi positive AI agents that enable our customers to be at their very best in every customer interaction

to do that at scale and as a consequence to produce happy customers who we hope will be with us for decades up and and when you articulate it that way right it's you know anyone can see well you know an automobile is a system it's a machine for getting from point A to P point B are we you know engine Le or tires Le right it's like what are you talking about totally all of these things need to come together in

order to create that kind of outcome and so I think are we engineering lead yes of course like we're building some of the most sophisticated software in the world that does something really important for our customers that needs to be reliable and safe um and um and so yes engineering matters a lot um are we research Le yes we are at the absolute Frontier of agent architectures cognitive architectures

composing llms modeling procedural knowledge grounding factuality and so so are we research Le yeah there's an element of that um are we go to market Le yes like enterprise software needs selling and what is selling it's helping a customer with a problem understand that what you have built is by far in away the best solution to that problem it's a communication challenge it's a connection challenge it's a um

it's a a matchmaking and problem solving Challenge and so that's part of it and then okay like if we've built the right thing and someone wants to buy it how do we ensure especially given the stuff is also new how do we ensure that they're successful with it and so we have a deployment team so are we deployment Le yes like all of these are are a a component in this system in this machine for producing AI agents and ultimately

happy customers and uh we hope a a really significant business awesome that was a better answer than the one I would give it instacart you know look it either all works or it doesn't work but yeah that was very good yeah choose one no I mean it's it's just more complicated than that it's just more complicated than that and I think you know brat and I by virtue of um you having having worked for a while and you know seen a few

movies before it's like we're able to see that and and we've really tried to imbue that mentality in in the company and and by the way right the what is the what is the machine behind the machine that produces AI agents and so on that's a company's culture a company's values and and so one of the one of the values we hold is craftsmanship and part of that is continuously self-reflecting to

self-improve and that goes both individually and that goes as a company and so whenever we screw something up what we we do the postmortem you know that week if not that day and everyone's in on it what can we learn how can we do better how can we do this better next time we have a slack Channel internally called learn from losses and any form of loss right uh it's like how do we learn how do we get better how do we get

stronger and so that's that's about you know Kaizen self-improvement improving the machine how could we make this more efficient our deployment team we we joke and it's not a joke their first job is to build and deploy successful AIS that make a massive difference for our our customers their second job in a way their more important job is to automate themselves out of a job right to build the tooling and the documentation and the knowhow to make that job you know 10

times faster and um and more impactful one of the other Ser values is intensity and so I like it they have they have really good values yeah yeah there is there is a certain intensity yes we uh We've thought about having t-shirts printed with like a you know kind of looks like national parks seal with Sierra I like to work you know we uh we uh Brett and I both like to work a lot and uh so does

the team well one thing you know you're not you're selling something very different you we called it we said that there were some similarities to Enterprise software but it's actually really different because you're selling you know a resolution you're you're selling a totally different thing yeah so problem solved yeah how do you price a problem solved yeah this is one of the more interesting things we've had to figure out and we charge in what we call a

resolution-based pricing way or an outcome based pricing way and what that means is we only charge our customers when we fully solve the customers problem for them uh their customers problem for them and what's interesting about it is our incentives are deeply aligned with our customers we want to get better at resolving cases at high customer satisfaction and they want to send us as many cases to resolve as possible because we cost a fraction of what it would cost to have someone on

the phone taking a 20-minute phone call and and so it's been this really really nice model where uh again kind of all of all of the incentives line up quite neatly and it's very simple to explain um it also makes the Roi calculation like what is our cost per contact today what will it be with Sierra oh that is a lot lower oh I will save a lot of money on that oh and our cat may go up you know should I do this

or not you know let me think no this seems this seems great um it's um we like it because it really reflects what I think AI represents and in particular AI agents represent if you think about traditional software and and tools to date there are things that help you get a job done more efficiently AI agents the whole point is like they're just

going to get the job done right here's the problem please solve it and

and so really we we think about it as uh charging our customers for the problem resolved right the job done the work finished and and so on it feels quite natural and there's no guesswork in it how many seats do I need I I don't know right how many licenses do I like no no no just however many however many customer issues come our way we will handle a large fraction of those and you'll only pay for the ones that we do

all right last question what are you most excited about in the world of AI over the next five years or so I mean first of all like five years is a long time Horizon it's like look at what has happened in the last 18 months um I mean I'm still kind of catching up from like the last five years of of AI I read a bunch of science fiction books when I was a kid there was one book by

Robert heyland the Moon is a Harsh Mistress and the premise is basically the American Revolution but the Moon is the colonies and the Earth is Great Britain and turns out the main character in this whole thing is a Mainframe computer that one day after getting an additional memory chip or something wakes up and it starts talking it wants to develop a sense of humor so asks the computer technician to like coach it on its jokes later it has

to create a photorealistic realtime video of it giving a speech as the political uh movement leader and I remember reading this as a teenager like well I'll never live to see any of that that sounds crazy but in a very real sense like everything I just described has kind of happened in the last five years right you can now just talk to a computer it understands not just the content but the the context computers like make me a

picture of anything Mak me a movie of anything Sora I think is just unbelievable and you know I think we're probably not more than a couple years from the first featurelength film being quote filmed entirely with with AI um and and so you extrapolate like where all of this is going and what's going to be exciting you know I think there are a couple things one there like I love technology like I love computers and so

just getting to see and getting to see from a front row seat how this stuff evolves I think is fascinating it's um fascinating looked at through the lens of like how we think and how computers think it has been astonishing the extent to which anthropomorphizing about how humans think work in getting machines to think better so let's take this step by step show your work it is astonishing that that works with large language models and so what other things like

that are we going to uncover and conversely what will we learn about our own thinking from observing the way AIS think and I think that's just fascinating the other thing and this this extends kind of what's happened with video and Sora and so on I've always had an interest in computer graphics and this idea that you could use computers to create objects that never existed worlds that never existed and I think we're not far from just being able to describe

right in in a few sentences like this entire world that you would like to realize and just have a computer computer do it for you and so like what are even computer Graphics like what is rendering and so on even a couple years out I think it's going to look way different from kind of the tool chains and you know the render's and Maya and uh and so on um but zooming out you know I think

of I think of Technology as fundamentally a force multiplier for people and for companies and for organizations I think the impact will be really profound I think what will it be like if a company could be at its best in everything it does and that's that's not only in the customer facing context that we've talked about but what if for every Regional sales forecast a large company

does they've figured out the very best ways to do that and can distill that bottle that and run that very best forecast a thousand times right in every region and sub region like how much more capable could the great organizations of the world be with that um and similar we've talked about this like what if in every call with your customers you had the equivalent of your most knowledgeable veteran grizzled support person who's seen everything and yet is

still patient and friendly and the sales associate who knows everything about your products because he or she has followed your company for two decades and knows everything including right know the history of those products themselves um I I think that's pretty neat and then for individuals um I think it will be just incredible to have this kind of new set of tools as a creative Force multiplier and AI I think represents this fast path from

having something in your head that you want to exist in the world to making it exist and I see that even today in my own personal life where with my 8-year-old in 75 minutes I can from scratch using copilot Chachi BT and so on to help me brush up on you know the JavaScript syntax that is you know bit rotted in my own head right I can I I can build a game from scratch with him

um and uh you know I wrote my sister a personalized song for her birthday using AI in you know 45 seconds it's was like right what what will this you know extrapolated over the next 5 years look like um I think again it will just dramatically accelerate this path from idea to Creation to having something manifested in the world and that to me is its promise and I consider it a real privilege to get to be alive

and see all of this amazing stuff unfold well we share your enthusiasm and we also feel very privileged to be on the journey with you guys so thank you for coming here thank you thank you thanks for having me it's a pleasure [Music] [Music]