Podcast Transcript - Qualcomm Driving On-device Generative AI to Power Intelligent Experiences at the Edge

00:00:12 Peter

Hello everyone and welcome to another episode of "The Counterpoint Podcast". I'm your host Peter Richardson and joining me today is a very special guest from Qualcomm to talk about on-device AI and actually, on-device Generative AI more specifically.

So please welcome Ziad Asghar who is senior Vice President of Product Management and Head of AI at Qualcomm. Welcome to the show. Ziad, great to have you with us today.

00:00:38 Zaid

Thank you for having me, Peter. Excited to talk to you.

00:00:40 Peter

Great. So generative AI has been one of the, you know, hottest topics in the last few years, actually. Now we've discussed different aspects of the technology on previous podcasts, and we'll put some links in the show notes in case listeners want to find out about those. But today, though, we want to dive a little bit deeper into how to run generative AI on the types of devices that people interact with every day as opposed to supercomputers in cloud data centers and I guess this is really where Qualcomm has some very special capabilities, which yeah, hopefully we're going to get into in a bit.

But just to kind of kick us off, can you help us by defining some of the terms that we're going to be using here today, I mean, for example, we're mainly going to be talking about generative AI, but how should we think about generative AI versus machine learning or other types of AI that people might be familiar with?

00:01:35 Zaid

Sure. I think I'll start more with you know, where we have been as a technology and you know as a technology and as an industry, it's always

been about telling the computers when you do this and if else then right. It's been a very clearly defined way of interacting with the machine. What's changed is that we have this umbrella term which is Artificial Intelligence which really encompasses machine learning and deep learning, and then Generative AI, and essentially the idea with Artificial Intelligence or AI, is to be able to give machines the capability to mimic some of the human behaviors by allowing them to be able to learn and then to behave in certain facts in certain ways, the way that humans do.

And then the term beneath that you can think of it as deep learning where basically you are able to take large amounts of data and allow the machines to be able to learn and essentially apply those techniques and automatically do certain behaviors.

And then I guess the step below that is the ability to be able to take neural networks and apply them to these use cases and that's machine running where the machines are able to take vast amount of data and apply neural Nets to be able to do artificial intelligence like use cases. And then Generative AI as a subset of essentially those techniques where you're actually applying machine learning and deep learning techniques to generate content.

You're generating text and images and video and 3D in the future. Where essentially this used to be the realm of human behavior, right? Generating content and generating art. Well, now I can go and tell a machine. Hey, here's a text, and make me a picture that essentially depicts exactly what the text says.

So, that's the really amazing thing about generative AI. It can generate emails for you. It can generate code for you. It can generate things that in the past required direct human attention. So, of course, there isn't a day that goes by in news, in tech media places where people are talking about new techniques and new ideas on how to be able to do these Generative AI use cases even better.

00:03:32 Peter

Hmm. OK, that's very helpful. Thanks. And you know, of course, smartphones have been running AI workloads for quite a few years now and you know, for example, scene recognition in photography, for example, something like that or maybe even, you know, optimizing the radio propagation characteristics, you know, those sorts of things.

Those use cases don’t go away. They don't get replaced by generative AI, right? So, they'll still be around.

00:03:56 Zaid

Absolutely. So, I mean we've been, you know, using Artificial Intelligence as a technique in multiple technologies, right? I always call AI a horizontal technology. What I mean by that is you can actually apply it in camera, in audio and speech and video and modem technology. And we've truly been leading the pack over here where we've applied it for example, to camera to be able to allow you to see in the darkest of rooms and yet be able to get, you know, a good picture out of it or to apply it to improve image quality.

And then, of course, we've taken it all the way down to where we're applying it to modem technology as well to be able to get better signal in worse of conditions, you know, in the hardest of channel conditions as well. So all those use cases, we're applying it to audio, video, camera, all of those Previous technology stays and continue, but now the real I guess excitement that comes on it that now you actually have use cases that are directly using Artificial Intelligence to be able to add value to be able to add new experiences to the products that we are so dear to us like smartphones that we use on a very, very regular basis.

So yes, all those use cases continue, which means the device of course needs to do even more than what it was doing in the past from an Artificial Intelligence perspective. And you know what that means, Peter, is that essentially whoever can do more AI processing with lesser power is going to come out looking great in this new revolution era.

00:05:16 Peter

Ok. And that, that actually kind of brings me on to the, you know, one of the focal areas of Generative AI that a lot of people have been talking about and we see this in, in terms of results from you know, companies that are creating the chips that go into the data centers for training these big, large language models. And that's just kind of the scaling issue. So you know I think in general, the larger the model, the better its performance, although I guess there are some counterexamples to that, and if we scale the model size, you know the number of parameters involved, there's maybe a kind of a consequent need to increase the scale of the computing resources used to train those models, and maybe also by implication, maybe you can disabuse me this notion.

Perhaps the results required to interact with those models to get some output from them and you know, and you look at some of the powers that are being, you know, the compute power that's being used to train these models and it's kind of staggering, right? So, you know, even the top ten supercomputers from a couple of years ago are not up to the task of training these models.

So how should we think about this scaling issue? Are we sort of on a trend where you know the models just get larger and larger and larger or you know, are we sort of reaching a point where actually we're now kind of beginning to look at different options and different ways of doing this?

00:06:24 Zaid

Yeah, I think it's a very interesting question. So I mean one of the things like you pointed out, we are seeing that new models are getting bigger. But I think they get bigger and then at least the way I look at it, they start to become somewhat smaller. So, let's take for example of a text image model like DALL-E.

Right, DALL-E started I think it was about 10 billion parameters and then now what's being utilized extensively stable diffusion that's about a billion parameter and it still does an amazing job of, you know, those text to image like use cases. So one thing that is happening is that when you add a new modality at that point in time, the model is pretty large.

But then with all the whole tech community working on these technologies, what is happening is that the algorithms and the way the models are constructed and the data is curated to be able to create that model. All of those things are getting better and so with time, those models start to actually become smaller in size right. I think if you take another example for text to text, you know ChatGPT 3, it was about 175 billion parameter model.

But if you look at the Llama and Llama 2 like models from Meta. They are in the 7 or 13 billion parameter range and in some of the metrics actually you know, Llama does fairly well when compared to ChatGPT 3. So again, a similar trend like we talked about in the text to image case. So, I think in the beginning you can argue that it does become large the model, but as the algorithms and as the data curation gets better or as you make the model more you know demeans specific. Essentially, you're actually able to reduce the model size, and that's where I think it's a very interesting trend

and that's why it allows us to be able to do a lot of these amazing use cases on device in the future.

And then I would like to spend a little time talking on the training versus inference difference as well, right. So, of course, the model gets trained ones and these large models like you, you know, take a lot of energy and power to do, but honestly, inference is not that cheap either. So if you look at it from the perspective of some of the news that come out recently, open AI talking about, you know, it costs them about $700,000 a day to be running just ChatGPT 3, I think that's the other part of it. And like you were pointing out, it's a lot of computation that you have to do.

But that's where I think the promise is right. If you could do that large amount of computation on the device, which is what we have been driving as a company, and I personally have really seen the promise that has, and I think that really will give us and the consumers and the industry many advantages by being able to do this on the device.

But the other aspect that I also want to point out, even though we will not be doing full training on the device anytime soon, but there are some interesting techniques that are coming up like low-rank adaptation and all where you could for example take a model and be able to fine tune it on the device and it's actually reasonable and it's not as computationally intensive, which can really curate and optimize the experience for a given user on the device extremely well.

00:09:16 Peter

Hmm. Ok. And you mentioned Stable Diffusion, DALL-E. Now you know I saw the demonstration that Qualcomm had at MWC back in February where you were running, Stable Diffusion on a Snapdragon device. That was, yeah, I didn't have any access to the cloud. I think it was in flight mode and it was producing pretty impressive results. It was impressive. It was kind of neat. But what do you see is the kind of day-to-day applications for generative AI on, say, on a smartphone device. What do you think will be the main ones that we'll see initially and then how those will develop overtime?

00:09:46 Zaid

Yeah. I mean, I think stable diffusion than what we showed at Mobile World Congress really showed I think the possibility. It showed the fact

that you know you can't do these use cases on the device. But you know we are in touch with each one of our customers. There's a huge slew of amazing use cases that they want to bring on the device which we believe we'll be able to do just this year and early next year as well. You'll see a lot of those.

From a smartphone perspective, right, it's a very interesting device because again, this is where the data is being generated to great. Right. If you're talking about, for example image, this is where your camera is sitting and you're doing the camera. You're capturing that image on the devices. So for example, one great use case that people are looking at is, you know, we could take this picture exactly like you know, we have here with your background, Peter, and you could just tell the device to put you in front of a very different background and it can just essentially generate that background using Stable Diffusion or ControlNet like models which we can actually by they were on today because since Mobile World Congress we're actually showing ControlNet running entirely on the device as well and that's by the way and it's a 1.5 billion parameter model, so even larger than what Stable Diffusion was, and you know, you can essentially do that on the device.

And it's a great use case whether you are a young person using the device or somebody else. But the one that really is very exciting to me is a true Virtual Assistant likes to use case, right? We've talked about it for the longest time, but now you can just envision, right?

I can say something into my device we use a model like Whisper to be able to change that to text and then that text can be going into a Llama 2 like model where you essentially run the query and you're able to actually generate an answer that's back in text domain and then you apply that to a text-to-speech where you essentially are able to generate speech from that text.

And you could then even put that into what we call like a face model where a person is able to, you know, move the lips and everything on your screen, say exactly what was basically asked about. So essentially what you have created is a pretty complete end-to-end virtual assistant model and now you can envision the possibilities, right? If it changes the way that we interact with the smartphone.

If it changes the way that you know we used all the different apps underneath in a day-to-day basis, right? I like today write to you. I always

give this example that let's say you are setting up a reservation at a restaurant. You know you can go to an application to be able to find the ratings of restaurants and then you will basically go to that restaurants app perhaps to basically schedule that reservation.

And then you actually figure out and go to the navigation app to figure out how do you get there. Well, your virtual assistant could tap into all of those underlying apps and essentially do all of those things for you. And I think that's where it becomes really interesting as to how we interact with them. Of course, for people like us who use the device for a lot of work, you know, you can do, for example document summarization or you can do for example, you know you can have a draft of an e-mail generated by the phone for you and they can just review it and add it and modify it and send it out quickly. You can have it basically you take a video conference call with your phone and you can generate the transcript for it, for example.

So really I think the use cases are many and as we interact with our customer partners that are coming up with new ideas that we are working with them to be able to launch on the upcoming products.

00:12:53 Peter

All right. I mean, we're talking here a little bit in a sort of a either or sense or a binary sense, you know either. In the cloud or on the device, but I think you know if I listen to what you're saying, we're more likely to own it with some sort of hybrid model, right. So where some of the workload is done on the device, maybe some of the inferencing, maybe some of the learning is done on the device. So to use your kind of personal assistant example, you won't, you know, I use this sort of term AI in a very kind of general sense, you want the AI to kind of learn about your habits and things that you like, you know, preferences and so on, but you may then require it to interact with, you know, other networks. So you need some sort of cloud interaction at that point. So, I guess we're really talking about this sort of hybrid model, right?

00:13:34 Zaid

Yeah, it's a good point. So essentially the way we are envisioning how AI works in the long run is basically what we're calling hybrid AI. What that means is that you know, as our devices become a lot more capable, which we are driving right now, you'll automatically be able to do a lot of this processing on the device. But additionally, what you can also do is that,

let's say, you know, a model size above a certain level. You can essentially say I'm going run this in the cloud, for example, or you can also have a certain degree of interaction between the cloud and the edge to be able to still continue to offload the cloud even further.

For example, you can do some of the token processing on the device and then the remaining in the cloud and be able to offload it by having perhaps a smaller version of the model run on the device and a larger version in the cloud. So, there are many techniques that we are looking at.

We're working with our partners to really be able to bring this hybrid division to fruition, but a big component of that is of course how much we can do on device. And there are many myriad of benefits of really doing that as well because just imagine, right, I mean when you do some of this processing in the cloud it's running on large graphics engines that are burning hundreds of watts of power and then you were there at Mobile World Congress and you saw we were running this use case on a handheld device and you know our team basically showed it the whole day and essentially it was we never even plugged the phone in the whole day.

So, you know we do it in milliwatts of power compared to hundreds of Watts of power. And I think that really changes the way you are able to bring generate AI to the masses.

00:14:58 Peter

Yeah. And I think that's a very kind of interesting, but also an important point from a sustainability point of view. Cause I think one of the criticisms or concerns about these huge, large language models is the amount of power that they consume when they're being trained, but also the inferencing part as well. I wanted to touch on a bit more about two aspects really. One is hardware, one is software. So, Qualcomm is perhaps best known for the Snapdragon line of application processes.

So, you know how is Qualcomm, leveraging its capabilities on these silicon chips to really kind of unlock generative AI here? I mean, you know, you have different compute elements within the chip, so are they used individually or are you using you know some combination of the CPU, GPU, how should we think about that?

00:15:43 Zaid

Yeah. We think really about AI and the hybrid sense, even on the device. Or heterogeneous sense I would say. And the way that we have set it up is we have multiple engines on the product to be able to do AI processing like you pointed out, CPU is an engine that can do AI processing. So can the GPU. And then so can the NPU or Hexagon processor that we have on our product, and then I'm responsible for driving the one technology roadmap at Qualcomm so essentially we create these core technologies like the Hexagon processor.

It's really designed in a way that it's essentially maps to the architecture of a neural network. You have scalar components, you have vector components, and then you have the matrix components within the hexagon processor. So, if you look at it from the perspective of power consumption. For any sustained use case, Hexagon processor is going to be able to do that AI workload at far lower power than CPU and GPU.

But the ease of use in some cases as to how familiar the developers are to the CPU programming model versus the GPU programming model allows a lot more developers to be able to start to use our silicon fairly quickly by leveraging CPU and GPU as engines.

We of course have very good run times like Snapdragon neural processing SDK and even tensor flow light and other well-known frameworks in the industry to be able to use Hexagon processor as well, but just a number of people who are familiar with how to leverage those capabilities on a CPU and GPU is larger and I think that's why those engines are also extremely important. And so the way we're thinking about this, Peter really is depending on the use case, there might be smaller networks. They're actually quite fine to be able to run on the CPU or those networks that might have spotty performance, right? You run it for a little bit, then it goes away.

But I think when we talk about sustained use cases, which is use cases like text or text and others, you would have to be able to run it on a specialized engine like the hexagon processor really to be able to get the right power consumption in the device that's not plugged in and that's why we have set it up this way such that we're giving our customers and partners and developers all the flexibility to be able to use the right engine. So our hexagon AI engine directly, you talked about the software is set up to be able to leverage the CPU, the GPU and the hexagon processor equal.

And you know, we've spent quite a lot of time creating our Qualcomm AI stack. And you know, the way we have created it is such that it's actually

common across all of our product lines whether smartphone or AR, VR or PCs or automotive or IoT. And essentially what you can do is and the power that it gives our customers is that you can really just, You know, build it once and deploy anywhere.

So, you build a great model, let's say for text-to-text generation. Well, you can create it for the smartphone and then take it and run it in the automotive context or run it on IoT. And I think that power that comes with Qualcomm scale is just amazing, right? Because at any given point in time, Qualcomm has more than a billion devices that are deployed. We bring in a few 100 million devices every year, so the scale that we can bring, you know with our hardware and software. We really are able to democratize these use cases in a way that no other company can in my opinion.

00:18:52 Peter

Excellent. Thank for that. And maybe that's a kind of a good segway into sort of talking about the partnerships that you have with you know other companies within the ecosystem and the developer community. So yeah, clearly you're kind of working with a whole host of different players. Can you talk a little bit about where? you know where you see that going, both from, you know, sort of the large players, but also some of the smaller developers.

00:19:20 Zaid

Yeah. So, we definitely are working with pretty much every you know level in our ecosystem of course. So you know we have you know we're working with Meta for example on bringing Llama 2 like models which are great text-to-text models and we are able to bring them into smartphones into PCs, into automotive, any product. And essentially that uses our technology.

We have a pretty close relationship with Microsoft. And at Microsoft Build, we actually showed many of the capabilities for the developers also to be able to run models like Stable Diffusion on a PC. And you know in the PC space, Peter, it's really interesting.

We are really the only vendor that has a specialized accelerator for AI processing, the other vendors do not have that capability. So like we were discussing on the last question, they are forced to run those cases on CPU and GPU.

And as you go into the realm of generative AI that's running nonstop, that's sustained use case rather than, you know, spotty use case, you cannot get the power to be right when you're running it only on CPU and GPU.

So, that's another advantage that we get so working very closely with Microsoft to bring, you know, Co-Pilot like use cases to the Windows Snapdragon product line.

We are working, of course, with all of our customer base very closely. We of course are very much engaged with Google as well on all the different vectors of how we are able to bring these use cases to the device. At the same time with our customers partners, we are coming up with new use cases that we are meeting with on a regular basis and they have these amazing ideas and want to make sure that we are able to bring them on to our upcoming products and then from a developer perspective like I mentioned at.

You know, we have essentially shown how you can do it on PC. Similarly, we have some very good frameworks like the Snapdragon Neural Processing SDK. What if you're a developer who's using Tensor Flow Lite or Python, We have ways to be able to leverage that capability and be able to use the full capabilities off the on the device and allow the developers to be able to get all of that benefit in real use cases.

00:21:29 Peter

Ok. But something that we've not really kind of touched on yet is privacy and security. Whether we're running these generative AI models, whether it's on the device or in the cloud. And I guess one of the attractions of doing this on a device is you can confine the data to, you know, some secure area on the device itself and not send, you know, sensitive data up to the cloud where we may have less control over it. You know, how should we think about that? And does Qualcomm have a particular perspective on this topic?

00:21:57 Zaid

Yeah, I mean, this is exactly why you're seeing on device generative AI. We are getting so much traction. I mean exactly what you said there Peter right? If let's say you put in a query and let's say as time goes by, we're more reliant on these generative AI-like clients that are running on your

device. And let's say you're putting in personal information in there, you really don't want that going to the cloud.

And you know in past days, we've seen news about people putting in code snippets into some of these generative AI-like tools to be able to create and generate code, and you can imagine that some of that code can be proprietary. So, the real promise with generative and why I strongly believe that you know the center of gravity for generative AI processing is really shifting from cloud to the device more and more every day is because of some of this privacy concern.

But there's actually a lot more additional benefits of why you want to do this on the device, right? So we talked about a little bit of the power benefits in the past where you know, like we said, large engines running hundreds of watts of power. But think about what's happening, right? What is happening now is the model size is like we talked about which used to be hundreds of millions of parameters. Now there are billions of parameters, so the model size is increased by 10 to 100 if not more. Second factor that's coming in is that there is a lot more use cases using these generative AI capabilities. You got text, you got video in the future what I see is from trend perspective you'll have multi-model AI coming up, so you'll have not just text to text, you'll have text to video.

Or you'll have image to video or you'll have image to image. So the capabilities will become multi model rather than just a single modality in what we do. So, a lot more use cases are coming up. There's a lot more people are excited about generative way use cases. So, they're coming in and using these capabilities. So, when you look through these three factors that are really multiplicative in nature. You really cannot scale the cloud capability to be able to run all of this in the cloud, and that's why generative AI needs on device to be able to reach its full potential. That's the only way to do it and that's why, you know, we're super excited about it because we can actually do these use cases.

We can do, you know, you know Cristiano has mentioned we can do more than 10 billion parameters this year on the device, so the capability that we have is really massive. So that's like the second major benefit of being able to do it on the device. So like we talked about power, power is basically a proxy for cost and essentially you know you are going to be burning quite a lot of cost and you know I talked to customers and they tell me look, I want to do this use case and I have it running in the cloud, but I'm limiting

it. I'm not letting people do as many queries as they want because well, every query is costing me money.

Well, if you do it on the device It doesn't cost money, right? So it's really very powerful. But you know the most compelling benefit that I see is even more than that, which is really an optimized and curated experience for each and every person.

What do I mean by that? You know your device has your contextual information available on the device. It knows kind of your likes and dislikes. It knows kind of your age, it knows your queries from the past.

We plan to use that contextual information to be able to do exactly the techniques that we talked about earlier, like Laura, which is low-rank adaptation and others to be able to fine tune the model for each and every user, but you know what that means. It means that you will have generative your experience that is far better on device than it can ever be on cloud.

Because you don't want to send that information up to the cloud. Again, it's very private information to each and every user. So the benefits of on device AI are just manifold. And I think it really allows us to be able to put generatively use cases in the hands of each and every person that has a smartphone that has a PC that has an automotive device, right. It's just his amazingly powerful in that sense.

00:25:38 Peter

I wanted to sort of touch on some other variety of edge devices so you know you mentioned there. You know, smartphone PC auto that there are some others maybe, but if you are creating tailoring as a model for one user, how portable will that be across different domains. So smartphone is the thing that I sort of carry with me all the time. But when I launch my PC am I going to be having the same experience there as on my smartphone or if I jump in the car am I going to have the same? How transportable is this capability?

00:26:07 Zaid

Yeah, just like I mentioned earlier, with the Qualcomm AI stack, we basically have this ability to really build once and deploy anywhere. So as you are taking it from one Snapdragon product to another Snapdragon product, you will have the ability to be able to take that model from handset

domain to the PC domain. And what we are creating is this ability for devices to really work together.

Much more seamlessly and you can literally take a model to say from the other side too. You can take an IoT model and bring that to smartphone side also, but like you point out, right that model that's specialized for you, that model that specialized for me.

I can actually take that from smartphone and you get a new device on your PC and you take that model that's already ready to go and you basically give access to that and the new device is able to actually use that model that's best for you. So you know what happens with time that model on the device becomes more and more specialized to each user, whereas the model in the cloud is more and more distant from it. And that I think is the true promise of on device AI.

00:27:03 Peter

OK. Well, beginning to wrap this up now Zaid. So maybe you know thinking about the next five years, let's say, what are the you've talked quite about some of the use cases that obviously you're excited about, but are there, you know two or three that really kind of you know you can't wait to see emerge?

00:27:21 Zaid

Yeah, I think I really like the fact that the technology that we create provides human benefit. It provides social benefits, right? So one of the things that we've been looking at is how we can apply generative AI to healthcare or how you can apply generative AI to education domains, how you can provide, you know, apply generative AI to, for example, elderly care.

So one of the you know what I am quite excited about is like if you look into the future, if for example take the example of a classroom and you know you are able to understand how well each and every student in the classroom is really able to digest the material that's being given to them. And that could be, you know, by virtue of a camera or something that sits in the classroom.

You take in that information and with generative AI, you are on the family able to change that Lesson plan based on exactly how well each and every

student is getting the material. This is the power of it, right? They have a tablet in their heads. That tablet is looking at it and they can see that the student is somewhat, you know, not getting the material you slow down the pace of delivery, it's an optimized education delivery plan for each and every student. You can never have that happen, right? So generally I really opens up these amazing, you know, capabilities in the education domain.

Now and similarly in healthcare, if you have an elderly patient, you know you can have a device in their home that again gets all the information from the sensors that they might have on their device. The you know, the blood pressure information, heart rate information and the temperature information at the same time, it's able to converse with the patient and essentially see if they might be incapacitated in any way, they don't have to wait to go for the next appointment before which the therapy can be adjusted. You can actually do it on the fly. Of course, working with the healthcare provider. So you can actually do that. I mean, in automotive domain, you can have the experiences of a, you know, a Knight Rider like experience, right? And essentially you can be talking to your car, you can be having that amazing experience of it being a true assistant to you.

Alright, you're coming into work and you tell it well I got these five things I got us solved today. So can you please schedule these meetings? Let's get those people to send out this information by this time today, right, all of that gets done. Your commute becomes a productive time for you, right. And then of course, if you look further out in time, there are even more exciting use cases you can think of. You know, one of the domains that we have which is AR, VR really lends itself extremely well to generate AI because you're creating this.

Virtual world with virtual objects around you and you can actually take that information and cues from what the person likes and generate all of that. Well, what if we could have an avatar of you that actually goes and represents you in meetings that you are not able to go? And actually sits in a Teams call or in a Zoom meeting and is able to converse and say exactly what you're going to say because it's trained on your data and essentially is able to represent you and then come back after the meeting and say, well, you know, these are the questions I could not answer. So please make sure you send out an e-mail addressing those ones and gives you a running commentary of what happened in the meeting. The possibilities are endless.

And really, we think we're really scratching the surface at this point in time with Generative AI.

There's just a lot more to come, and I think what we can do on the device, the curation, with the optimization for each and every user, is really going to evolve the way we entertain ourselves, the way we work and just really change the productivity of each and every person.

when I also say that. You know, it's really a very good force multiplier is the way I think about generative AI. Because it will take away the mundane, it will take away the routine work from you and really free up humans to be able to do what they do amazingly well, which is essentially novel ideas, innovative work. So I think that's where I see generative AI really taking us in the future.

00:30:55 Peter

Well, I think that's a fantastic place on which to end this, thanks again an excellent and very insightful discussion. So glad you could join the show today and share your knowledge with us and with the audience. Yeah. Thank you very much.

00:31:09 Zaid

Thank you for Having me, it was great talking to you.

00:31:11 Peter

Thanks and for our listeners, thanks for tuning in. As always, you can head over to Counterpoint Research or your favorite podcasting platforms to listen to this and previous episodes. And if you'd like to know more about on-device generative AI and talk to our analysts, please do drop us an e-mail at press@counterpointresearch.com. And with that, thanks very much.

See you on the next one. Bye now.

