Deloitte.

User Friendly: Season 8, Episode 3
January 2025

The impact of on-device AI: Enhancing performance, privacy, and user experience

Host: Hanish Patel, Managing Director, Deloitte Consulting LLP

Guests: Vinesh Sukumar, Head of Generative Artificial Intelligence and Machine Learning Product Management, Qualcomm

Baris Sarer, Artificial Intelligence Practice Leader for the TMT industry, Deloitte Consulting LLP

Hanish Patel: As we all know, artificial intelligence, more commonly known as AI, is evolving rapidly, and we're on the cusp of a revolutionary shift with on-device AI.

This groundbreaking technology brings AI processing directly onto devices like smartphones, PCs, and more—fundamentally changing how we interact with technology.

But what exactly is on-device AI? What are the benefits, and how can key players across the ecosystem adopt this transformative technology to enhance performance, privacy, and user experience?

Today, on the *User Friendly* podcast, we will dive into the evolution and the impact of on-device AI. Joining me to unpack these exciting developments is Vinesh Sukumar, head of Generative Artificial Intelligence and Machine Learning Product Management at Qualcomm, and Baris Sarer, Artificial Intelligence practice leader for the Tech, Media, and Telecom industry at Deloitte Consulting. Vinesh, Baris, welcome to the show.

Baris Sarer: Great to be here.

Vinesh Sukumar: Thank you. Good to be here.

Hanish Patel: Alright, given recent times, when someone mentions AI, minds immediately jump to Generative AI. But there are various types, as we all know. And Baris, I'd like to open up with you for maybe a bit of a primer on the different types of AI and from there, let's kind of talk a bit more about what is on-device AI, and what does it exactly mean and how does it differ from, say, traditional cloud-based AI?

Baris Sarer: Sure. So, GenAI is a step in the evolution of various AI and ML [machine learning] technologies [that] burst onto the scene a couple years ago, obviously; but if you look back at the development, research, and technology evolution behind it, it goes back decades.

Now, what is on-device AI? It's a whole different question. It's an exciting shift, or rather maybe an emerging marketplace or approach, in how we perform AI processing. Instead of relying on cloud servers, which require data to be transmitted back and forth, on-device AI performs computations directly on the end user's device, whether it's a smartphone, laptop, or any other edge devices.

And this brings a lot of different benefits. Obviously, since you're doing it locally, it's near real time, reducing the latency associated with compute. And then you're not really dependent on an internet connection to be able to use AI. The data is local. In addition to offering faster response times, it offers greater privacy too.

Because your data is not getting transmitted in the public domain, and then you have better service availability. You're able to use your AI without the internet connection, as I mentioned, which is essential for applications that require uninterrupted performance, like autonomous driving or real-time translation. What I would want to point out is on-device is not an either-or proposition. There's a right place to use high-performance cloud servers to compute AI, especially all of the training that's going on today that definitely requires very substantial compute capabilities, that at least today, only exist on cloud.

But it's a combination of leveraging the compute on-device, leveraging the compute on the cloud, and executing the right workloads at the right place, using the right model at the right time. All of which, in my mind, comes together as a new architecture, which I guess [is] increasingly being recognized as hybrid AI architecture in the industry.

Hanish Patel: So, firstly, thank you for that, and then I just want to maybe dig in a bit deeper as you talked about hybrid there and some of the benefits—like you said, just having it local, not needing an internet connection.

If we think about on-device AI and, as you mentioned, hybrid AI architectures, and they're considered to be somewhat an evolution on the AI landscape, can you dig in a bit deeper?

And I really want to turn to the both of you on this one around the main business benefits and the end-user benefits when we think about that kind of hybrid AI architecture as well as on-device AI.

Baris Sarer: So, service availability I mentioned when you do AI locally, you don't even need an internet connection, and there's so many different use cases across industries where that is critically important.

Privacy by design: You're keeping the data local and, as you know, privacy concerns got elevated over the past decades and with the advancements in AI—and the data-hungry AI—it's become even more top of mind for both consumers and enterprise users.

And then security, endpoint security offered by processing AI locally on the edge, reduced latency. And also, I've failed to mention cost reduction is an incredibly important element of it, and this is a little bit counterintuitive. Over the last decade, we've seen a huge shift from on-prem workloads to the cloud.

Now, what we're arguing with on-device and hybrid is we may see a little bit of a shift back or redistribution of workloads. And if you're a software company and you're deploying a lot of AI workloads to the extent that you can deploy some of them on-device, you're effectively externalizing those costs or the cost of running those workloads, and therefore there are some great benefits for software companies to get out of this.

Vinesh Sukumar: I'll probably add to what Baris mentioned here. When you look at on-device from an AI-enablement standpoint, to a large extent our focus of attention has been on immediacy or better quality of service, reliability, personalization, privacy, security, and cost. But there's always going to be instances wherein on-device AI may not really support all elements of it.

There could be instances wherein you want to look at distributed processing. The processing could happen at near edge, which could be your own personal workstation, or near edge, [which] could be within a premise-based server or completely on the cloud.

And the question really becomes at what point do you really make that decision to do a hierarchical shift towards where hybrid becomes extremely critical and extremely important. One such example would be in automotive environment there is something called a *shadow* mode.

A shadow mode is something wherein you want to understand, given a situation, an environment, how would a driver and how an autonomous agent in a car would act, and they try to mimic these actions. If the autonomous agent in a car equals the action taken by the user, then you don't have to share anything towards the cloud. In that case, things are looking good.

But there are going to be instances wherein the agent has taken a different action and the action was wrong compared to that of the user, or the user has taken an action, which was completely different and led to an accident; we record that.

So, the expectation becomes if you want to have some continuous improvement or quality leverages over a commercial deployment, you kind of share those elements in cloud, you optimize the models, and then you make sure that you do an over-the-air upgrade to continue to increase the prediction accuracy of these models.

There are going to be instances, especially, let's say, in enterprise environments—and enterprise environments could be in financial or health care markets wherein you want to take a certain action on your financial transactions, or you want to have some kind of a summary of your financial transactions for the last six months.

So, the intermediate request of translating, let's say, a voice-based command into a textual input, the textual input goes into a large model. The large model understands the intent of the user. Once the intent of the user is understood, it needs to take a subsequent action of collecting the data of financial transaction for the last six months and then do a summary of it, and that information is resident on the cloud.

So, in that case, you have a specific context, you have a specific request about the user that goes to the cloud, extracts the information, and then comes back to the edge. So, it's going to be a combination of edge plus cloud-based processing.

So, in summary, I would say hybrid is inevitable. The question really kind of becomes how much of a workload distribution would be on the edge and how much would be on the cloud, and that totally is dependent upon a use case.

It could be an automotive environment, financial transaction, health care, or even common consumer space, and depending upon the key performance indicators or experience indicators that distribution changes accordingly.

Baris Sarer: And Vinesh, maybe I want to add on top of that. As you think through the hybrid architecture, a critical component is going to be the orchestration of those workloads. Would you agree?

Vinesh Sukumar: Well, absolutely. It's important that this orchestration makes that decision. It is not predefined, in a way—because predefined is one way to go look at it from a hybrid standpoint. But the other element is, can this orchestration self-learn? At what point do you make that switch to a near edge or a far edge, or completely towards the cloud? So, yes, absolutely aligned with you, Baris.

Baris Sarer: Yeah, and to put a finer point on that, the critical thing will be when we talk about end-user devices, obviously we're talking about resource constraint environments that are used by end users.

So, to optimize between the performance of the system, the resources available to other end-user applications, and the complexity of the use case, we're talking about a fairly sophisticated orchestration that happens real time, managing both cloud and on-device resources. I think that's where the future is, to your earlier point, about hybrid AI is inevitable. That's going to be a critical enabler of that journey, in my mind.

Hanish Patel: I want to stick with where you both talked about the orchestration, and if we think about that, how is the market actually evolving to then support on-device AI? And specifically what trends are driving the adoption of endpoint AI across the various devices? I'll start with you, Vinesh, on that.

Vinesh Sukumar: A great question. I'll probably simplify this and kind of answer this in two phases. Phase one is kind of looking at the system approach. System approach is, what are we doing at a base level from a hardware perspective? What is happening on the software side, and what is happening at the system intrinsics [level] to really get to a much better user experience from an application standpoint?

So, starting at the lowest levels of hardware, hardware is where all the magic happens. The question really kind of comes back as, do you happen to have enough computer resources? Do you have enough bandwidth? Do you have enough memory? Key blocks to really get these AI experiences fully enabled.

When you look at various AI use cases, they either anchor on latency or they anchor on higher performance, or they anchor on better quality of service. If you are in a streaming kind of AI use case or the gaming kind of use case where you're running this AI experience on a portable device in performance per watt, where energy efficiency becomes extremely critical.

So, once you understand these key performance indicators, the question really becomes how do you architecture this from a hardware perspective to have the necessary investments in place to get it accomplished?

When you go one layer above that as a software, as someone said, every hardware architecture lives or dies by the software, and it's a true statement. The question really kind of comes back as, is the software easy enough to use? Is it portable enough to get you the performance you're looking for?

And do you have the necessary "knobs" in your toolkit, which could be libraries, could be your operators, to really eke out the highest performance that you're looking for, for any ecosystem practitioner?

Could be a developer, could be a research scientist, it could be an app developer—any of those kind of audience. And last, but not the least, is the applications. Now applications is where it gets a little bit of a tricky thing.

Herein, our applications collect a lot of data about the user, and then depending upon a static training model that's available, you provide a certain inference of the data that's made available. The question kind of always comes back to factors. Can this application be dynamic in nature?

In other words, if a user is asking for a certain response and the response is not good, can the application self-learn so that next time when the same application is invoked by the user, the same mistake is not happening.

That's a difficult problem to solve, but there's a lot of research happening in that space to get it done. So, there's a lot of excitement to really make sure these apps are self-learn and they get personalized to the user perspective.

Now, the second component of it is an agentic workflow. Agentic workflow is wherein the humans or any kind of common user does not have to specifically touch an application or invoke an application.

That's where I think the big talk is all about in the future is you happen to have a plain display, a plain intelligent machine, and based off using your voice as your primary input translates your intent into actions.

Those actions go invoke certain instances of what you're looking for and get things done. But again, I think these are evolutions that are expected in the future, but there's a lot of excitement on getting things done on the edge for sure.

Hanish Patel: Vinesh, as you covered that sort of evolution… Baris, what are the trends that then you are seeing through the adoption of that, across those various devices?

Baris Sarer: Well, to kind of pick up from where Vinesh left—and I'm going to apply a little bit of a Deloitte lens to it. We really opened our eyes, if you will, to the opportunity as some of these small language models started becoming available, maybe late 2023.

And then also the fact that some of these chips coming available earlier this year—we realized that it was actually technically feasible, and not only feasible but attractive, to start pushing some AI workloads to the edge.

And our journey with Qualcomm really started with a proof of tech where we decided to demonstrate the power of on-device AI by building a CRM-like mobile application and demonstrate that.

And the application worked like this: Two business people talking to each other. You use the application on the phone to listen to the conversation, capture the audio, translate that to text, and then—using a small language model running natively locally on the app—summarize it, identify the key action items, identify the individuals involved in the conversation, pull their contact information, turn this into an email, and capture the minutes of the conversation to summary and the action items in the body of that email. And when internet becomes available, email it to them.

And this was really well received, it's developing really fast, and we're seeing the traction in the software vendor landscape, and obviously to Vinesh's earlier point, there are so many different enterprise use cases as well. And as the market confidence develops and the solutions mature, we're going to see an uptick in the enterprise segment as well.

Hanish Patel: Thinking about what you said about it being a nascent market and things emerging, Vinesh, could you tell us a little bit more about what Qualcomm's approach is when it comes to enabling AI on the chip set and the types of AI models that can now run on these devices, given the sheer kind of early stages of this industry to a certain degree? It'd be great to cover that off for our listeners.

Vinesh Sukumar: Absolutely. When you look at Qualcomm, to a large extent, Qualcomm has been always synonymous with telecommunications like [the] 5G and LTE kind of landscape. But one thing is the AI journey in Qualcomm began about 10, 15 years ago.

Wherein the focus of attention to a large extent was enabling simple use cases around audio and (with time) into camera, because it was mostly anchored around consumer-based AI. And with time, [we] expanded that horizon into vision-based applications, which were traditionally around detection, segmentation, classification, image quality enhancements, elements around it, and got deployed across IoT [Internet of Things] streams—AR [augmented reality], VR [virtual reality], including handsets.

Now, obviously with time, this morphed into Generative AI. And to a large extent, Qualcomm's journey in Generative AI, I would say, started about three years ago, and we actually made an announcement of enabling GenAI.

We did a "myth buster" showcasing that creating synthetic images using text prompt can be done on-device.

To a large extent, at that point of time, there was a lot of craze in the market that when you look at creating synthetic images or synthetic text, we were able to create something very similar.

It took approximately 13 seconds to get it done. So, obviously there was a wow factor that this can be done on-device, but the quality of service was quite large.

And by the time we commercially deployed it, we were able to get this down to less than four seconds. So, over a period of, I would say, eight months, we reduced the latency by more than 200% give or take.

And this is possible only because we had invested the entire stack on both the hardware side and the software side to eke out every performance. And when you happen to have this foundational experience enabled to a consumer, you can use this foundation for doing a lot of stuff.

It could be using your voice or a text as an input, you could create a synthetic background on your handsets. You could do image-to-image transformations.

So, many such experiences we were able to get done using the vision portion of GenAI. We were also able to expand this into the productivity segment. The anchor point was how can you make the on-device analytics much more intelligent?

You could use with simple use cases, like if you happen to have a lot of numerical data, how do you come up with recommendations of what kind of charts actually make sense for the type of the numerical data you have? If it's scientific data or kind of an alphanumeric long-extension data, what kind of graphs actually give a pretty good visual representation of the conclusions? That was all done using AI.

You could also expand this into pretty much on human-to-human interactions, especially on meetings wherein you could look at live translation, live transcription. The transcription gets summarized into a nice meeting summary along with actions mapped to a speaker and then when these models have access to plenty of calendars, what will be the next sync-up meeting that needs to get done? Again, great productivity tool.

And you can go to the other extreme, too, from an on-device. Again, typical use cases would be in the automotive environment. Automotive environment kind of comes in two phases. One is the ADAS [advanced driver-assistance system], which is outbound; another one is infotainment within the car itself.

And here, one of the interesting use cases is you're seeing [that in] most of the cars that are being rolled out in '25 or even late second half of '24, you've got these displays that are pretty much in the front of the dashboard and all these displays have voice as a means of interaction.

It'd be really nice if I talk to my dashboard and say, "Exactly what is the problem?" It understands the environment, it knows exactly the history of the car, it tells, "Hey, the problem is A, B, and C, and do you want an appointment on so-and-so date because I have access to your calendar? Do you want me to get it done?"

So, these kinds of things really help. Again, you're identifying a problem, which makes the user experience better, and then kind of get into newer areas, which was never done before. And this could be some of the agentic work that we talked about wherein you can look at creating knowledge graphs to make it much more personal to the user.

I think these are good investment areas where AI is actually pushing for use cases, on the productivity side, content consumption side, content creation side, and plenty more.

Hanish Patel: You mentioned a number of use cases, like you said, from a vision perspective, productivity, and clearly even one could argue a safety perspective when you think of automotive.

Baris, I want to turn to you as well. In some of the examples that you're seeing, whether you've got any you want to share of just how clients are using some of that on-device AI technology in their real-world applications.

Baris Sarer: Yeah, absolutely. One I'm going to mention is with one of our contact center software clients. Their use case specifically for on-device was two parts.

They wanted to do translations, but real-time translations would have very stringent demand on latency, and they were struggling to get that type of performance out of cloud-based deployments and that's what got their attention in terms of on-device AI as an option. And obviously, the other attractive part was the cost component.

Another interesting deployment, and this is more internal to Deloitte but something I'm really excited about, we partnered up with our own IT organization to take a knowledge base of 3,500 typical IT problems that would ordinarily get folks like me on the phone with our IT help desk.

We created a knowledge base local to the laptop, put an enterprise RAG [retrieval augmented generation] solution there also on-device and a small language model, and with a very performant solution, were able to get 80% of our most common IT support requests handled by a bot running locally, powered by a small language model on-device.

That was one of the use cases that I continue to believe is really representative of the potential of the solution. And then the third one is memorable for me. I met with a software company.

They developed DJ software, and it was memorable because the use case was really interesting. They came to the conclusion that given where DJs sometimes played could be a beach or a basement or a place that you wouldn't take good quality internet connection for granted, then you want to still be able to do your AI-enabled DJ moves, let's say, whether it's effects or cuts and edits, and things like that that you do in a live setting. You want to be able to use an on-device model to enable those, even in the absence of an internet connection. That was one of the ones that stayed with me.

Hanish Patel: Thank you both for those examples. And as I kind of reflect on those, it's also kind of got my head spinning around just how fast things have moved on and where they could be going. So, I want to maybe end with a question that really kind of looks at the future.

And specifically, what kind of innovations or developments can we expect to see, when it comes to the realm of on-device AI? And how do both of you see this technology shaping our future? And I'll start with you, Vinesh, and then come across to you, Baris, on that.

Vinesh Sukumar: So, this is the million-dollar question that everybody is trying to invest on, but I'll give you my two cents. One of the biggest challenges you have in AI is AI lacks common sense. What I mean by saying that is the component of reasoning to understand a certain situation and make a prediction as a function of that situational awareness is lacking.

So, how do you infuse that, is one thing. There's been a lot of focus, there's a lot of investment happening, and can that be enabled on-device? In other words, we're trying to make sure that these systems which have built-in AI, can they be dynamic in nature? Can they understand the user? Can the personal and the emotional intelligence of the user be baked in along with the environmental intelligence?

And it cannot be static in nature, it has to be dynamic in nature and can it be done through self-learning? And that's a big topic. To a large extent, it cannot be resolved in a day or two, and there are going to be different phases.

The extension towards it was about agentic flows. The agentic flows always touch on how do you do device deployment and how do you transition that to hybrid? And as Baris was talking about it, the question really becomes is the intelligence in the orchestrator strong enough to make that switch between an edge, near edge, far edge, and a cloud?

And it can be seamless in nature when you happen to transition towards, let's say, in this case, to a cloud. Can the contextual information be carried along with it so that the user experience is seamless? And these are difficult problems to solve.

And last but not the least, there's a lot of stress on governance and safety and guardrails, and how do you really make sure that some of these policies are fully enforced? And can it be commercially deployable?

We're putting a lot more emphasis on translating these governance policies into what needs to happen from testing purposes, and then translate that into investments across the stack to really respect these policies and make sure that we can deploy it for both consumer and for enterprise applications.

Baris Sarer: I totally agree with Vinesh's sentiment on this. The other thing we probably need to think through is for on-device AI to be successful, it's critical that it offers a compelling value proposition to software developers, whether these are companies or independent developers and startups.

And if you look at the software market, most software applications are SaaS [software-as-a-service] applications that are made available through a web interface, and enabling on-device AI through a web browser is, in my mind, a critical unlock. And the good news is that we've seen some breakthrough in that space.

To me, it's a really important breakthrough because as we start enabling on-device AI through the web browser, then you're talking about gaining access to the larger part of the market, which is all SaaS solutions and then e-commerce applications that run on the web. Most enterprise applications today are web-based. It really opens up the largest part of the software market to on-device.

And this is something I'm excited about. Again, we're in the early inning, but the market is evolving so fast that I think maybe six months from now that's going to be table stakes, and we'll be talking about something else, which makes me really excited about this emerging market we're in.

Hanish Patel: So, I'm just reflecting on the conversation here, and it's amazing to hear just how on-device AI is frankly changing the game and allowing for real-time local computation on our devices and, as you both highlighted, be that at the consumer level or at the enterprise level, and just everything it's doing around productivity or just efficiency and privacy as mentioned before.

And as you both talked about what the future holds and some of the innovations ahead, it's pretty exciting to think about that potential. As it not only allows for less dependency on cloud in the example that both of you cited, the budding DJ, or whether it be alerts coming off your car appropriately, it clearly opens up a lot of possibility in many applications.

And with those ongoing advancements, when I think about, from a hardware perspective, the AI models themselves, it's clear that on-device AI is set to play a real key role in the future of technology, again, both for enterprise and for consumer.

And with that, I really want to thank both of you here, Baris and Vinesh, for joining me today to share your perspectives and really opening up the aperture of thinking for our listeners around just how AI has evolved, particularly when we think about hybrid and on-device and what that future potentially holds. And with that, to all our listeners, until next time, happy listening.

Explore more episodes of *User Friendly* at:

userfriendly.deloitte.com

Deloitte.