AI360 podcast

Season 1, Episode 12: On-device revolution: A deep dive into hybrid AI

Host:

Rohan Gupta, Principal, Deloitte Consulting

Guest:

Mark Szarka, Senior Manager, TMT, Deloitte Consulting

Rohan Gupta: Hello everyone, and welcome to another episode of AI360. I'm your host, Rohan, and I'm really excited to have Mark on as our guest today. Mark's been a longtime collaborator and friend, and I'm excited to have him on the pod. Mark, you want to quickly introduce yourself?

Mark Szarka: Thank you, Rohan. Happy to be here. Quickly, my name is Mark Szarka, AI leader here in the firm. I wear a few hats: first, is defining the way that we'll be doing work at Deloitte using AI at the core of our delivery. So helping transform our business as well as a leader in hybrid AI—and we'll talk about that here in a minute.

Rohan Gupta: So, Mark, I know just coming right out of the gate, you said "hybrid AI." What is hybrid AI, and what role do devices specifically play in hybrid AI?

Mark Szarka: Yeah, hybrid AI. So that is something that we've actually been playing with even since last summer, which is the theory around these models are getting more performant and the compute is increasing. So can you not only focus on bringing AI workloads up in the cloud, but down in the device? So we've been doing this since last summer, and hybrid AI is the theory of can we have a world in managing the world of interplay of distributed compute, so I can have enough workloads up in the cloud, but also start to bring down things down to the device. So you could actually start to realize benefits such as cost reduction, enhanced privacy, reduced latency—and really AI on devices is not new. There's been vision models. There's been speech-to-text models. It's been going around for a long time. But there's a lot more interest now with these more memory-intensive language models that we can start unlocking new use cases on devices.

Rohan Gupta: So you kind of hinted at this. What are some of the drivers for why hybrid AI is finally taking off? You said that we've had these models on device for a long time, but what makes this moment different?

Mark Szarka: Yeah. So there's a ton of investment and a ton of research going into not only can I focus on taking a GPU to be much larger and much more performant up in the cloud, but also the same things happening on the device. So can we actually increase and double the number of tops? So basically the number of calculations that can happen on a device year after year. And that same theory of Moore's law is applying now to the device level as well. So things that I couldn't do on a device last year… with what's coming out in the next couple of months, and even the years after, the compute will continue to increase on the devices, as well as the research that's going in from refocusing on "Can I train a large language model?" to actually "Can I focus on making the inferencing of that large language model more performant?" So things such as can I do compression techniques. So something like quantization, which basically means can I shrink the number of compute points or decimal points inside of a neuron inside of a language model? Can I make that smaller and smaller? Can I change things like model architectures so that the number of attention heads that occur in the model or the types of things such as sparsity, which is do I actually need all the layers inside of the model at any given time. So those are very technical things, but the focus on inferencing efficiency and increasing compute is making all of this happen—in addition to the fact that AI itself isn't really cheap, but ISVs [independent software vendors] and others are really incentivized to start bringing workloads down to devices so that there can be a balanced interplay between bringing feature richness as well as efficient economics.

Rohan Gupta: So, Mark, can you talk a little bit about what you're seeing in terms of early use cases and their benefits?

Mark Szarka: Yeah. So a few of those use cases are… they're obvious, and [what] we've already started to play around with is could we actually bring AI coding onto device? So could we actually take an IDE plug and then build those and actually be able to run code implementation? So like Python, JavaScript, Java, etc. and actually be able to do the inferencing locally? That's one clear, obvious one. I could also start to—we actually also built things such as IT service support, IT service management, on the device. So rather than needing to go out and service tickets and get you out to a call center, you can actually start to serve and manage PC support directly on your laptops. Those are just a couple of examples of productivity type of things. We can start to think about and extrapolate that to if I was a doctor, or if I needed to update patient information real time, I could actually just speak directly to my phone. It would update that onto the phone. Don't have to worry about the information leaving the device, but also keeps doctors more productive in things like field service. I'm out in the field, no connectivity, and actually ensure that those technicians maintain productivity throughout their entire jobs. So that. There's a whole range of industry things, there's a whole range of use cases, a whole range of productivity use cases. And that will continue to compound as we start to increase their features to.

Rohan Gupta: This sounds a lot like offline maps, but for all of AI, which is incredible. So, what are some of the success factors to start getting this right and maximize the value of hybrid AI?

Mark Szarka: Yeah, absolutely. So the number one thing that we think is, you know, important beyond anything is the orchestration of all of this. So thinking about the analogy of we went from data centers to the cloud. That all happened because there's a platform to compute on, right? Similarly here there needs to be the interplay in managing between the device variables. So there's a bunch of different chipmakers, there's a bunch of different models. Everyone's going to have a different range of devices. Rather than

ISVs needing to build for each individual variable, there will need to be the orchestration component of not only how do you provide and serve up that service to a ISV, or a user, or someone building; there needs to be the ability to actually serve that and manage all those variables in an intelligent way. There's going to be things like dynamic data, dynamic inferencing, etc. to be able to bring this all to life. And we see that as being a key to unlocking all of this.

Rohan Gupta: That's incredible. We asked some of our participants this a few times about predictions, especially for newer areas. I think you qualify as well with hybrid AI. So what are your three predictions for hybrid AI over the next couple of years?

Mark Szarka: It seems obvious, but there will be an increasingly disproportionately higher volume of AI features available to those companies, those users, those end users, or even anyone that has a higher compute device. Right? So there's legacy devices, but that increasing set of features and that will all be driven by orchestration. It will unlock who actually has access to those features because of the economics. Now, that ties to the second point, which is the interplay between the AI features from ISVs and others, as well as it will tie directly to the OEMs and device refresh cycles, because there will be an interplay there of the "chicken or the egg" supply/demand equation—we'll have to monitor that. But as we start to build in more and more, the demand for the increased set of compute on down the device will increase that refresh cycle. And then lastly, just the continued monitoring of the shift to the importance of inference. We'll start to see some really interesting changes in model architectures down for the smaller set of compute devices. Not GPU racks up the cloud, but down to the device, we'll have models that will have a quarter of the footprint, same sort of outputs. It's the interplay between power, speed, and accuracy to be able to balance all this. So it'll be really interesting, a very fast-moving space, but we're very excited about what's coming ahead of us.

Rohan Gupta: Fantastic. Mark, thanks so much for being on AI360. I'm really excited about your prediction, and I'm sure we'll have you back for a future episode.

Mark Szarka: Awesome. Thank you Rohan, thank you for having me here today.

Visit the AI360 Podcast Episode Library to learn more.

This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to deloitte.com/about.