Counterpoint Podcast Transcript: GenAI – The Next Big Thing in Smartphones

[Mohit] (0:00 - 0:33)

Hello everyone, thanks for tuning in to yet another episode of The Counterpoint Podcast. We have a super interesting topic today as we deep dive into the world of generative AI and the future of mobile technology. AI Phone is the new buzzword today, and we'll explore the possibilities, the challenges, and the exciting future of AI in your pocket.

My name is Mohit Agrawal and I am joined by my colleague Tarun Pathak who is the Research Director for Devices and Ecosystems at Counterpoint. Hi Tarun, welcome to the show.

[Tarun] (0:34 - 0:36)

Hi Mohit, thanks for having me today.

[Mohit] (0:37 - 0:57)

Thanks, and before we deep dive into the discussion, I think it will be good to talk about the definition of AI smartphones. I mean, if you look at AI smartphones, they have existed for over half a decade now. So, how do we differentiate between a phone that has AI features versus a generative AI smartphone?

[Tarun] (0:57 - 3:29)

Right. So, Mohit, good question. I think right now what is happening within the industry depends on who we talk to, there's a lot of confusion on how groups within the industry are defining AI or even a GenAI phone. So what Counterpoint Research is doing, how we are defining it, we are defining the GenAI smartphones as a mobile device that leverages large-scale, pre-trained generative AI models.

All these models we have been hearing, LLMs. But the key takeaway here is we are defining something to create content that is original, and that can perform a wide range of tasks that can lead from contextually aware tasks. We expect these devices to have multi-modal capabilities, which means they will be able to process text, image, voice, and even videos.

We have just heard about the latest update on that to generate a variety of output. And I think what you asked in how is it different from AI versus a GenAI phone, I think one of the key differentiations here is largely AI was meant to basically automate certain tasks.

Like you have tasks, you are automating them. Conversational AI largely focuses on understanding and responding to what me or you are inputting in that particular device. So, they are reacting to what we are saying.

But the GenAI, what it aims to, it is more like creational. What it is doing is it is creating an output that is based on diverse data points, diverse inputs, what we have been giving it to them. And the output is very different, right?

What we could have imagined on that. So that's the major, major difference. And, one thing that we are also doing at Counterpoint is we are not putting any hardware numbers on a GenAI phone.

So what we are also assuming that these hardware functionalities will evolve over a period of time. But at least these devices should have a performance which is similar to the current flagship smartphone to effectively run this GenAI model. So this is what we are assuming internally.

But this is one space that is going to basically evolve over a period of time as we analyze.

[Mohit] (3:30 - 3:49)

Thanks. So, let's look at a case of a smartphone wherein imaging has been enhanced using AI. We are probably erasing the unwanted people in an image, or maybe enhancing the image itself while we are taking a picture. So is that AI smartphone or is that a generative AI smartphone?

[Tarun] (3:50 - 5:00)

Good point. So one is like if you are, let's take a couple of years back, we were leveraging AI for camera enhancement and the basis on the profile. We were like using AI for the image post-processing, right?

But now we are reaching at a phase where we are using AI portrait, let's say. So for example, OPPO Find X7 Ultra, so what they have done is they have done a very impressive on-device AI feature called AIGC Eraser. So the point is what AI was doing earlier, GenAI is one-step ahead.

We are not only creating a nice image, but we are also able to edit in a way that would not have been possible, like what we are seeing in some of these devices, like S24 series, Xiaomi Ultra, we have seen 14 Ultra, we have seen in the OPPO Find X7. So a lot of these examples, but we believe this is just a starting point. These are very, very early use cases and this will evolve over a period of time like we have been discussing.

[Mohit] (5:00 - 5:19)

And one of the things that I have particularly noticed is that there's a lot of talk about LLMs out there. LLMs also come in various sizes, but more importantly, in the definition of Counterpoint, how do you treat LLM? Is it going to be running on the device or on the cloud? How does it all work?

[Tarun] (5:19 - 7:36)

So, for example, from the security perspective, definitely we'll treat the LLMs that are working on-device, that makes it a very nice fit into a generative AI discussion. But we believe that the future will be more hybrid. There will be certain models that will be running on the cloud, so it's more like a hybrid as of now.

But I think the discussion here is the capabilities and the performance of large AI models. We believe that those will be largely determined by the quality and the quantity of parameters that will be fit into it because what we believe that the smartphone hardware design is quite demanding already. You need to take into consideration constants such as heat dissipation, power consumption, battery life, and as well as the PCB footprint, right?

So, the parameter size of the LLMs running on the smartphone can't be too low. We believe right now in our definition, we are taking it, obviously it needs to support multimodal, multilingual features, and we are taking an inference speed of for a 7 billion parameter on-device LLM with a token rate of 20 tokens per second, which we believe will match the average reading speed of humans. So this is what we are taking.

Right now, we believe an LLM can run at even 8GB of RAM, DRAM, but our preference here is more on the 12 GBs to run these models even more effectively. But I think one very important point here to note is we are looking at a few GB models running in isolation as well. So practically within one year, we'll have four to five models running in the background, which will be a combination of SLMs and LLMs, and maybe a one or two app-based model as well.

We believe that it all depends on OEMs, how they are going to look into the entire use case that are going to come up. But at least certain hardware parameters like 7 billion, 12 GB of RAM, these we believe are foundational to make sure that experience of those models on the device is good enough.

[Mohit] (7:37 - 8:25)

Sure. And just for our listeners, I clarify that SLM is small language model and the LLMs are the large language models, but again, they are running into billions of parameters in any case. So thanks for elaborating on the hardware specs that you spoke about.

And again, you also feel that many of these will be running in the background, just the way apps do today. So mobile apps, multiple mobile apps are running in the background, and many times some of them are simultaneously working. But would that mean that the price of the smartphones will go up?

Because again, you're talking about hardware specs that would be required, which is 12 GB is what you're saying, and even more going forward. So many LLMs will be there. So again, what will happen to the pricing itself?

[Tarun] (8:26 - 11:01)

Yes. That's a billion-dollar question, right? So that's everyone right now on debating. And then I guess me and you already had multiple discussions on that. I think there's no fixed answer to this right now. Obviously, there is no brainer if the component that goes into the smartphone becomes more powerful.

Obviously, there's some cost that is attached to those and your bills of material, bomb cost will go up. But the point is, how industry is going to basically monetize this entire opportunity? Are you going to realize that value upfront on the device that says, hey, this is your generative AI device.

This is, let's say, $50 costlier than your premium phone, because as of now, the entire trend is on the premium side of things. We believe that GenAI smartphones will first diffuse into the premium and then further diffuse into the lower price tier starting 2025 onwards. But again, the point is whether the OEMs will realize that value upfront from the users or there will be like more models evolving, like, let's say, generative AI as a service.

Samsung has also hinted a bit like until 2025, these things will be free. And then who knows, you'll have an AI as a service running on your phone. If you need a better service, you might be paying, let's say, $5 a month or $10 a month.

And that's how there will be a revenue sharing between the OEMs or they might plan to realize on that. But we believe that looking into the launches right now, we have not seen OEMs breaking that cost into this as of now. There will be a lot of R&D that goes into this.

We have seen what Meta is doing with the foundational model, what Google is doing. There are almost like more than 100 plus LLMs right now in the market. A lot of these companies are doing big billion dollar investments in this area.

So, again, the question is how they are going to realize. We believe that hardware, there is a less probability that phones are going to be, obviously, there will be some delta, maybe, let's say, they'll decide, ok two or three percent will realize from the hardware cost and the rest of them will go ahead and make a subscription. If someone is going to have a basic AI service, someone is going to have an advanced AI service going forward.

So, there could be multiple models that can be evolved. Assuming you have a good use case is coming into picture one year down the line, two years down the line.

[Mohit] (11:02 - 12:05)

Certainly. And I think that's a very valid point. We don't know how the pricing is going to be in the future. But again, we can only think about various scenarios that could be there. But looking back, I've been following the phone industry since the time of the feature phones, a long time back. And there were these monochrome phones.

And so every time there was a new innovation that came to the market, there was an upgrade cycle that would kick in, for example, from monochrome to color. And I still remember when color came in at a certain price point and suddenly the market shot up like big day. Then there were these smartphones.

And then after the smartphones, the apps came in. So there was always something or the other that was driving the market. And in the last couple of years, we have seen that the market has been subdued for smartphones.

So, what do you think? Will this again drive the upgrade cycle? Will this lead to a shortening of the replacement cycle? And we'll see the growth coming back to smartphones.

[Tarun] (12:06 - 15:33)

Right. So we believe that this will have a positive impact on the overall smartphone ecosystem, whether it is components, whether it's value chain, because it's not just hardware. There's a lot of software that is going to be at play here.

And we are talking about experience and experience will be largely driven by your like how your software is driving that experience. So the point is, we believe plenty of it is tied up with the use cases. Let's assume there is a killer use case that is evolved and then users are talking about it and suddenly people started upgrading about it.

And this also depends on how OEMs are marketing those features. For example, Samsung has done a great job with the circle to search, right? We remember like even in our smartphone discussions and the retail chat, we have seen people who are coming to the store and say, hey, I want a smartphone that has that circle to search or how Motorola is leveraging Moto AI from the camera perspective.

So, it all depends on how OEMs are going to market these use cases. Right. And that is going to drive a big upgrade. And this upgrade will be very different from the technology upgrade, because you have been earlier over a period of time upgrading devices mostly on the feature. Right. So this is a feature or tech transition 4G to 5G. There is definitely an upgrade. This upgrade will be a bit different. This will be more like an experience-based upgrade.

So, for example, if what the larger hypothesis in the GenAI discussion is, all these times we were very much adapting to our smartphones. Now, in future, our smartphones will be adapted as per our choice, as per our personality. Now, the question is, if my smartphone, the same smartphone I'm using, a GenAI phone, is like adapting to my user needs and preferences, I have a greater incentive to upgrade to those phones rather than the ones where I am supposed to align to the smartphone.

So, yes, this opens up a window of opportunity, I think, for the smartphone market. But like we discussed that everyone depends, everything depends upon how the entire GenAI opportunity is being positioned to the user. As of now, we are also doing a consumer research on how consumers are perceiving the GenAI on phones and very interesting comments are coming, users saying, hey, this is something that was earlier as well, I think very much similar to what question you asked in the beginning.

People are confusing between AI and GenAI, so they think it's the same. But then use cases will differentiate, like, for example, even in the imaging, right? You are that blur, taking someone out of the picture, you are erasing it, making it more customized.

You could not have done in the past, maybe in a better way. So it all depends on how consumers perceive, as of now, they think there is no big deal. And once consumers realize the use cases are out there and then suddenly you see consumers saying, hey, we are going to upgrade because this phone makes more sense for my needs and preferences.

And then we have OEMs who need to position this as a value for users, because it's all experience based preference that is going to change in this area going forward.

[Mohit] (15:34 - 15:48)

One thing I can't help notice was you mentioned Samsung, you mentioned Motorola and so many other brands, but one key brand that was missing was Apple. So what's Apple doing in this space or are they lagging behind?

[Tarun] (15:48 - 17:50)

Yeah, so I think if you look at it, it's a very typical Apple, not many in the past two decades we have seen Apple coming out and then they're always late to a certain extent, but they make sure they do it right. So I believe for a closed ecosystem like Apple, they need to look into things in a different way than how an open ecosystem like Android thinks, right? So it's more like you go it out, you launch it, let consumers react to it and then take the feedback and then fine tune it and then it gets better and better.

Apple is like we are going out, but we are going ahead with the best in the market. Right. So that that makes them a little bit behind. Here in the GenAI space, I don't think they are far behind. It's something we just started right two quarters back. And then everyone is there a lot of speculations out that the upcoming WWDC, there will be a lot of AI focus.

In fact, during the last week, Apple earnings call, there was like more than 15 times mention of what analysts have been asking from Tim Cook and Tim Cook do not want to go ahead of the announcement. So what he said, we are incredibly optimistic about the application of AI in this space. So it seems like the way they are going to look at it is more use case based, because obviously, if you look at traditionally, Apple doesn't talk about hardware in a big way.

They do not have the sensor numbers and camera megapixels or the memory to their name. But it's all about experience. And this GenAI differentiation is also about experience. So I think Apple, if they get it right, that could actually market that GenAI features to the consumers on a broader scale. So we believe that there will be a lot of announcements during this WWDC on the new iOS upgrade that will talk about AI features on the new iPhone.

[Mohit] (17:51 - 18:13)

Thanks. And we did speak about now probably most of the OEMs. But again, if we look at the ecosystem that is much broader than the OEMs themselves, so you have the chipset players, you have different players who are and then on the on the LLM side, we do have a lot of players. So can you touch upon some of the strategies that probably some of these chipset players and other ecosystem players are focusing on?

[Tarun] (18:14 - 22:03)

Right. So if you look at Qualcomm and MediaTek, they have been doing a great job in terms of awareness, starting with Snapdragon 8 Gen 3, Dimensity 9300 from MediaTek, I talked about multimodal capabilities and they have taken a different approach in basically making more awareness about this GenAI. Qualcomm, on the other hand, they have already launched multiple as well, making sure the OEMs will launch a broader set of devices, not just focusing on the premium.

We have seen multiple SOCs from Qualcomm that are also targeting at $300 to $500 as well. It has 8s Gen 3, you have one in the 700 series. So the point is these actually the Qualcomm and the MediaTek, because we are talking about the experiences based on on-device, right?

So, there's a lot of onus on these companies to make sure that the experience live up to their expectations. And you need a very powerful device for that. These companies have done a great job.

Now, the second thing in the value chain is all about the OEM. So, for example, like we discussed, Samsung has been leading the AI trend, showcasing some consumer-friendly use cases like Live Translate, Circle to Search, automatic summaries of long text, auto-filling. And then they use multiple models like Gemini Pro and then they use Gemini Nano for that.

And then similarly, different OEMs have different strategies. So, for example, Xiaomi, they are also developing their in-house MILM with 13 billion parameters. OPPO has also showcased innovative use cases like I mentioned, AI Erasers, Phone Conversation Summary.

They have their own GPT with 180 billion parameters. They have established OPPO AI Center. I think they're doing a great job on that. Vivo, on the other hand, they have also introduced their own BlueLM with multiple billion parameters. They have introduced a blue kit for the developers. So in short, and not to talk about Honor, I think they were the best during the MWC.

I guess you were there and you saw the devices, Magic Series first hand and the way they are promising like how AI will cut the entire thing of one step at a time and multiple use cases. So I think the OEMs are there. They realize that it is going to be a big thing if they get it right.

But still early use cases, still early days. So it's more like a chipset. And not to forget the third part of it, the memory guys, right? We have SK Hynix and Microns of the world. They have introduced the DRAM capacity to support these AI smartphones. So everything is working in tandem right now.

There is a lot of developer support that is happening to create a use cases. Maybe in future, you do not need to download the app from the Play Store. Your assistance becomes so powerful that you do not need a single app on your phone.

All you need is a voice input and then output based on and not even input. It could be like a predetermined output for you based on your maps, indication, what you are currently doing. I'm sitting in the office. My office closes at 6.30. I've been booking Uber for the past six days. At the same time, maybe these guys just book the Uber without even me letting know that at 6.30 or something like this. So there can be plenty of use cases, but the point is, will those use cases be good enough for users to upgrade or meaningful going forward?

So, I guess it's an ecosystem play at here, right from the component players to the developers, to the OEMs, to everyone in picture right now. So it's an interesting segment after a long time in the smartphone area.

[Mohit] (22:04 - 22:34)

Yeah, surely. And the interest is not only in the smartphones, but in adjoining areas as well. So at CES, for example, I saw that Rabbit was another device that was launched.

And then we had Humane AI that was showcased at MWC as well. But again, if you look at, there are different kinds of LLMs that are also developing across different regions. For example, China is slightly different from the rest of the world. So do you think by region, we'll have differentiated experiences in terms of the user experience?

[Tarun] (22:35 - 24:15)

Yes, in fact, very good question, Mohit. So what we are saying, I was looking to the announcement, what Ola has done, although it's a different like Krutim AI, they have developed in-house India-owned AI. So what China is doing, why those are new models and there are multiple, like more than 10 models that are happening in China.

Yes, there will be regional experiences like for 5G, we have been talking about Korea. It was a different use case that emerged in 5G. China was different. India will be like different, like fixed wireless. So in the same way, I think the regional use cases and approach will, like for example, based out of India, I can see a big use case in the vernacular side of things, right? Going forward, because here the dialect changes every 100 miles, right, in India.

So going forward, maybe you do not need to know. So your smartphone is good enough to basically help you navigate these regional changes and variation. And you have a vernacular stack, you have an India-based stack that goes into these phones, OEMs are adopting it, and then you are getting those.

So, I believe regional approach will be there, but like any other smartphone development, it will be at the hands of a few companies. I can see Meta, Google, obviously Microsoft, and then these companies taking a lead to start with. And then obviously, we'll see how the regional models evolve and coexist with the larger global players.

So, it's not like one regional LLM will be deployed on the device. It will be a coexisting of multiple models like we see on a device and then see how the use cases are emerging from those models.

[Mohit] (24:16 - 24:29)

Yeah, thanks, Tarun, pretty interesting. So before we wrap up, I would really like toget into our forecast for generative AI smartphones. So what are our forecasts? So what is it that you're expecting in the future?

[Tarun] (24:30 - 26:15)

Yes, so like we did a press release last month, and I think we'll have an update coming out in the next one month as well. What we believe is the smartphone share of the overall smartphone will reach 11% by 2024 and 43% by 2027. And by 2027, this translates into roughly 550 million units.

But the installed base, we believe, should surpass one billion by 2027. So there are four major trends that we are expecting. 2024, we expect this GenAI trend to start, obviously, with the Samsung and Apple entering into, we expected Apple entry into this space.

2025, we think the broader use cases will emerge. 2026, we think this GenAI will diffuse towards the lower priced tier, especially $300 to $50 or so. And then by 2027, we believe this, the entire shift of the use cases, everything comes together and the entire installed base will cross one billion.

So that's the roadmap we are seeing. But again, it all depends. It will change dramatically if some killer use case emerges, awareness increases and how OEMs will basically market this feature going forward, especially differentiation between AI and a GenAI phone is very, very important to realize, because like we say, 11% in 2024, AI phones are already 80%.

So almost four in every five smartphones is already an AI phone based on the SoC capabilities. But yeah, GenAI is something we are discussing and talking as a great differentiator going forward.

[Mohit] (26:15 - 27:10)

Thanks, Tarun. Thanks for the insightful discussion. Good to know about the entire forecast and how this entire exciting field is going to evolve over the next few years. And I do feel that we are just scratching the surface right now. And there is a lot that we do not know. We can only think about what can happen in the future.

And I mean, we'll stay tuned on this particularly interesting topic. And for all our listeners, this another area of interest is generative AI PCs. So stay tuned. We will be coming up with another episode for generative AI PCs in the near future. And you can also listen to all our previous podcasts on all major podcasting platforms such as Apple, Spotify or even on Counterpointresearch.com. This is Mohit signing off for now.

Look forward to seeing you in the next one. Thank you.

Thanks.