Podcast Transcript - ChatGPT and Generative AI: Differences, Ecosystem, Challenges, Opportunities

00:00:06 Maurice

Hello everyone and welcome back to The Counterpoint Podcast. I'm your host, Maurice, and today we're going to be talking about AI and specifically Generative AI. Joining me today on the podcast, I have Associate Director Mohit Agrawal based in Amsterdam and Akshara Bassi based out of India. Hey Mohit, how's it going?

00:00:27 Mohit

Good. How are you?

00:00:29 Maurice

I'm doing good as well and hello Akshara. I hope you're doing well.

00:00:31 Akshara

Yeah, I'm doing awesome. Thanks for asking.

00:00:36 Maurice

Perfect. So, let's get right into it. Generative AI has been a very hot topic over the last few months, and at MWC or Mobile World Congress, Qualcomm, for example, showcased an on-device generative AI solution called stable diffusion that can generate an image offline based on text input.

Another example of this has been Adobe's Generative Fill in Photoshop, which can essentially generate or replace a background based on the image that you're working on. And another example is Google, which has its own generative AI model, and there are many other companies as well that are focusing on this really hot trend.

But, you know, why is this such a hot topic? Since last year basically, everybody was talking about the Metaverse, and now seemingly the focus has really shifted to this generative AI and specifically since the launch of

ChatGPT. So, my first question actually goes to you Akshara. What can you tell us about Generative AI?

00:01:37 Akshara

So Generative AI is a type of Artificial Intelligence under the umbrella of Artificial Intelligence that can create new content, including text, images, audio, videos. So, the models work by training on existing data and then they use that new data that the model has learned to generate new examples which are in similar to the original data that they have trained upon. So, generative AI is focusing on creating new content while traditional AI which we have seen over the past few years have been more on analyzing data and making predictions, but not generating new content per say.

So how do they work as the models are trained on big data sets and they require large compute power. Usually, they are trained on in data centers and then those models can be used for a variety of applications from a smaller smartphone level applications to large enterprise level applications.

So, as you just mentioned, ChatGPT is one of the most famous examples right now that we had of Generative AI, which can hold a contextual conversation and is basically it's trained on a lot of parameters about the new one is approximately trained on 180 billion parameters and it can generate text, code, audio, video and images and which is almost undistinguishable right now from human written text.

00:03:08 Maurice

I see, that’s all very fascinating and a lot to comprehend, but I want to bring in Mohit into this conversation as well. You know, Akshara mentioned a little bit about ChatGPT already. Can you tell us a little bit more about any differences between ChatGPT and generative AI?

00:03:26 Mohit

Thanks, Maurice. For most people, Generative AI and GPT are the same thing, and that's probably because the term generative AI became popular when ChatGPT was launched. In reality, Generative AI, though, and GPT

are both types of AI, but and then they can be used to create new content. However, there are some key differences between the two.

So generative AI is a broader category of AI techniques that involve generating new content based on patterns learned during existing data. The content could either be text, image, video, music, or code. On the other hand, GPT or generative pre-trained transformer is a specific type of generative AI model based on transformer architecture developed by open AI. And this is used only for generating content around text and code and I must also add it is one of the most powerful and versatile generative AI available today.

00:04:25 Maurice

You know there has been a lot of interest now in, you know, ChatGPT and generative AI and all these big language models that are being used. But there's also an issue about bias in there in the data sets, right? Is there anything you'd like to add about that that has been going around.

00:04:45 Mohit

Yeah, you brought a very important challenge around the bias and the bias definitely exists in the data sets itself. Whatever data sets that we are using to train on the output is likely to be modelled on it. To give you an example. If you look at, there are a lot of roles that are more male-oriented and if we are using the same data to train our models, the output is likely to be more driven towards men and instead of being gender neutral.

00:05:14 Maurice

And I want to get a little bit more into which companies are actually working or active right now in that generative AI space. Can you elaborate on that a little bit more too?

00:05:25 Mohit

So, we know all the usual suspects, whether it is Open AI, Google, Facebook, Microsoft, they are all very, very active in this space.

But if you look at companies like Adobe, they are doing a significant amount of time and are a pioneer in the field of generative AI. Their

sensei technology is used in a variety of products like Photoshop, Illustrator and so on. Then we have Hugging Faces. It provides open-source tools and resources for generative AI. Their transformer library is one of the most popular out there for generative AI. We have Autodesk that is using generative AI to automate tasks in the design and engineering process. For example, Autodesk generative design technology uses generative AI to automatically generate a variety of options on a given set of criteria.

There are many other companies like Uber, Salesforce that are using AI or generate AI to improve their services and operations. Companies like Expedia, Kayak, OpenTable, all these companies have trained large language models to improve upon the customer experience and remove friction in the customer journey. Having said that, it's not that generative AI is limited to big companies alone.

Gates Foundation is funding 50 projects in middle- and lower-income countries, and some of the projects are very fascinating. One of the examples and that really touched me was a group of scientists in Uganda they planned to build a ChatGPT-based application to provide information to farmers on crop diseases and that too in their own native language. Imagine how powerful that would be that the problem so far has been that that a lot of knowledge has been there. but not available in the native language and hence the lower strata or the people who do not know the other languages have not been able to benefit from it.

Another project in Brazil, there an NGO is planning to use these large language models to develop a support bot for psychologists and lawyers helping women who have faced gender-based violence. Akshara has been doing a lot of work on researching the generative AI ecosystem, and she can talk more about it.

00:07:43 Akshara

Thanks Mohit. So, I think. You very beautifully explained how the companies using the models. But if I just take a step back. So, if I go from the stack, who are actually creating those models, so we have the obvious companies, the big techs, which is Google, Meta, Microsoft with acquisition of stake and Open AI.

But then we have also other companies which are working on that whether they be the Chinese companies or other companies which are specifically in Europe, there are a lot of companies that are working on, also creating their own models, for example in China it's about Beijing Academy of Artificial Intelligence.

Then we have Alibaba, we have Baidu opening up, in Europe we have actually Luther.AI, Laion and Bloom, which is a parent company is Bigfin so it's a nonprofit sort of company which is working upon AI-creating models.

So, these are the companies which are actually creating that infrastructure, basically, those AI models and the other companies either are licensing from them or are using it through the open API sources.

And then if I go back down the stack again, where are all these models trained? As we mentioned earlier in our podcast that they require a large amount of computing resources. So where is that compute sourcing coming from? Who's providing the infrastructure? So the companies have two obvious choices either they buy their own infrastructure, the IT data center infrastructure or they go to cloud companies, so Meta, AWS, Oracle, Alibaba, Microsoft, Google, Cerebras, Baidu are kind of the top cloud service providers which are giving AI as a service or supercomputing as a service depending upon the companies requirements.

And the most kind of right now the most hot topic in the specifically on the hardware side of things, if I come who's providing that infrastructure itself. So, NVIDIA is the current darling considering, they actually have almost monopoly on the AI chips. So, AI chips require specific type of chips because there is a lot of computes happening parallelly.

So, the models can train themselves, so NVIDIA is the highest, but the traditional chip companies like AMD and Intel are also catching up with their AI offerings or AI-specific or focused offerings. And then we have companies like Broadcom and Marvell, which are making the AI chips on behalf of certain vendors, whether it be Google, Amazon, or Microsoft, which means they do a custom silicon chip.

And then so pretty much all of this is coming from the hardware angle and if I just again step a bit more back and see who's manufacturing that

hardware, then it is primarily TSMC where most of these chips are right now being fabbed.

Intel is of course using its in-house foundry, but rest everything is actually being fabbed on TSMC right now, so this is kind of the whole ecosystem that we have talked about from you know software providers or the application providers to hardware and manufacturing.

00:11:12 Maurice

Awesome. That's a lot that you just unpacked there. But I wanted to more specifically focus on one thing because you mentioned a lot of US companies actually that are in that space and only a few Chinese companies. But I wanted to, you know, for you to just dial in a little bit more on, what specifically the Chinese companies are doing in this space?

00:11:36 Akshara

Sure, so Chinese companies due to the ecosystem or the unique ecosystem that China has, what they are trying to do is they are trying each companies trying to build their own AI model instead of a hybrid mix that we have seen in US and in Europe where the companies are, you know, either licensing the AI models. For example, GPT 3.5 or 4.0 for their purposes and integrating it out. So, for example, like we mentioned Adobe Salesforce, most of them have integrated GPT but Chinese companies, each of those major Chinese either e-commerce player or a cloud player, have built their own proprietary AI models.

That's like a kind of a different take to what Chinese companies have taken. And secondly, there is another what was the perspective coming from a macro point of view where the companies due to the bans of the chip companies specifically by U.S. government, so they are facing a lot of obstacles, actually sourcing those chips and consequently the hardware that can train their AI system.

So, I think that will be a bit of challenging kind of headwinds for the Chinese companies, but it's definitely the ecosystem is definitely kind of building up in China and US both.

00:13:02 Maurice

Thanks, Akshara for giving us a little bit more of a deep dive into the Chinese market, but I wanted to actually now move back to Mohit and ask you a little bit more of, you know generative AI is extremely powerful and you know if I'd like to quote one of my favorite Marvel superhero Spiderman, there's "with great power comes great responsibility."

But what are the challenges associated with this? You know, with this accelerated progress that we're seeing with AI like what's happening there?

00:13:41 Mohit

Thanks, Maurice. It's a great question. Generative AI is still a new field and there are many challenges in it. The top two challenges are high cost and potential for bias. Bias is something that I alluded to in your previous question as well. And then there are many other smaller challenges that are there, which we will need to address over a period of time.

So, generative AI models, they require a large amount of data to train, and this can be a challenge for businesses that do not have access to large data sets. So again, there could be a divide digital divide that can come in between the large companies and the smaller companies just because the smaller companies may not have enough data to train on.

Then the other challenge is the training complexity is really high. So, for training the generative AI models, we really need a very high compute power, and these are time-consuming and also depends on the number of parameters that we require to train. So, as the number of parameters go up, there is an exponential increase in the computational resources and the training times that we need.

We already spoke about the challenge of bias. So, if you look at training data sets and if the training data sets itself has a bias, then the output is likely to be biased to give you another example, if the model has been trained on lot of images that have white people in them, then most likely the generated images will also have white people. There could be cultural and gender bias as well. The other challenge that is there is the intellectual property right itself.

So, these generative AI models, they are creating new content such as images, videos and music. We need to have clear laws and need to develop clear intellectual property rights for generative AI models, because the creators of the images or the content that was there, they need to be compensated because the base for the generative AI output is the content from the creators themselves. So how do we compensate them is something that we really need to find a way for powered on?

Then there are issues around deep fakes and inaccurate responses with confidence. There are issues around cultural sense also, so I use chart GPT multiple times, and I'm surprised the kind of answers that it gives me with confidence and sometimes it's giving me the wrong answers but with such confidence that I tend to believe that what I know is wrong and what it is telling me is probably right. So, we need to have a way of finding out how find the responses that are accurate and some of them which are not accurate.

And then there are other challenges around AI-specific infrastructure that needs to be developed for training these AI models and within the companies itself, we will need to have the culture and the organization to support it.

00:16:46 Maurice

And you know, just to your point, you're getting confused using some of this generative AI feature is, you know, we're not going to lose our jobs. I don't think anytime soon you know someone needs to definitely still take a look at what's being produced and a human needs to go through that data to make sure everything is correct.

Because a lot of times there is still something going on there that we need to definitely look at. But again, you mentioned infrastructure and I think AI infrastructure is actually a real key point in this. Akshara, can you tell us more a little bit about the current hardware that is being used for AI advancements?

00:17:32 Akshara

Yeah, sure. So, like in my earlier description, I did talk about the ecosystem of AI. What are the companies doing? Who are the major players? But all at the end of the day, it comes to is, no matter what the

companies' ambitions are about their AI, whether they want to integrate it across the whole suite of, you know, services and offerings. Or maybe, you know, just a company wants to bring out a specific cool AI Kind of you know, an add-on feature or maybe a new feature which the consumers can try, or enterprises can try, but this all can happen if an AI model exists in the first place.

So that's where this AI infrastructure kind of is very crucial and what we have seen this year is and probably it will kind of spill to the next year also, is most of the companies who are actually invest in IT infrastructure have transitioned their investments in AI, Visavi the journal purpose compute that we have seen in past years. So, whether I talk about Amazon, whether I talk about Azure or Microsoft Google Meta, we talked about AliBaba or and we talked about Oracle, so any company you take it.

So, what we have seen like we just ended up Q2 earnings season as it is like only called. So, what common thread across every company was the investments in their AI infrastructure and so where it stems is right now the hardware infrastructures. It all comes to compute and the compute is in scarcity right now because as I said, NVIDIA is the top. It's pretty much the company which is making off shelf chips.

AMD, Intel, and Marvel are catching up. But still, they don't have a feature which can compete with the NVIDIA's top-of-the-line or the latest product portfolio. And in fact, we just like I think we are hearing that you know, NVIDIA is sold out till the end of next year also that's the scarcity of AI chips that is going on in the market right now. So, I think these companies who are building out AI infrastructure would be super crucial and the second thing we would see also is the new business models that are coming up.

So, we are seeing a lot of companies as I said, they are either offering AI as a service or supercomputing as a service or even GPU as a service. For example, we have seen there is a new not new, sorry, but a very specific niche cloud player which is only a company called Core V which is just giving GPU as a service. That's it. Nothing more, nothing less.

So, I think the companies are also having a bit of trouble finding right AI infrastructure at right cost. That they can, you know, they can train or

make their own proprietary AI models and then use them to either create a consumer-facing app or an enterprise-facing app, so I think this will kind of continue and the 2nd constraint is coming from the production also.

So why NVIDIA can't you know produce more and more chips because it's coming from at the foundry level, which is TSMC because AI chips need a specific level of packaging which is CoWoS. So, right now, capacity is not as at scale as NVIDIA would like to or TSMC would like to. And then the other component, specifically memory they use HBM memory which is High Bandwidth Memory and also right now there are very limited suppliers which are providing HBM. And the AI chips. HBM is a necessary component for these chips. So right now, that is also in scarcity.

So, I think these all kinds of ecosystem challenges and constraints are causing not the AI infrastructure to scale as per the demand is there in the market, right now.

00:21:48 Maurice

That's an excellent recap. I think of what's happening right now in that ecosystem. And Mohit if you could add on to this a little bit, Akshara mentioned a little, a little bit on this already, but can you expand on what chipset players are actively doing right now in terms of this?

00:22:07 Mohit

So as Akshara already mentioned that NVIDIA has a head start as most of the focus right now has been on the compute layer of generative AI. However, there are many other chipset players that have announced their intention to integrate AI within their chipsets that are intended for different applications.

So, there are companies like Cerebras, Esperanto, Tenstorrent and Graphcore sambanova, AMD and Intel that are focusing on AI hardware. Each of these companies are trying to solve the challenge of ever-growing AI models in their own unique way.

So, Cerebras for example. It's a company that is known for helping with COVID-19. Research is on a commercial expansion that would fit its

hardware against the compute infrastructure built on NVIDIA's GPU and the other thing that we have to keep in mind is that not every chip requires 180 billion parameter model to run. Scaled-down AI models work very well. And this has been demonstrated by Road maps of chipset companies like Qualcomm and MediaTek.

The other thing that we have to do is make a distinction between AI training and AI inference. AI training is where a set of data is fed into a model, and it learns from the data set to become more capable of making predictions. While AI inferencing is where the application uses what AI has learned during training, the compute requirement for AI inferencing is much lower than that of AI training.

Another big opportunity is for AI at the edge, where computation is done on the devices, Where the data is generated instead of taking the data to the data centres to run AI models. An example is autonomous cars where latency is critical. Many chipset players are addressing or working on addressing the needs of AI at the edge. So, AI at the Edge is a huge opportunity in itself, and as AI becomes more prevalent, real-time AI insights will happen at the edge for many use cases.

00:24:11 Maurice

Really fascinating! Thank you, Mohit. Now, before we wrap up the podcast, I had one last question I wanted to ask Akshara now I wanted to ask you a little bit more about what's happening in the future. So, can you tell us a little bit more about how the ecosystem will continue to evolve?

00:24:31 Akshara

I think in future, firstly the companies would start at least would kind of mature into the AI infrastructure or they would have you know trading models which are either on an iterative basis they are like training constantly and you know are kind of updating themselves. That's where I see like the future of AI model going on and but as Mohit mentioned, inferencing will also become lot crucial because once the models are trained, they need to bring out the outputs which are very contextual for the user and are super smart in terms of either it's whether it's new content creation or no, even giving out answers which are contextual in nature, they understand my context

And that's where I see like the companies who specifically focus on, you know, user consumer health devices will kind of ramp up because inferencing will happen at where the consumer is sitting or where the end user is sitting. So, that's where I see a big play you know the companies who make use on devices and of course, the infrastructure for the user devices, whether it be Qualcomm, MediaTek and the devices like you know those vision glasses or the AR glasses that we see, or whether it's my smartphone, it will truly become you know from a just like a predictive maintenance sort of stuff or a predictive answering to an actually a smart assistant for me.

It will truly become a very customized thing and I think most of these companies who are at the forefront, where they make the consumer infrastructure have already demonstrated in their either the roadmaps or with the POCs for their products and chipsets that how they are going to integrate AI and hence make AI actually prevalent across the consumer because they can't always go back to cloud and you know wait for Internet connectivity or any kind of connectivity for that matter to give an answer. So, the future is you know maybe you're not connected to the Internet but still, your smartphone is working smartly. I mean, that's where I actually envision the future. Or as we call it in our tech terms, it is edge.

It's everything lies at the edge, so I think edge computing and specifically edge AI compute will become more important. Specifically, once the companies have built out their back-end AI infrastructure, so I think more of this probably later next year, 2024 second half, we would see lot of you know announcements and devices coming on which cater to specifics of these devices and use cases.

00:27:22 Maurice

Yeah, that's actually very interesting because you know you and Mohit both talked about this topic, AI at the Edge, and Edge Computing and how that's all really evolving. But however, for the sake of time, I think we need to save that for another podcast, so I think we'll end it here. Thank you, Mohit and Akshara for joining the podcast and sharing these really great insights with me.

00:27:50 Mohit

Thanks, Maurice, for having me. It was a pleasure talking to you and Akshara.

00:27:55 Akshara

Thanks, Maurice. For having me bye.

00:28:00 Maurice

And for our listeners, thank you for tuning in. If you have any questions or wish to chat with our analyst on generative AI or any other topic, please do reach out at press@counterpointresearch.com. You can also listen to our previous podcast on your favorite podcasting platforms such as Spotify Apple Podcast, Google Podcasts, and others.

You can also follow us and subscribe on our YouTube channel, Counterpoint Research for analyst takes and other coverage, and that's it. For now, I'm signing off and see you on the next one.

Take care.

