Rashmi Mohan: This is ACM ByteCast, a podcast series from the Association for Computing Machinery, the world's largest educational and scientific computing society. We talk to researchers, practitioners, and innovators who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned, and their own visions for the future of computing. I'm your host, Rashmi Mohan.

If you've been using the popular generative AI tools in the market only to craft perfect emails or grammar check your documents, you're selling yourself short. Having been in this field for much longer than the recent popular wave of interest, our next guest can tell us a thing or two about large long language models and their applications. Professor Edward Y. Chang is an adjunct professor in the Department of Computer Science at Stanford University since 2019. Edward has been a director at Google Research in the past and president of HTC Healthcare, amongst many other roles. He's the founder and CTO of Ally.AI, an organization that is making groundbreaking moves in the field using generative AI technologies, technologies in various applications, most notably healthcare, sales planning, and corporate finance.

An accomplished author of multiple books and many highly cited papers, he has also won numerous awards, including the Google Innovation Award, the coveted XPRIZE, Tricorder, and the Presidential Award of Taiwan for his work. He's a fellow of ACM and IEEE recognized for his work in contributions to scalable machine learning and healthcare. We are so lucky to have the opportunity to speak with him. Edward, welcome to ACM ByteCast.

Edward Y. Chang: Thank you, Rashmi, for your invitation. It's my great pleasure to join the podcast.

Rashmi Mohan: Likewise, we are really excited to speak with you. I'd love to lead with a simple question that I ask all my guests, Edward. If you could, please introduce yourself and talk about what you currently do, as well as give us some insight into what drew you into the field of computer science.

Edward Y. Chang: Okay. I was born in Taipei and came to the US to attend college and I first attend UC Berkeley major in Operation Research, which is kind of optimization techniques. Then I work in a software company intrigued by programming. Then I consider my background was insufficient, so I went back to school and joined Stanford to receive my MS and PhD and the timing was perfect. When I was in Stanford, I was classmates of Larry and Sergey, and they are founders of Google. After I spent about seven years in UC Santa Barbara, received my tenure, then I joined Google because they have so many machines can do parallel processing.

At the time I worked with Fei-Fei Li to annotate ImageNet. Once we have received so many images and my lab started to parallelizing some mission-critical machine learning algorithms, including Subperfetum machines and LDNs and so on and so forth. It has been a very exciting journey. The recently I started

to study consciousness because as Yusha Bengio mentioned about four or five years ago, and the current AI pretty much focuses on computation, which is modeling humans' unconsciousness, to be able to do reasoning and planning, we have to think about how to model human consciousness and GAI seems to attend getting into the realm of human consciousness. So this is really a very exciting era.

Rashmi Mohan: That's great. I mean, I think our audience would be super excited to hear about this topic. I don't think we've covered this in great detail and it's so relevant in today's day and age. But I want to go back a little bit, Edward, to what actually was the most driving force for you to pick computer science, even in your undergraduate education? Were you exposed to it when you were younger?

Edward Y. Chang: I think initially it was just to receive a job, so I needed to start coding. So that was pretty straightforward. I really get so interested in coding because computing is really the foundation of science. Many different sciences we need to collect a lot of data, analyzing the data, and to be able to get some insights. So information processing is a big application which drive me to dive even deeper into computer science methodologies.

Rashmi Mohan: Yeah, no, I mean I hear you. I think many of us at the time when we probably picked computer science, early exposure and definitely the excitement of an up-and-coming field that had a lot of jobs was the motivator to sort of get into it, but it sounds like you found a lot of very interesting areas to delve deeper. As I was reading about your previous work, one of your very, very early innovations I was very surprised to read was you're credited with inventing the DVR, the digital video recorder, which really transformed the way in which we created and saved content. I would love to hear more about that phase of your work.

Edward Y. Chang: Okay. When I was in the PhD program at Stanford, my advisor, Hector Garcia-Marina, asked me to work on infrastructure very similar to Netflix, pretty much just a streaming video. This was about 1995. The bottleneck is not really on the server side. The bottleneck is really the internet to the home, the last mile. At the time, 1995, we are kind of accessing the internet using telephone line. So even I have developed this kind of media server technologies, it was not practical at the time. So the second part of my thesis, and we say, "Well, if we really cannot do real time information streaming, can we buffer some data in local devices?" The best devices to do buffering will be the disk, hard disc. So if we buffer some information already on the disk, with some initial latency, we can do streaming at home. For real-time TVs, we can pause the TV and the TV program can save on local disk, and then we come back to resume and we can fast-forward.

So that was a very simple idea and the implementation itself was not extremely hard. Basically, then we were replacing this technology using disk instead of tape to revolutionize the VCR technologies. The professor Patrick Harahan was also a major pusher behind this situation because I took his course and he

encouraged us to come up with these kind of interesting devices to help the world. At the time there are multiple companies' participants come to Stanford to see the demo, and one of the visitors later started a company. I think people probably still remember TiVo was started after about two years after I published my paper.

Rashmi Mohan: Yeah, no, of course. I certainly remember TiVo and what a revolution that was. What is also interesting, Edward, is that as somebody who was in academia and working on your thesis and your PhD program, did they just take your paper and kind of run with it at that point, think about how they could productionize it and make it into a product? Were you involved in that process at all? Did you have to collaborate with companies outside that were trying to make a product out of your idea?

Edward Y. Chang: I think unfortunately at the time, because we don't have the sense of filing a patent, and I think those founders just took the idea and they started the company. So I really didn't get heavily involved in the development of the product, but subsequently, after I joined UC Santa Barbara as a faculty member, Sony was very interested in collaborating. They planned to enhance this VCR only supporting one TV to be able to support multiple TVs at home. We work on the prototype to be able to support up to 20 devices at home. But I consider because the device is very cheap and end up all the family today every TV they have one digital VCR, but technology-wise, actually one VCR can support 20 TVs. So that's a situation. Then after that, I think the internet bandwidth becomes really, really high, getting increased faster and faster, and this bandwidth problem gets resolved and the research issue is no longer challenging. So then I switch my focus to machine learning.

Rashmi Mohan: Got it, yeah. I was going to ask you how did that transition happen? So the next phase of your career was your work that you did at UC Santa Barbara and then at Google, is that right?

Edward Y. Chang: Yes. When I joined Santa Barbara, I said, "Well, to pursue my tenure, I need to have a very exciting research topic and digital VCR definitely is something in the past." I say, "Well, maybe using machine learning to identify photos or processing video will be interesting." After working on the application for some time, then I consider I really need to get into machine learning because that's the foundation of object detection and object recognition.

Then I started working on the topic for some time. As I mentioned, I knew Fei-Fei Li pretty well, and when I joined Google, I collaborated with Fei-Fei Li and sponsored Fei-Fei Li about 250,000 US dollars to sponsor the project. The reason I joined Google at the time was we had so much data. In university, you just have no machine, no devices, no resources to process those data. Google has so many machines, so many CPUs. At the time we haven't started using GPU, so with the MapReduce run on so many GPU at the same time, so I joined Google to start develop Paleo machine learning algorithms. That was a transition in

about 2'06. Between 2'06 and 2012, my major focus was making those mission-critical machine learning algorithms to be able to run MapReduce on Google infrastructure.

Rashmi Mohan: Got it. What better place. At the least at that time, the scale of data as well as the infrastructure that Google had could not be rivaled anywhere else. So was your primary focus at that point improved performance of the machine learning algorithms? Is that what you were sort of focused on, or was it accuracy? I'm trying to understand what are the key problems that you were trying to solve?

Edward Y. Chang: Yeah, my focus was to improve the machine learning infrastructure accuracy to try to power Google's different applications and myself focus working on the Q&A system. So we need to do semantic parsing, natural language processing, natural language understanding, so with the robust machine learning algorithm, it would be extremely helpful. But at the time, this is about 2'08, most of the algorithms Google employed, they use a linear algorithm or sub-linear algorithm, which means they only want to process the data once. My colleague told me, "You don't want to use a much more complex algorithm support by the machine. The computation complexity is N-squared, and N-squared means if you can process 1,000,000,020 instances in one second and N-squared means now you need to spend 1 billion seconds to process all the data, and you just cannot do that in Google. Google has a lot of data."

So when I started to do a parallel machine learning on this quadratic machine learning algorithm, my colleague actually advised me not to do it because they say, "Well, you cannot work on something which is very time-consuming." But really, machine learning with big data was the trend. So eventually, AliceNet was extremely successful. Then people saying, "Well, the accuracy is so drastically improved, so now we are willing to put into a lot of money to parallel our computation to improve accuracy." The GPU then was started to be utilized and the cost was not as high as using a lot of CPUs and the entire field of using a CPU to process big data just took off around year 2014. If you look at Nvidia stock price, and you could see after the image ImageNet was published, AliceNet was very successful. Then about two years after that, Nvidia stock was increased by tenfold because of this paradigm shift.

Rashmi Mohan: Yeah, no, absolutely. I've been on a rocket ship since. But it's amazing that you were literally at the inception of this whole transformation and the use of a GPU for processing. It's pretty fascinating that you were there. What was the next transition like for you, Edward? What took you to your next role after Google?

Edward Y. Chang: Yeah, the next role was I consider it was a time for me to maybe contribute to my birthplace. So I say, "Well, maybe I should go back to Taipei and try to educate or maybe mentor the local students, youngsters. So also at the same time, can you improve the infrastructures of Taiwan's computation?" So I went back to Taipei to join HTC. At the time, HTC was a very good cell phone manufacturing company. Recall HTC was a manufacturer of the Pixel, maybe

Pixel 2 or Pixel 3. Of course, later the competition becomes extremely tough. HTC was no longer competitive and we sold our entire cell phone division to Google.

So during the time when I was at HTC, initially I contributed to these mobile phone applications. One notable application, even today, iPhone still doesn't have the application, was the 360 degree panorama. If you just take 20 some photos, you capture this entire sphere of 3D. That innovation was we used some sensors on the camera to try to capture the movement of the cell phone. Then we instruct a user the direction that need to point to take picture. So this way when they take a spherical picture, there wouldn't be holes in the middle because we know exactly which direction the photo has been taken and we know exactly where the information need to be acquired. So with about 20-some shots directed by gyroscope and accelerometer, we can capture precisely the entire sphere. That was extremely well-done and the Pixel has that featured today, but because we filed a pattern, so iPhone, we haven't seen that.

Then I move on to work on healthcare. The motivation of healthcare was Taiwan has a very good healthcare system and in about 30 years they have collected so many medical records. So as we already learned at the time, with big data, you can really improve the accuracy of many things. Healthcare diagnosis is one application then later I focus working on, and that was my major focus on HTC during the second half of my tenure over there. Then we can probably discuss about the IoT devices I work on at the time, and then we enter a competition called Tricorder and won the second place in the world.

Rashmi Mohan: Wow, that's amazing. I mean, I like how your work around image processing and machine learning led you to HTC. I mean, and of course the motivation to do more in Taiwan, but really led you to improving the quality of the camera and pictures that an HTC phone could take. But also the transition to healthcare was a very interesting one, and driven mostly by what you're saying, the record keeping in the overall healthcare system in Taiwan. That sounds fascinating. So moving into healthcare, what was the work like when you were... I mean, did you do more healthcare-related work while at HTC, or by then had you had sort of started to think about doing things on your own or getting back into academia?

Edward Y. Chang: Yeah, I started the healthcare project by an kind of accident because at the time a professor at Harvard University, and he would like to join the competition hosted by XPRIZE. XPRIZE is a foundation, they encourage Blue Sky Innovations. So some well-known competition they hosted early days was self-driving vehicle, and the latest competition was sending human to Mars or sending robots to Mars. So in about 2010, they started a project saying, "Well, there are a lot of remote areas, they are lacking medical devices and the doctors cannot do precise diagnosis. So can you put together a device which is very, very light in weight?" So their constraint to us is like five pounds, and with five pounds you

need to be able to detect or diagnose about 15 diseases including HIV, liver problems, and diabetes and those kind of diseases.

The challenge at the time, of course, the weight is a big problem. The second challenge is, if you want to do those kind of diagnosis at home, definitely you need to have some machine learning algorithms and to collect data to do supervised learning and you can do classification in the remote area. Once you have done the classification, the data can be sent to the cloud. A doctor let's say do a remote diagnosis on a patient in Africa, once the diagnosis has completed, the data can be sent to the cloud. Any doctor in the world can review the diagnosis and assess the quality. So we consider that a very kind of humanity kind of project, so we started working on that. Since at HTC we are really good in making devices light, like you'd make a cell phone very light, so we have the edge in putting devices together, and also because of my machine learning background.

This Harvard professor, Professor Pan, was delighted to work with us, so he kind of built a consortium with all the hospitals in Taiwan so we get a lot of data. On my side, I will focus on machine learning and also device manufacturing. At the end, although we won the second prize, and we consider actually we should have won the first price, and the reason we won the second prize was the first prize winner, they came from a family of five and they used very rudimental kind of devices and the methods, they came out with a Tricorder at home on their dinner table using 3D printers. We spent so much money and effort. So I actually raised a lot of money from Taiwan government. So at the end, the foundation may be seeing, well, this is Goliath fighting with David, so that doesn't make any sense, so they couldn't give us the first prize.

But anyway, the whole process, we learned a lot of good experience and that paved my way back to academia. We had a good collaboration with UC Berkeley at the time. HTC in about 2020 started working on virtual reality. So using virtual reality, we can scan a patient's brain and a surgeon before the surgery can fly into the brain to see the detailed structures. The brain surgery, or any kind of surgery, a surgeon want to remove tumors at the same time they want to keep the benign tissues intact. In the past, without this virtual reality visualization, a surgeon need to kind of imagine what the 3D structure like by taking a look at these 2D MRI images, and oftentimes they could make some mistakes or suboptimal surgery planning. Suboptimal means you have a path you goes in that you have to destroy some benign neurons, so after surgery, the patient may be malfunctioning speech or other functions. We definitely want to avoid that. So that also kind of collaboration with Berkeley and with the virtual reality, and Stanford invited me to host a panel and eventually I moved to Stanford to start my adjunct professorship.

Rashmi Mohan: Amazing. I mean, I love the story that you talked about, David and Goliath. Congratulations on the Tricorder victory because that sounds like an amazing innovation. I completely hear you, right? I mean, when you talk about

somebody who's built a very homegrown solution to a problem in comparison to a corporation that does it, I can see how that vote may have swayed, but it sounds amazing. So going into a little bit more about the device that you built as a part of the Tricorder, I mean was that used for more commercial usage as well, or is it mostly just a POC?

Edward Y. Chang: I think at the time it was a POC, but the challenge to us to make it commercialized was two-folds. One was we have to go through FDA. Every single device we have to go for FDA to get approval. The second interesting thing is whenever we have a software update, like say on our cell phone, we can just upload a new version without any trouble. But in medicine, you cannot do that. For every new version, even a minor revision, again you have to go through FDA. So this entire process is really not scalable. So we work with FDA to try to have this kind of regulation changed, but it was very, very slow, so we couldn't commercialize in time.

But at the time, the foundation of Gates, Bill Gates Foundation, and also from India we have some colleagues, they say, "Well, India FDA was not that strict," and so therefore we actually transferred the technologies to other entities and they started to working on those devices. I know China has a company called iHealth, they also have this kind of healthcare IoTs and try to do diagnosis and try to do a kind of treatment in the rural areas.

Rashmi Mohan: Got it, yeah. No, and that's great. Certainly I know countries like India obviously need it as well, because there is a lot of rural population and access to healthcare is a challenge. So I can imagine that this would be widely used and appreciated. So yeah, it's great that you were able to find a home for it, even if the FDA process was cumbersome.

ACM ByteCast is available on Apple Podcasts, Google Podcast, Podbean, Spotify, Stitcher, and TuneIn. If you're enjoying this episode, please subscribe and leave us a review on your favorite platform.

But your part, which is almost like the next phase of your career, when you were back at Stanford and back in academia, is that when you started to sort of get back or into looking at artificial intelligence and LLMs?

Edward Y. Chang: Yes. In the beginning, we didn't really pay attention about LLM, right? We work on natural language processing kind of technologies, but always we run into a lot of corner cases because during semantic parsing or language understanding, even we have a perfect model, often we encountered exceptions. Say we have an airline reservation system, but the customer may say something we didn't expected, then the robot just cannot continue. So we always run into limitation until GPT-3 launched last year and we say, "Well, GPT-3 functions much better." Since at the same time I was working on this conscious modeling and GPT-3 was able to do reasoning, to me it was a really remarkable situation.

Then I looked at GPT-3, and now GPT-4, and we know a lot of people saying, "Well, they still have some limitations, especially the model have some biases because of training data." Suppose all the training data we input is from CNN instead of Fox News, of course the answer be probably tilt to the left. So the model has inherited biases because of training data that has to be addressed. The second issue people understand is this hallucination. Hallucination can be randomly or could be kind of non-logical expression, so how to mitigate those. I think a lot of method has been devised like chain of thought, tree of thought, but I think we came up with an interesting breakthrough. We named the platform SocraSynth.

The idea was initially very simple. We say, "Well, LLM is so powerful, so knowledgeable, and also its knowledge representation is polydisciplinary." So when we import data into LLM training, we don't say this is a physics book and that's biology. We don't. When LLM was trained, they have no boundary of knowledge, they just put everything together. So even with today, we ask a question, it's about computer science, LLM doesn't know this is a computer science problem, it just simplifies the answers to answer our question.

So in this kind of representation without disciplinary boundary, or we call polydisciplinary, it actually can synthesize new knowledge. It can synthesize something we call unknown unknowns. Suppose we agree LLM can synthesize unknown unknowns, that is a big problem because human's knowledge is limited. If we don't know, we don't know, how can we even ask questions? So I use analogy to describe a situation. Suppose I'm a 10 years old kid, I go to this Nobel laureate award ceremony and there are 1,000 Nobel laureates sitting there. If I'm a person asking questions in front of this panel, there's no way I can ask any interesting questions and get insightful answers.

So the solution is [inaudible 00:25:40] human is not qualified to ask questions. If we want to get insightful information, we can only be a moderator. We put such a matter on the table, then ask Nobel laureates. They can do a debate, they can do discussion. We are just sitting there to listen. So that's a key insight we obtain. We say, "Well, you get hallucination. Yeah, maybe the algorithm may not be robust, but most of the time we asked the wrong questions or our question, the context we provide to LLM, was not precise enough, so therefore we didn't get good answers." So that was the motivation, and end up we have four algorithms together with this kind of debating setting. I think we make really interesting progress.

Rashmi Mohan: My gosh, that sounds absolutely fascinating. I think the part that was particularly interesting to me is when you're talking about the boundaries, the interdisciplinary work, that because of the fact that LLMs are not classifying the content into specific human defined subjects, if you'll, there is no boundary in terms of how you can bring two concepts together. Versus us humans, we may think of something, like you said, classified as biology or physics or psychology. But really, the areas of intersection that we don't expect at all is something that

the system that you're building can uncover for us. Kind of blows my mind. So in terms of SocraSynth itself, Edward, what stage of the product or the idea are you at now and what is the role of human beings in this process?

Edward Y. Chang: Yeah, it's already kind of ready to be utilized by various applications. In fact, I'm consulting multiple companies. It has been used in healthcare, used in sales planning, and also investment banking. So give a very quick example, suppose we want to diagnose disease, and we can actually get a lot of ground truth from US CDC. So they have this mapping of symptoms and then diseases, right? But it's interesting when we input a set of symptoms into Bard and also GPT at the same time, we say, "Okay, can you diagnose this patient with the following symptoms?" They actually came up with different answers. I say, "Well, you can have a debate why you come up with these answers and why your prediction are different from each other." So they actually provided their justifications. They also say, "Well, because the information you provided, the symptoms, they are not sufficient, so you have to ask more questions." Finally, they say you have to conduct certain lab tests, like blood tests and so on and so forth.

That was interesting results I obtained because whatever we obtained from CDC was called ground truth, but the ground truth, actually they have mistakes. This is really tying to this very good paper published by Johns Hopkins physician last year. Analysis was saying in the US about 5% of diagnosis is erroneous or misdiagnosis. This is a huge problem, not only liability, but human's health. So you cannot take CDC's data as ground truth, there are some errors in there. I have done this AI research in healthcare for 10 years, I always treated those data as ground truth. Now I open my eyes and I say, "Oh, if I input to SocraSynth, allow agents to debate, they will have nuances and tell me the previous diagnosis may be wrong." The human role in this situation is zero, just a moderator, because we are so limited. If people consider we are smart, that will be counterproductive.

Demis is the CEO of DeepMind and gave us three examples saying humans should get out of the way. The first example is AlphaGo. AlphaGo compared to AlphaGo Zero, AlphaGo Zero is carried of all the human experiences, AlphaGo Zero wins over AlphaGo. Another example is AlphaFold, protein folding. AlphaFold 1 uses human heuristic to be a model in the middle. In AlphaFold 2, Demis is saying, "Oh, let's carry the human heuristic," and the score of AlphaFold is much higher than AlphaFold 1. It's like AlphaFold scored something like 50, AlphaFold 2 scored like 90-something. The last example is self-driving vehicle. If you have human knowledge, like put a map in the middle, try to instruct the driver how to drive, no, it's not going to be very effective because human sensors and human heuristics always encounter exceptions. So that Tesla will say, "No, forget about human in the loop. We just do end-to-end training." So they got a much better, much more effective self-driving algorithm.

So in short, human, because we are limited in knowledge, even when we have one or two PhDs, we cannot compete with LLM, which has multiple PhDs at the

same time synthesized into this kind of polydiscipline representation. Therefore, the treaty of SocraSynth is human can only be a moderator. You be kind of in a very passive role. We can evaluate their reasonableness, their logic, whatever, but we better don't contribute our ideas. That's kind of very sad, but that's what we have learned so far.

Rashmi Mohan: Wow. Okay, so many pieces in there that I want to dig into a little bit more, right? Let's start with the simplest, is have you found that, and I don't know how you would evaluate this, but the precision in terms of the diagnosis is better by using a system such as SocraSynth?

Edward Y. Chang: Yes, that's true. So I just give a quick example. So let's say 14 symptoms, one is headache and the other one is fever. Then GPT is saying, "Well, you should ask additional question, like do you have those two symptoms happening simultaneously?" That I have never thought about. Also, they say, "Why didn't you ask is your headache kind of periodic or only happens once a day or whatever, and is your headache getting better or getting worse?" So all those kind of refined questions, most of physicians didn't even think of. They just said, "Do you have headache? Do you have fever? Do you have runny nose?" But the machine even asks for correlation, timing, duration, severity, all those details.

So it's interesting, [inaudible 00:31:53] the process I was impressed. This is just one example. There are many other examples like sales planning and investment banking. Investment banking, they're supposed to say, "Oh, I want to invest in this company, I want to buy stock." It has this SEC, they filing the petition, you want to say whether their petition is accurate or not. Based on that, you want to make a judgment whether you invest or not to invest. We really get a lot of insights from this SocraSynth debate process.

In the monologue Q&A, you may not be able to get good answers because during the monologue discussion, the system like GPT will give you the default biases, the default biases of the model. But if you do a debate, you can be free from the model biases because you force one agent to take positive position, you force another agent to take negative position. You are forcing them to be biased according to your will, rather than taking the model's bias. We push into a positive and push into a negative, then we have very intelligent modulate with contentiousness. Eventually, after they debate for a while, they reduce their contentions, come up with some compromised conclusions, and that conclusion can provide much more insights than the default monologue Q&A session.

Rashmi Mohan: Yeah, no, that was going to be my exact next question, which is inherently it sounds like a system like this would remove bias or at least reduce bias in a significant way. That's kind of what you were just talking about.

Edward Y. Chang: Yes, it also reduce hallucination because you say, "In the debate process, the two agents will keep on arguing with each other." During the argument, basically, your Q&A is extremely focused. The Q&A pair will be formulated by

the agent, not by the human. When human formulate a question, it can be very fuzzy. The second thing is when you started to do this debate on and on, after rounds and rounds, the contextual information is improved. The context is getting better and better and better so the debater on the two sides can delve into deeper and deeper, deeper insights. So therefore, hallucination has no room to exist. Also, I have a treaty, I say, "Well, do you ever have the same nightmare twice, exact the same nightmare twice?" The question is probably not, right? So if you don't have the same hallucination twice, it means you will not have the same bad argument twice generated by LLM. Then after this refutation debate process, all the hallucination will be disappeared.

Rashmi Mohan: Got it. It sounds like a perfect system, Edward. I'm just curious, what are the risks? What are the downsides? What are the challenges that you're facing?

Edward Y. Chang: Right now, we are doing evaluation because every round of debate, we want to make sure the quality is very good. Luckily, we use the Socratic method. Socratic method has been there for 2,000 years, but strangely people don't use it. We use the Socratic method to evaluate every argument's reasonableness. Here we say reasonableness is basically evaluate its logic, and we face a very interesting problem. A lot of people saying, "No, I want to evaluate whether it's fact or it's truth." Unfortunately, after doing the research for some time, I actually consider there's no facts in the world. I mean, the same event happens in somewhere in the world, if you look at different newspapers, the stories, narratives can be different. So there's no way I can know the facts unless I go to the news location to eyewitness what happens. Even so, maybe I won't be able to see the whole thing. I couldn't understand the causality, for example, I cannot tell who is right and who is wrong.

So then I say, "The only thing we can do is to evaluate reasonableness, the argument's quality." I think it turned out to be reasonable. So we have GPT and Bard, for example, doing debate. Then we have inferior LLMs that do evaluation, because when you do evaluation, you don't need to have knowledge too much. You just need to make sure the logic is extremely tight. We published a paper, we showed the evaluation is quite consistent. So we are pretty happy about not only generating content, also we are able to evaluate the quality of the content. Of course, there are much work to be done in the future.

Rashmi Mohan: Yeah, no, that sounds great though. Is this work that you also push forward via Ally, the company that you are the CTO at?

Edward Y. Chang: Yes, a company came to look for me because they considered the idea to be intriguing. Actually took them sometime to really understand the powerfulness of the SocraSynth system. There's a great potential to apply to many areas.

Rashmi Mohan: Got it, I understand. One of the things that I was reading was that there's a program called Stanford OVAL, I think.

Edward Y. Chang: Yeah, the program was established by Professor Monica Lam, who was my partner when even I was back in HTC and she would visit our company for collaborations. When I joined Stanford at the beginning, because Monica was interested in healthcare, we see a synergy there, so I joined her lab. So yeah, I think right now we are taking different approaches to address the problem of semantic parsing. So we had the same mission and we try to address semantic parsing quality, but we are using different methods to address the challenges.

Rashmi Mohan: Got it. It's absolutely fascinating work. I'm just curious, what's next as you sort of make progress? What are the next set of goals that you have?

Edward Y. Chang: On the side, I know you asked me about my hobbies, I write a lot of poetry and I also generate a lot of art and taking a lot of photographs. So I'm actually using GPT currently with the D to create some interesting art pieces and I just published a poetry book. So this multi-agent scenario, we talk about SocraSynth that can debate, you can have three or four agents and each agent is a persona. You can ask them to have a dialogue and you can create a fiction, create a novel out of this SocraSynth platform. So I say, "Well, maybe after my scientific endeavor, when the system become more mature and I'm getting into writing a novel using the platform." At the same time when you start working on different applications and we will discover additional research challenges need to be resolved.

The final grand challenge is if we really consider polydisciplinary representation have new insights into the knowledge, if we know LLM may have some unknown unknowns human cannot tap into, then my final goal is to be able to, even I don't know how to ask questions or there'll be a trick which I can trick LLM to tell me something which is I don't know, maybe I don't even understand, and it can teach me to explore some unknown unknowns, and that would be really remarkable. This is maybe changing the world of research, that's my final dream.

Rashmi Mohan: It is. Yeah, it truly is. I do have to ask though, I mean, because often the most colloquial question is, oh, AI is going to take over the world and take over all our jobs. In a lot of the way you described the product, if the human is out of the loop, how do you see human involvement in products and systems like this in moving forward? What kind of role would we play or what kind of jobs would we have?

Edward Y. Chang: So human can probably be only the moderator. But the moderator, a more skillful moderator or a more knowledgeable moderator can get much better results and much better insights. So humans still need to improving our own knowledge and hopefully, because we can work with LLM, so our knowledge acquisition can be much more efficient. So hopefully in the future we will see some new careers or new pathways to be able to support our own livings. In the short term, for example, data science, computer science, AI really can do a lot of work to replace human beings. So in the short term can be pessimistic, but in

the long run, there should be some new applications, some usages, and the human can be employed.

Rashmi Mohan: Yeah, for sure. Like you said, I mean areas that we've not thought of combining or bringing together, maybe those will be uncovered through these systems and give opportunities for us to explore new jobs and new careers and passions. It's fascinating, can be unnerving at times, but definitely very exciting. It's amazing that the work that you're doing in this field is going to uncover some of that. But speaking about passions, I spent some time on your YouTube channel, Edward, and I was fascinated with your photography work. I know you were just talking about it. I saw the one with the bald eagle and the snake, it was really quite something. So I was just wondering, what are your other hobbies and passions? You spoke about poetry, which is quite unique. I've never heard of a computer science researcher also being interested in poetry, at least I haven't come across anybody. So what else do you do?

Edward Y. Chang: Yeah, poetry is really my biggest subject. Since I was a child, I was extremely interested in literature and philosophy, and poetry is a very good kind of platform for a busy person to write thoughts because a fiction would take a long time and poetry typically is much shorter than the fiction. So for me, I think this is a much better kind of media compared with other medias. Amazingly, with GPT, and GPT can help me to write even better poetry, for example, collecting historical facts, and maybe sometimes can help me to have better rhyming, better choices of words. So this is really kind of interesting situation, [inaudible 00:41:48].

Rashmi Mohan: Very cool. Did I also see, did you climb EBC? Were you at the Everest Base Camp?

Edward Y. Chang: Almost two years ago, yes. That was a very good experience. Because my oxygen levels deprived, that was when I was study consciousness very, very intensively, and when the oxygen level get deprived, you are pretty much in the unconscious situation. That experience was interesting. Schrodinger once said, well, a lot of situation, we are unconscious, even let's say our vision. We look at a person, we pay attention to some object we are focusing on, but our peripheral vision still is processing data, but it's under our unconsciousness. But if our peripheral vision suddenly sensing a car driving towards us, then there will be a quantum jump to elevate that event from our unconsciousness to consciousness.

So just like Einstein has always said, innovation doesn't come from consciousness. Innovation comes from your preparation in consciousness and push all the information into unconsciousness. One day, when you are whistling on the Everest, you're unconscious suddenly something pop up. Then you say, "Oh, I realize what you just told me from unconsciousness." A lot of people have that kind of experience. You cannot wheel into innovation, you can wheel into

preparation and when the time is right, unconsciousness will tell you innovation has arrived.

Rashmi Mohan: That's such a great way for us to bring this interview to an end. I mean, I love the analogy that you drew there. But for our final byte, Edward, what are you most excited about in this field of using generative AI and the work that you're doing with SocraSynth and healthcare over the next five years?

Edward Y. Chang: Yeah, so far, the very practically, LLM has improved my productivity by 10 times. That means the next five years will be equivalent to 50 years of my previous years, previous endeavor. So this means I cannot squander any minute of my day because it's become 10 times more precious. I'm going to move forward to continue polishing SocraSynth. Interesting enough, I was writing scriber to scribe information for my investment banking company, and of course, SocraSynth and also ChatGPT helped me to write code, be very effective. But then I say, "Okay, now you scribe my SocraSynth site, getting all my papers, and GPT tell me what are the major shortcomings of my current algorithm or systems." Remarkably, GPT gave me three new assignments, new insights, which I will be working on in the next few years.

Rashmi Mohan: Wonderful. Well, thank you so much for taking the time to speak with us. It's been a wonderful conversation, Edward. Thank you for speaking with us at ACM ByteCast.

Edward Y. Chang: Thank you so much, Rashmi. Thank you, bye-bye

Rashmi Mohan: Bye-bye. ACM ByteCast is a production of the association for Computing Practitioner's Board. To learn more about ACM and its activities, visit ACM.org. For more information about this and other episodes, please visit our website at learning.acm.org/bytecast learning.acm.org slash B-Y-T-E-C-A-S-T.