CRAIG: Hi, I'm Craig Smith and this is Eye on AI.

Currently the largest AI system in the world is China's WuDao 2.0, a sparse, multimodal, large language model with 1.75 trillion parameters, trained on 4.9 terabytes of images and texts. While not directly comparable to OpenAI's groundbreaking dense LLM, GPT-3,WuDao is a more massive model and, what's more, much of its code is open source.

This week, I talk to Tang Jie, a professor at China's Qinghua University, who leads the WuDao team, about how the model was built, why it is unique and what his team plans for the future.

Connor Leahy, who was on the most recent episode of the podcast talking about large language models, joined me to lend his expertise in talking to Professor Tang. The audio is not always clear and so I encourage anyone interested to read the transcript, which is available on our website, eye-on.ai.

I hope you find the conversation as mind-blowing as I did.

CRAIG: Could I have you first introduce yourself and then I'll have Connor introduce himself, and then we'll get to the questions.

Do I call you Jie or do I call you professor Tang or what should I call you?

JIE: Just call me Jie.

My name is Tang Jie. I'm a professor in Qinghua University, with the department of computer science. I'm also the associate chair of this department. And my research interests include social network, data mining, machine learning, and knowledge graphs and, of course, general artificial intelligence.

Now, I just focus on the WuDao project. The WuDao project is a pre-training language model project. We have the ambition to pre-train the world's data, I mean, all data. All kinds of data.

JIE: So, we get data from China, most of the web and, also Wikipedia page and the knowledge graph page.

JIE: And also, we get some data in English. So, after cleaning, we have more than 1.5 terabytes cleaned Chinese texts, not the web page, but clean text, and 1.4 terabytes English text and also two terabytes of images.

JIE: And recently, we'll probably double the data, and will pre-train another model.

JIE: We have a team with more than 100 people working on that. And we have some very exciting new algorithm and also the model itself and many applications.

CRAIG: I want to introduce you to Connor in Germany. He's with a group of unaffiliated researchers and since he is familiar with language models much more than me, I thought it would be helpful to have him ask questions.

Connor. Why don't you introduce yourself?

CONNOR: I have a much less illustrious history than professor Tang. I'm one of the founders of a bunch of researchers interested in open-source AI research and several other projects. We're most well-known for ongoing project to produce and a release large open-source language models, which is still ongoing, but we also do a bunch of other cool stuff.

CRAIG: So Jie, when did the Wudao project start and what was the motivation?

JIE: We started this project last year, last October, or maybe September. or We want to build the largest the pre-training model. I have the idea in Qinghua three years ago, which means 2018. I have the idea in your mind. slogan, I want to in All the data in the world. Last year, the end of September, or maybe the beginning of October, we officially started this project.

CRAIG: Okay. Connor, why don't you go ahead with some of your questions?

CONNOR: Sure. Over here in the west, it's been a bit difficult to get details about the Wudao model. So, I was interested if you could answer a few questions about some of the technical details. You decided to use the GLM training objective, rather than a traditional LM, or mass language models. I was wondering if you have tried a LM or MLM, and if you found any benefits of using GLM compared to these other architecture s.

JIE: Yeah. So, we actually have a paper on Arxiv for the GLM, BERT is that auto encoding model and GPT-3 is an auto-regressive model. BERT is better at, for example, the natural language understanding and GPT-3 basically is better at generating something. Answer some kind of question if you ask. For example, dialogue system or something like that.

JIE: GLM basically unifies all of these together. So, it's a more general model. So, you can use GLM for all the different things. So, we have tests for NLU, natural language understanding and also generation, and even conditional generation. We get better performance. So that's it that the reason we want to promote the GLM model.

CONNOR: Wudao is a multimodal model. Could you elucidate a little bit how Wudao uses multimodal input and or output?

JIE: Okay. We actually use Chinese language and English language and also image. So, we actually combine all the different things together.

JIE: For example, you can put input some Chinese words, and it can answer in English in principle. And then of course you input English and it can answer you in Chinese.

JIE: On the other hand, you can input, for example, Chinese and the system can generate pictures for you. We have a demo for that.

JIE: So, you can put any kinds of words, any kind of words. For example, you can say, I want picture of a cat with sunglasses and the system will draw a cat sunglasses. But of course, DALL-E released by open AI can also do that. We verify that our performance can be much better.

CONNOR: I was curious how the architecture differs from DALL-E if there's any specific difference.

JIE: It's totally different, okay. To pre-train the model, we actually organize the image into different tokens also combine the texts and tokens together. All the code and the demo actually are online.

JIE: If you are Interested, you can try them and test the performance. It's quite open. The name of the system is CogView, COG VIEW. So, in that system we actually did some super resolution, which means we first generate some kind of token, like a load. And we up sampling to generate super resolution so performance is quite good. And we actually use this one to inversely generate an image caption, and to do the ranking. DALL-E generated 16 images and select which one is the best and show that best the one. But we only generate eight 80 images and then we use inverse image generation and that way, we actually can generate verb -perfect images.

JIE: We also did something for some specific domain,

JIE: for example, we collaborated with Alibaba to generate some images using GAN in that domain. That performance is a super good. you actually cannot distinguish, it's a real picture or not the generated one. So, we are actually quite different from DALL-E.

JIE: we have some technical contribution. So, for example, the autoregressive transformer like transforming itself. We also have precision bottleneck relaxation, PB relax, which makes the model, generally more stable.

JIE: Second, is a sandwich layer normalization, we call it Sandwich LN. That can also make the model much more stable. So, if you check our neural IPS paper, you will see that.

JIE: We are working to generate the video. We have the first version. Hopefully we can generate video using text based on that.

CONNOR: Wow. Very impressive. What is the plan with Wudao? Do you plan to continue to scale these types of models? Do you plan to release some kind of commercial application release the model to researchers? What's the plan?

JIE: We do have the plan to continue to increase the scale, for example, maybe a 10 terabyte and 100 terabytes. I will tell you that in June, we actually already trained a 100 trillion parameter model

CONNOR: wow, that's incredible.

JIE: It's not convergent. We didn't train the model until convergence. We just proved that we have the ability to do that.

JIE: So that’s our plan, but it is not urgent. We can do that quite easily. Yeah. We have a supercomputer. We can do that. It's in our plan. It's in our schedule.

JIE: We actually want to do some of application.

JIE: So, we build the association with more than 30 big companies in China. We hope that we can use the model to build some application to help them with millions or even billions of users. So that is more urgent for us.

JIE: Because we do want to show the benefit of the model to all the users. And also of course, to all the readers. So that's a second plan. And the third plan is we do hope that we can build some novel models so for example we want the generate video so that actually it's quite new. OpenAI didn't do that. And also,

we hope that the generative result could be correct. So, for example like GPT-3 and the GPT-2 they generate text and the image, but actually most of the generative result is not correct.

JIE: It's the same here. So, we hope that we can combine knowledge under the model itself together and maybe double check and make a generative result more accurate. We cannot make sure that it is correct, but at least more accurate. That is very important to us.

JIE: And also, we are working on fine the model. We call the P-tuning tool. So, this algorithm we've already published this paper and also, we have the code available. The P-tuning with less than 1% of the data can fine tune the model better.

JIE: So only one 1% of the data. So that'd give another direction or a totally new direction. Because fine tuning is very important for the pre-trained model.

JIE: And we are working with some chip company, and we're hoping that we can build a new underlying architecture for the big model. So that's another thing. So, we are working for that. So ideally finally and hopefully we can make this model do something beyond the Turing test that’s our final goal

CRAIG: On the number of parameters, so 10 times more parameters in GPT-3, but how do you measure its performance compared to GPT-3 and DALL-E. You've given some benchmarks, but do you generally consider it twice as capable or twice as useful? 10 times as capable or useful.

CRAIG: How meaningful is the parameter metric?

JIE: Compare with DALL-E, they did some kind of evaluation on MS COCO and they are using the matrix F I D. We did the same thing, and our performance is clearly better than them.

JIE: And second, we put it online. We can generate a picture in less than one minute. It's quite open. The performance is quite clear, much better than the DALL-E. So, you can test it. But the demo itself only support the Chinese language.

JIE: But for the first one for the evaluation, we did that in English, comparing with DALL-E using MS coco. For Wudao system compared with the GPT-3, we didn't do that because we cannot get the API.

JIE: So, we can only test our method on our machine using our data and also, we have English and the Chinese and also images. They only have English as, as far as I know.

CRAIG: Is there a plan to release a public API or a licensed API?

JIE: We do have the plan, but it probably will be later because it's expensive to host the API. We already have many partners in China. They just use the harder disk and copy the model to their company. The reason is that the model is too huge, and we do not want to host this.

CRAIG: And if scale is the differentiator for these models, what do you see as the limit for model size?

JIE: We already published more than 20 models based on Wudao. And we published those models and also API for those smaller models. That's quite open. You can just go to our website and check, the model, in Chinese and English, and also API. You send some kind of request and then access the model.

JIE: You can easily do that. But for the largest one, as I said, is difficult to, to host that because is simple expensive. But in the future, we will do that.

CRAIG: The hardware that you're training this on, you're using the Sunway supercomputer. Is that right?

JIE: Yeah, Sunway. The largest supercomputer, the largest. For those smaller models, like a 10 billion and the 100 billion, we train just a normal architecture.

CRAIG: You've talked about pre-training the world is Wudao 2.0 training ongoing. And is there a lifelong learning element to this? And does that mean that this podcast will find its way into the pre-training data?

JIE: Yeah. Very good questions.

JIE: We do have the plan to build something like a lifelong learning system, which can pretrain all the data continuously continue to improve the model. But we didn't train the largest one, just as I said, the 100 trillion model, but we didn't start with that one, because we want to find something novel some kind of new algorithm to do that. So, I still do some research on that.

CRAIG: Just generally China's AI research is advancing rapidly. What direction do you think is likely to lead to a breakthrough? And maybe it's what you just said.

JIE: I don't think we are the strong we are still learning from experts from United States and learn from experts from Europe. I believe we just work harder. So just, every day, probably 12 or maybe 14 hours, we just work. We generate so much data. We'll have more and more data, so maybe we can leverage there's reach of the data and build maybe a larger model.

CONNOR: Looking forward to the future, do you expect to see human level or superhuman level, even intelligence in our lifetimes?

JIE: Yeah, that's quite good question. Personally, I think that after 10 years or maybe 20 years, 10 to 20 years, I believe that the machine, the AI can do better than human beings in terms of most of the cognitive tasks, like as writing like a speech, if the task can be described into some kind of computational models. Okay. After 10 to 20 years, that is a generation or cognitive AI. I'm quite sure about that. So that's cognitive science. cognitive tasks. The machine will beat human beings. After 50 years, I believe that the machine will have consciousness.

JIE: we will have entered into a new world. And we will live with machines.

CONNOR: Some prominent figures recently, such as Stuart Russell from UC Berkeley have expressed some concerns about these kinds of things. They think that building extremely powerful, human, or human plus level, super-intelligent machines without any way to control them might be very dangerous. Even now, our agents and stuff do all kinds of bad things. we give those things superhuman powers. And that sounds very dangerous for the human race. Do you consider these risks?

JIE: Yeah, this is very important. We should have some kind of policy to control that. On the other hand, science. So, we should also see if it is possible. Can we really do that. If we cannot do that, why should we worry about that. If we can really do that with science, of course, we should worry.

CONNOR: Do you think it would be better if governments intervened more in research, or regulated research or hardware more? Do you think there's some kind of possibility for international treaties? Or do you think this is the wrong direction?

JIE: Yeah. I believe the government should do something for that. Not for pre-trained models, but for general AI. Because we are facing the generation of general AI, generation of cognitive AI. AI will probably be potential dangerous to a human being.

CONNOR: Do you think that's likely that maybe the governments would come together, do you think that's possible?

JIE: Yeah. I think it's possible.

CRAIG: You were at Cornell and the head of BAA I was at Microsoft and Baidu has a research center in California, so there's a lot of cross pollination, but at the same time, the two governments are working on developing military capabilities. About the research community is collaborative and it works across borders, but the governments are trying to take that research and they're not collaborative. Do you worry about that? How Do you see that playing out?

JIE: I'm quite positive personally, I'm quite positive to the situation. I'm quite positive to the science. Definitely, I worry about the collaborations. cutting the collaboration between China and the United States will, at least, delay the research in China. Because we do not have powerful chips.

JIE: We are quite open actually. I'm leading the team with more than 100 people and we build all the different models. We just release all the models, make it open to the world., to all the world. People can access the API because it's just science.

JIE: We build science for all human beings in the world. We want to build some kind of model, which can benefit all the people.

JIE: Even with such kind of a, serious situation, I still believe that we could do some even better then GPT-3. I defined three stages.

JIE: Stage one we follow GPT-3 and follow DALL-E. Stage two, from now or beginning of next year, we will do something special and some parts we will be better. And some part probably not so good, but we will build something comparable with Open AI and the stage three. Of course, we hope we can build the best. That's an ambitious goal.

JIE: Such kind of science is not for China. It's actually for all the world.

CRAIG: Is the entire Wudao source code on GitHub. Is the entire model public or just parts of it?

JIE: A few parts are not open because we also have some code on the supercomputer. You cannot use that code on the normal computer

CRAIG: So, this competition between china and the U S it's a government competition, but it's a collaborative competition among researchers globally. It's a more of an adversarial competition among governments and militaries. How do you handle that? Hopefully it won't lead to decoupling of the research communities, but I know in the U S there's a lot of pressure on the private sector on research institutes to cooperate with the military. A lot of the funding comes through DARPA. On the Chinese side, certainly a lot of the funding comes from the government. What choice do researchers have to avoid this adversarial relationship. '

JIE: To this question, different people will have different answers. I myself actually learned a lot when I visited the United States. So, like IBM and the Microsoft, I learned from United States. One coin, Has two sides. From the science, purely science, I only want to do something to advance the science.

JIE: So, we hope that we can do some science for the world, not just one country. On the other side, that is national research, which means we should do something on demand.

JIE: So that's based on the national project research plan. This is probably the same for all the people in the world.

CRAIG: You mentioned working on a novel algorithm, a new algorithm. Can you talk at all about where in the stack that algorithm will fit?

CRAIG: Is it something like mixture of experts to help training or is it a more basic algorithm like the transformer algorithm?

JIE: Yeah. That's good question. It is still on stage one. So, we are still working with the transformer system, transformer algorithm and we use the MOE so mixture expert. Actually, last year when Google published Moe, we actually already have the code. We also want to publish that paper, but Google released that paper. We have no chance to publish it. So, we actually, we developed the open-source code, we call it FastMoe. So, the code is totally open source on the website and each much faster than the MOE by Google.

JIE: That new algorithm actually underlying it's still transformer architecture, but the mechanism is different.

CONNOR: Thank you so much, Jie this has been really helpful and it's really nice to see the optimism on your side and I just really hope your research goes well and that we can continue to collaborate as a global science community.

CRAIG: That's it for this week's episode. I want to thank Professor Tang and Connor for their time. As I said at the outset, you can find a transcript of this episode on our website, eye-on.ai. I'm interested in hearing your thoughts on large language models, so please email me any comments at craig@eye-on.ai.

Remember, AI is about to change your world, so pay attention!