Video transcript

What are Generative AI models?

Over the past couple of months large language models or LLMs such as chat GBT have taken the world by storm. Whether it's writing poetry or helping to plan your upcoming vacation, we are seeing a step change in the performance of AI and its potential to drive enterprise value.

My name's Kate Soule. I'm a senior manager of business strategy at IBM Research and today I'm going to give a brief overview of this new field of AI that's emerging and how it can be used in a business setting to drive value.

Now large language models are actually a part of a different class of models called foundation models. Now the term foundation model was actually first coined by a team from Stanford when they saw that the field of AI was converging to a new paradigm where before AI applications were being built by training, maybe a library of different AI models, each AI model was trained on very task specific data to perform very specific task. They predicted that we were going to start moving to a new paradigm. Where we would have a foundational capability or a foundation model that would drive all of these same use cases and applications so the same exact applications that we are visioning before with conventional AI, the same model can drive any number of additional applications the point is that this model could be transferred to any number of tasks that gives this model this super power to be able to transfer to multiple different tasks and perform multiple different functions is that it's been trained on a huge amount in an unsupervised manner on unstructured data.

And what that means in the language domain is basically a bunch of sentences and I'm talking terabytes of data here, to train this model and the start of my sentence might be "no use crying over spilled" and the end of my sentence might be "milk" and I'm trying to get my model to predict the last word of the sentence based off of the words that it saw before and it's this generative capability of the model predicting and generating the next word based off of previous words that it seemed beforehand that is why that foundation models are actually a part of the field of AI called generative AI.

Because we're generating something new, in this case the next word in a sentence, and even though these models are trained to perform at its core a generation tasks, predicting the next word in the sentence, we actually can take these models and if you introduce a small amount of label data to the equation, you can tune them to perform traditional and op tasks things like classification or named entity recognition things that you don't normally associate as being a generative based model or capability, and this process is called "tuning", where you can tune your foundation model. By introducing a small amount of data, you update the parameters of your model and now it could perform a very specific natural language task. If you don't have data or have only very few data points, you can still take these foundation models and they actually work very well in low label data domains and in a process called "prompting" or "prompt engineering".

You can apply these models for some of those same exact tasks so an example of prompting a model to perform a classification task might be, you could give a model a sentence and then ask a question, is this sentence have a positive sentiment or negative sentiment? The model's going to try and finish generating words in that sentence and the next natural word in that sentence would be the answer to your classification problem. So, it would respond either positive or negative depending on where it estimated the sentiment of the sentence would be. And these models work surprisingly well when applied to these new settings and domains. Now, this is a lot of where the advantages of foundation models.

come into play. So if we talk about the advantages, the chief advantage is the performance. These models have seen so much data, again data with a capital D, terabytes of data that by the time that they replied to small tasks they can drastically outperform a model that was only trained on at just a few data points. The second advantage of these models are the productivity gains. So just like I said earlier, through prompting or tuning, you need far less labeled data to get to its task specific model then if you had to start from scratch, because your model is taking advantage of all the unlabeled data that it saw in its pre training when we created this generative task.

With these advantages, there are also some disadvantages that are important to keep in mind. And the first of those is the compute costs. So that penalty for having this model see so much data is that it's very expensive to train, making it difficult for smaller enterprises to train a foundation model on their own. They're also expensive, by the time they get to a huge size, a couple billion parameters, they're also very expensive to run inference. You might require multiple GPUs at a time just to host these models and run inference, making them a more costly method than traditional approaches.

The second disadvantage of these models is on the trustworthiness side so just like data is a huge advantage for these models, they've seen so much instructor data, it also comes at a cost. Especially in the domain like language, a lot of these models are trained basically off of language data that's been scraped from the internet and there's so much data that these models have been trained on. Even if you had a whole team of human annotators, you wouldn't be able to go through and actually vet every single data point to make sure that it wasn't biased and contained hate speech or other toxic information and that's just assuming you actually know what the data is. Often, we don't even know for a lot of these open-source models that have been posted, what the exact datasets are that these models have been trained on leading to trustworthiness issues.

So IBM recognizes the huge potential of these technologies, but my partners at IBM Research are working on multiple different innovations to try and improve also the efficiency of these models and the trustworthiness and reliability of these models to make them more relevant in a business setting.

All of these examples that I've talked through so far have just been on the language side, but the reality is, there are a lot of other domains that foundation models can be applied towards. Famously, we've seen foundation models on vision looking at models such as DALLE 2, which takes text data and that's then used to generate a custom image. We've seen models for code with products like Copilot that can help complete code as it's being authored.

IBM is innovating cross all these domains so whether it's language models that we're building into products like Watson Assistant and Watson Discovery, vision models that we're building into products like Maximo Visual Inspection or ansible code models that we're building with our partners at Red Hat under Project Wisdom. We're innovating across all these domains and more. We're working on chemistry. So, for example, we just published and released a MoLFormer, which is a foundation model to promote molecule discovery for different targeted therapeutics and were working on models for climate change building earth science foundation models, using geospatial data to improve climate research.

I hope you found this video both informative and helpful. If you're interested in learning more, particularly how IBM is working to improve some of these disadvantages making foundation models more trustworthy and more efficient, please take a look at the links below.

Thank you.