BRIAN KLAAS: Hi there. This is Brian Klaas. And in this lecture, we're going to look at how large language models work and how they fit into systems like ChatGPT or Bing Chat or Claude or Midjourney. Now, you are not-- I repeat not-- going to need a PhD in computer science to understand this. On the other hand, I'm also not going to show you how to create your own large language model or build one from an open-source repository. I'm not going to do that either.

And honestly, what most of what I'm going to say about large language models and artificial intelligence is going to be an oversimplification to anyone with an advanced computer science degree because this is a public health class. And my goal is to make this topic accessible and understandable, not perfect in all of its nuance.

And in this section of the lecture, we're going to take a step back and look at the big picture around AI, where we've come from so we know where we are and where we're going with generative AI in particular. So what is a large language model, to begin with? Well, I can explain it. Or we can ask our friend ChatGPT.

So ChatGPT with a basic prompt of "what is a large language model," tells me that it's a kind of AI that's trained on vast amount of text and learns statistical patterns and relationships within that text. OK, that makes sense. Now, if we wanted a more comprehensive answer, we could ask ChatGPT to explain it like they are a professor of artificial intelligence at Johns Hopkins University. So there's going to be a lot more jargon here. But again, it's talking about patterns and relationships between the patterns and the words inherent in the human language. And it learns those from a large data set.

And you'll learn more about prompt engineering in the next lecture. But it's also kind of interesting to see how does ChatGPT explain this if you're saying you're going to explain this to a group of high-school freshmen. So you're asking ChatGPT to take on a role, and it's going to explain it in a different way, which again, we'll learn more about a little bit later in the class. But again, the focus here is on taking a lot of information, a lot of text, and finding patterns and relationships in the language in that text so it can create a facsimile of a reasonable human response to the prompt.

So a large language model is really about modeling patterns, discovering and modeling patterns so that a system that uses that large language model at its core can then wind up giving reasonable responses to human inquiry. And this idea of patterns and artificial intelligence as pattern detector and pattern amplifier has been in existence through the entire history of artificial intelligence and major steps that artificial intelligence has taken every decade for the last five or even six decades.

So let's just take a step back. And before we look at actually how large language models work, let's look at a brief history of artificial intelligence so we can better understand the foundational knowledge and approaches that have been used that have led us to this kind of explosion in interest in generative AI, in particular, today.

So back in the day, in the 1980s, when artificial intelligence really started to actually become a part of practical computer science, artificial intelligence was really based on a series of rules, matching patterns based on rules. If you see x piece of data, then do y. Or if you see this very simple pattern, like the word "red," then do x.

Artificial intelligence has long been associated with the game of chess. Can you teach a computer to play chess and beat a human being? Well, in terms of rule-based AIs that they had back in the '80s, there were chess programs that you could play against.

But these were very simple algorithms, where it said, if the rook is in this position, do x. If the queen is in this square on the board, do y. These were long, long lists of "if this is true, then do this other thing" statements, hand coded by a human being with no room for improvisation or changing strategies or tactics on the fly because everything was hand coded to meet a very specific set of circumstances.

And then in the 1990s, machine learning really started to appear. And machine learning is just a means by which computer applications, software can create behavior by taking in data and then forming a model about that data. And the model really is just a simplification of some sort of more complex phenomenon in the real world.

So it takes that model and then executes on that model. Again, though, that model is based on recognizing patterns. And the AI application or software is about executing on that pattern. So the big distinguishing characteristic about machine learning was that the machines did this kind of pattern recognition all on their own.

So in terms of the chess example, with machine learning, chess programs were then able to learn strategies on their own and apply them to the playing of the game of chess, which is much more like what a human does on their own. And it was also during this time that we had our first really powerful example of machine learning applied to how humans input text and full semantically, syntactically correct sentences into a computer.

Translation and dictation software is truly foundational for building the language models which help current large language models like ChatGPT to understand how English works or how Japanese works or how Hindi works, for that matter.

And for those of you who may be old enough to remember, Dragon NaturallySpeaking speaking was like the first really big dictation software that came about and was publicly available. It launched in 1997. And this was really the first big step in taking machine learning and then applying it to everyday household computer use cases.

And machine learning ruled the roost, honestly, for about two decades, until the 2010s, when deep learning came into its own. So machine learning uses algorithms to parse data, learn from data, and make informed decisions based on what it has learned. Deep learning, however, structures algorithms or computer code itself into layers to create a kind of, well, what they call an artificial neural network, kind of like how our brain works, that can learn and make intelligent decisions on its own.

So machine learning was, again, a series of algorithms or rules or lines of software code. But deep learning was about layering multiple layers of algorithms or software code so that the machine could literally learn on its own. So deep learning just strives for independent decision-making with less ongoing human intervention than traditional machine learning.

And it also requires a lot-- and I mean a lot-- more computational power to do this work. And because dictation and translation software is so foundational to the original models that explain to a computer how the English language worked or how the Spanish language worked, this is about the same time that we got both Siri and Alexa, the ability for humans to speak clearly to a machine and have it do something in return.

And this ability really came about because of the advances in deep learning. On your phone, you can speak into the phone and have it do something and respond, or dictate a text message to someone. Well, while that's going on, there's all this background noise, which interferes and gets in the way of the computer understanding what you're saying.

But with the models developed through the technique of deep learning, computers were able to say, oh, this part of the sound I'm hearing is a human speaking. So I'm going to pay attention to that. And this other stuff, the car horn, the dog barking, and the wind blowing through the trees, although I'm picking that up, I have learned that's not what's important. That's what deep learning is about, being able to make those independent decisions. And that's why we saw the rise of tools like Siri, Alexa, dictation on your Android phone, whatever it might be during this time.

So fortunately, for deep learning researchers who are working on tools like Siri or Alexa or AlphaGo, around this same time, GPUs, or Graphical Processing Units, that are inside your computer, suddenly became much more capable and much more efficient. Now, you've probably heard of a CPU, or a Central Processing Unit. That's the main brain inside of your laptop. Windows or the macOS runs on that CPU. Google Chrome, Microsoft Word run on that CPU.

But GPUs were designed to create graphics. That's why they're called graphical processing units. And they were primarily used for CGI in movies and big special-effects movies, as well as your local Xbox or PlayStation. So as games on PlayStations and Xboxes suddenly became a lot more photorealistic in what they could render, researchers in artificial intelligence also discovered something very important.

So a lot of artificial intelligence data is stored in matrices, what we might call matrices from a mathematical perspective. But in software development, we call these vectors. They're not quite like vectors from geometry. But they are called vectors.

And graphical processing units use matrix math, use vector-based math to take a series of numbers and turn them into a three-dimensional object on the screen in front of you or in a movie. And as GPUs became more powerful and rendered more photorealistic results, back in the 2010s, artificial intelligence researchers realized, hey, wait a second, these things are really awesome at doing vector math.

And we store all of our data in vectors, or mostly in vectors. So why not tap into these GPUs instead of using CPUs, which aren't very good at vector math. And that's exactly what happened and launched a significant upgrade in artificial intelligence research.