## Video Transcript

## How Large Language Models Work

GPT, or Generative Pre-trained Transformer, is a large language model, or an LLM, that can generate human-like text. And I've been using GPT in its various forms for years.

In this video we are going to number 1, ask "what is an LLM?" Number 2, we are going to describe how they work. And then number 3, we're going to ask, "what are the business applications of LLMs?"

So let's start with number 1, "what is a large language model?"

Well, a large language model is an instance of something else called a foundation model. Now foundation models are pre-trained on large amounts of unlabeled and self-supervised data, meaning the model learns from patterns in the data in a way that produces generalizable and adaptable output. And large language models are instances of foundation models applied specifically to text and text-like things. I'm talking about things like code.

Now, large language models are trained on large data sets of text, such as books, articles and conversations. And look, when we say "large," these models can be tens of gigabytes in size and trained on enormous amounts of text data. We're talking potentially petabytes of data here.

So to put that into perspective, a text file that is, let's say, one gigabyte in size, that can store about 178 million words. A lot of words just in one Gb.

And how many gigabytes are in a petabyte?

Well, it's about 1 million. Yeah, that's truly a lot of text.

And LLMs are also among the biggest models when it comes to parameter count. A parameter is a value the model can change independently as it learns, and the more parameters a model has, the more complex it can be.

GPT-3, for example, is pre-trained on a corpus of actually 45 terabytes of data, and it uses 175 billion ML parameters.

All right, so how do they work?

Well, we can think of it like this. LLM equals three things: data, architecture, and lastly, we can think of it as training. Those three things are really the components of an LLM.

Now, we've already discussed the enormous amounts of text data that goes into these things. As for the architecture, this is a neural network and for GPT that is a transformer. And the transformer architecture enables the model to handle sequences of data like sentences or lines of code. And transformers are designed to understand the context of each word in a sentence by considering it in relation to every other word. This allows the model to build a comprehensive understanding of the sentence structure and the meaning of the words within it. And then this architecture is trained on all of this large amount of data.

Now, during training, the model learns to predict the next word in a sentence. So, "the sky is..." it starts off with a with a random guess, "the sky is bug".

But with each iteration, the model adjusts its internal parameters to reduce the difference between its predictions and the actual outcomes. And the model keeps doing this gradually improving its word predictions until it can reliably generate coherent sentences. Forget about "bug", it can figure out it's "blue".

Now, the model can be fine-tuned on a smaller, more specific data set. Here the model refines its understanding to be able to perform this specific task more accurately. Fine tuning is what allows a general language model to become an expert at a specific task.

OK, so how does this all fit into number 3, business applications?

Well, for customer service applications, businesses can use LLMs to create intelligent chatbots that can handle a variety of customer queries, freeing up human agents for more complex issues.

Another good field, content creation. That can benefit from LLMs which can help generate articles, emails, social media posts, and even YouTube video scripts. Hmm, there's an idea.

Now, LLMs can even contribute to software development. And they can do that by helping to generate and review code.

And look, that's just scratching the surface. As large language models continue to evolve, we're bound to discover more innovative applications. And that's why I'm so enamored with large language models.

If you have any questions, please drop us a line below.

And if you want to see more videos like this in the future, please like and subscribe.

Thanks for watching.