Let's start by understanding what generative AI is. At a superficial level or at a broader level, generative AI is all about AI generating content. Content in the form of text, images, video, or audio, given users input in the form of text, video, audio, or images. Now, let's confine ourselves to one of the applications of generative AI. That is, let's talk about deep learning models that are capable of taking users input in the form of text and generate output in the form of text. These models have been given a special name to them, which is large language models. In other words, large language models are deep learning models which are capable of generating text given a user's input in the form of text. Now, let's try to understand how these large language models work at a basic level. To understand this, we've got to trace back ourselves to 2017, where a few Google researchers are trying to tackle the problem of Google translation, or in other words, translation. As we already know, translation is nothing but taking a text which is in one language and translating it into another language. For example, converting French to English is an example of translation.

Now, earlier, there were some architectures called RNNs, which stands for recurrent neural networks or LSTMs, which are used to tackle this problem of translation. But there was one big problem with them. One cannot train massive models with these architectures, or in other words, one cannot build massive architectures using RNN and LSTMs, because inherently, these architectures are incapable of parallel processing. That is, you cannot put these models across the GPUs and train with a vast amount of data. So to tackle this problem, In 2017, Google researchers came up with a new architecture called Transformers, not only giving this architecture the ability for to parallel process, but also getting the accuracy which beats the RNNs and LSTMs by a large margin. Now, let's try to briefly understand how this transformer architecture works. Now, the transformer architecture at a broader level have two things. It is made up of two things or two blocks. First one is encoded block, and the second one is three-coded block. I'll give you a brief and layman introduction to how it works technically. Now, let's consider an example of translating a French sentence to an English sentence. So this French sentence is given as input to encoded blocks.

These encode blocks do some operations on this French input. And at the end of these stack of encoded blocks, they'll have some information in their memory about what exactly this French sentence is about. And this understanding, which is in the form of numbers, is given to the decoder blocks along with some hints, along with some of the words in the translated English sentence. And with the combined understanding of French sentence and English words, it will try to guess the next word, or it'll try to translate that particular French sentence into English sentence. So briefly, encoder blocks, try to understand the input sentence. Decoder blocks try to decode that input sentence into a language that we are trying to translate. In this case, it is English. Now, essentially, that's how a transformers work. Now, we haven't answered the question of how large language models work. So to understand this, we have to talk a bit about OpenAI. So right after the release of Transformers, some of the OpenAI, similar to Google

research, some of the OpenAI researchers have read the paper of Transformers, which is titled, Attention is all you need. And and thought to themselves, why not develop?

Why not take the decoder blocks of the transformer and train a model that is capable of taking some input text and generating some output text. So thus came an architecture called generative... Came an architecture which is known as GPT, which stands for generative pre-trained transformer, which eventually led to the popular tool, which is known as ChatGPT, as the world knows. Now, ChatGPT is based on GPT's third version. Gpt is basically trained on or trained using decode this block of transformers architecture. Okay, let's end this discussion here. I'll just summarize whatever I have said so far. Initially, in summary, we have We've talked about a couple of things. We have started by understanding what generative AI is, and then we have journeyed through an application of generative AI, which is known as Large Language Models. And then we have understood the architecture behind this, behind large language models, which make them, which basically make them run on GPUs with the ability of parallel processing. And we have talked briefly about how transformers work on a superficial level. And then we have talked about generative pre-trained transformers, also known as GPT, which is a decoder version, which is a decoder version of transformers.

So these are the four points that talked about. I hope you understand something from this. Thank you.