2. Fundamentals of text generation | Module 2

[00:00:12] Sil But how does a language model work? In this video I'll go through the fundamentals of text generation. Along the way, we'll learn what hallucinations are. In the last video I explained language models are designed to do one thing and one thing only.

[00:00:25] Sil For some sequence of words, what is the probability of that sequence appearing in nature? Language models are designed to produce this probability. Now, you might think that this probability on its own, just a percentage, isn't very interesting, but there's a lot you can do with it.

[00:00:42] Sil For example, United States of America is a very common sequence of words. United States of America. It's more common than the United States of Pizza or United States of Canada. In understanding that United States of America is more common than Pizza or Canada, we're implicitly saying that America is more likely to follow the United States of than Pizza or Canada is.

[00:01:09] Sil Now, this means that you can use language models to give you the next most likely word. For example, if you were to give a language model, the words "United States of" it might say that it's 60% probable that "America" would follow. And then perhaps it might say that it's 15% likely for "Mexico" to follow. And then some infinitesimal value for "Canada" or "Pizza". So given that it's, say, 60% likely that America would follow, let's pick that word.

[00:01:38] Sil Now we have the string "United States of America". Now we can give the language model the string back and see what it thinks is, again, most likely to follow. "United States of America is" or "the" or whatever word, and so we pick that word and we repeat this process over and over and over again.

[00:01:58] Sil Language models like Chat GPT or Bard or Anthropics Claude are always predicting the next word. We just ask them what word is likely to follow so quickly it looks like the model's writing. When you ask Chat GPT a question, you'll find that it produces word by word, by word by word. This is because we're continually giving a language model that string of text that Chat GPT has already produced. We give it to Chat GPT, it produces the next word, and we repeat and repeat and repeat.

[00:02:26] Sil If you make the model that produces this word after word after word big enough, that is to say if you give it enough computational power, enough artificial neurons, and you give it enough examples of words drawn from the Internet, it turns out that the words of pigs become informed by logic and knowledge of the world. It becomes a world model, so to speak. Its predictions become so good, the model even appears to become intelligent. But it's important to remember that even if the model seems like it could do math and write essays and all that other good stuff, what it actually is doing underneath the hood always remains the same. It's always predicting the next most likely word.

[00:03:06] Sil But sometimes the model isn't sure what words should come next, so it confabulates. It stretches its knowledge. It suddenly does what we might call a hallucination. It slips. Hallucinations are when the model appears to lie or produce false information. Examples of hallucinations have appeared in the news.

[00:03:25] Sil For example, a lawyer a few months ago, preparing for a lawsuit, asked Chat GPT for relevant precedents. Chat GPT appeared to comply: it produced list of valid
cases. But when the judge reviewed the precedents, the judge found that the cases were all made up. Chat GPT had lied.

[00:03:43] Sil But it's important to remember language models do not necessarily lie. It's important to remember they're not conscious, it does not have intention. The model is only ever trying to realistically continue the sequence you give it. So hallucinations are not malicious, but it's unlikely that they'll ever be solved and they are certainly a problem. The model can't ever know everything about the world, but it'll always do its best to continue the sequence that you give it.

[00:04:10] Sil Humans, of course, do the same. Have you ever stretched the truth in knowledge demands you feel less confident about? But in situations where truthful information matters, hallucinations are definitely problematic. If you're generating an article, for example, containing information you know people depend on, can you really trust the language model to do your copywriting for you?

[00:04:31] Sil It's important to be aware of hallucinations when considering Chat GPT in your newsroom. Hallucinations, to repeat, are not malicious, but language models are not perfect. We're slowly figuring out that there are circumstances where we can get the language model to hallucinate less, and this is including a technology called Retrieval Augmented Generation. You should look into it if you're seriously considering using Chat GPT in your newsroom.

[00:04:57] Retrieval Augmented Generation involves having the language model do research before writing an article. It's a bit like having the model act like a journalist. The model has a data set, for example it might be the archives of your newsroom. When given a certain topic, it will go through all the articles and all the vetted, truthful information that exists in your archives, and it will pull relevant pieces of information and produce an article that is less likely to have been produced by a hallucination. Retrieval Augmented Generation and other similar technologies like langchain and agents, ensure that your model is less likely to lie to you.