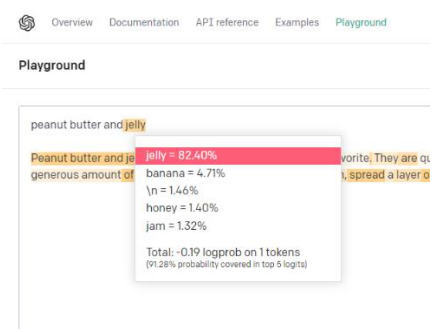
Part I of III in a series on Artificial Intelligence written by our colleague Alan D. Mead, Ph.D.

Artificial Intelligence 101 – How it Works

To be a good driver, you don't need to know deep technical details about your car (e.g., how fuel injection works), but it's hard to imagine a good driver who does not know basic details about how a car works. And, even worse, if you didn't know anything about how cars (and oceans) worked, you might reason, "Hey, my car can take me to Chicago, so it can also take me to Paris!"

A lot of people don't seem to have a basic level of understanding of how popular AI applications like ChatGPT and Bard work. This post will explain how **large language models** (LLMs) work with a minimum of technical details. After we explain how LLMs work, we'll examine some things they can do well and things that they cannot do well (and why).

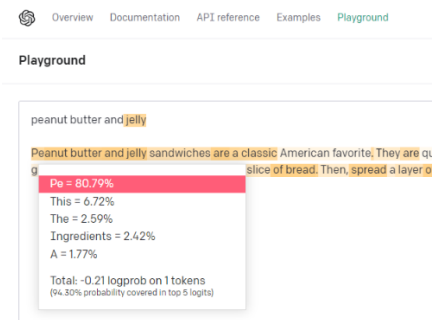
The first thing to know about LLMs is that they are good at **associating** written things (and soon, multi-model LLMs will be good at associating written, audio, and visual things). "Associating" means being able to predict which words a person would expect to "come next" when the model is given some words, like a question.



For example, if I say: "peanut butter and ___", you probably fill in "jelly". The words we send to an LLM are called a prompt. If we send the prompt: "peanut butter and" GPT-3 (usually) replies with "jelly" and some information about PB&Js. A special interface called the GPT-3 **playground** allows us to see the probabilities of the words it chooses. The word "jelly" is about 82% likely. Other responses include "banana" (4.7% likely), "honey" (1.4% likely), and "jam" (1.3% likely), and there are many other possibilities (not shown) with very small probabilities.

The model correctly predicts that "jelly" is very associated with "peanut butter" and the model's less likely responses (banana, honey, and jam) are also very good choices, compared to all the words in the English language.

This is what all LLMs are good at: processing one piece of written text to find associated words,



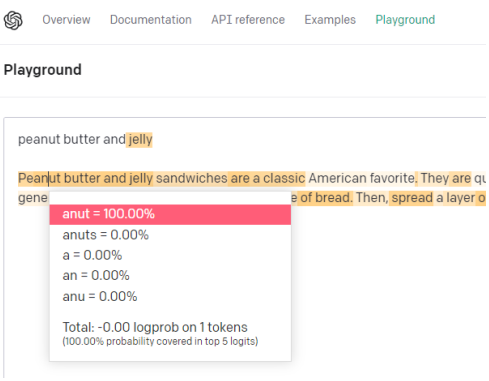
and "good at" means that model predictions often match what people would produce. When you give an LLM a prompt, the input is encoded so the model can process it, and then the model calculates a set of possible words and their probabilities. Then one word is chosen at random according to those probabilities. For example, if you prompted the model "peanut butter and" 100 times, you would get "jelly" as an answer about 82 times, on average.

rick@talentalignment.net www.talentalignment.net

But the model doesn't stop working once it gives you "jelly." It then adds the new word, "jelly" to the prompt and runs again.

This example shows a detail of how the model works. It does not actually predict the next word; it predicts the next **token**. When discussing LLMs, a token means a whole or partial word, a number, or punctuation. LLMs encodes input by turning words into a series of tokens. The reason why the models use tokens is fascinating, but beyond the scope of this post.

Continuing with the example, the next token generated by the model is "Pe" with probability just shy of 81%. Once that token is generated, the next token is "anut" with almost 100% probability. In fact, the other (tiny) possibilities are smaller fragments that would also spell out "Peanut." Because the model is composing text about this sandwich being classic.



The model continues this process, generating more and more of the response, until it stops. Either the model randomly chooses a **stop sequence** token, or else it hits a maximum number of tokens. If you have interacted with LLMs, you've probably seen this happen when the response just cuts off mid-sentence.

In summary, a LLM composes a response. The response is built up, one token at a time. For each new token, the model calculates probabilities based on your input prompt and the tokens it has generated do far. And when the model finishes this series of predictions, the response is generally syntactically correct English that is associated with the input prompt.

"Follow" our LinkedIn page for the rest of our three-part series featuring Alan:

"User Beware – AI Models Can Hallucinate and Deceive!" (Part II)
"Six Rules of Engagement for AI Users" (Part III)

Check out our thoughts on advancing strategy through Culture Change while you're there!

rick@talentalignment.net
www.talentalignment.net