04-01 What are tokens?

Video Link: https://learn.microsoft.com/en-sg/shows/generative-ai-for-beginners/introduction-to-generative-ai-and-llms-generative-ai-for-beginners?WT.mc_id=academic-105485-koreyst

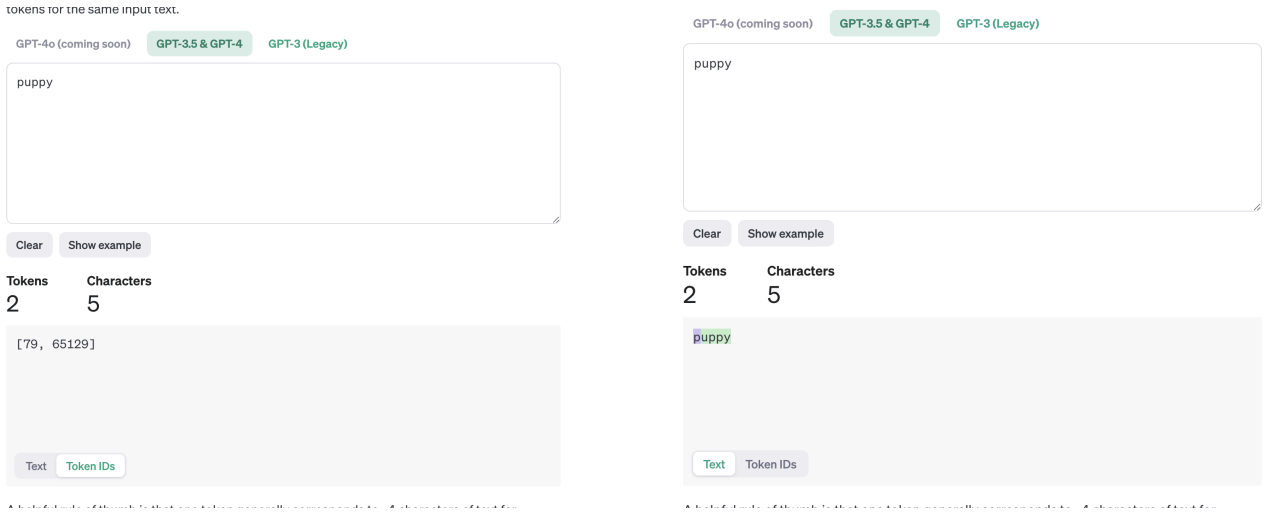
OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text

The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).

https://platform.openai.com/tokenizer

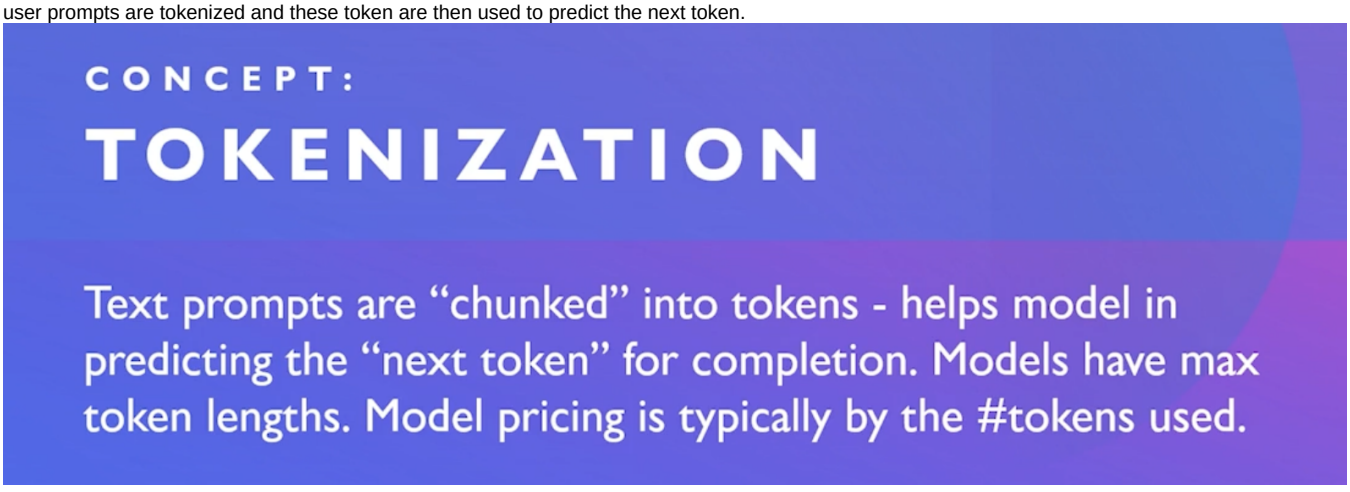
Can play around in the above url and understand how it is tokenized(although it still look like black box 😃)



Try
"My favorite color is red."
"My favorite color is Red."
"Red is my favorite color"

here red would have different token Id

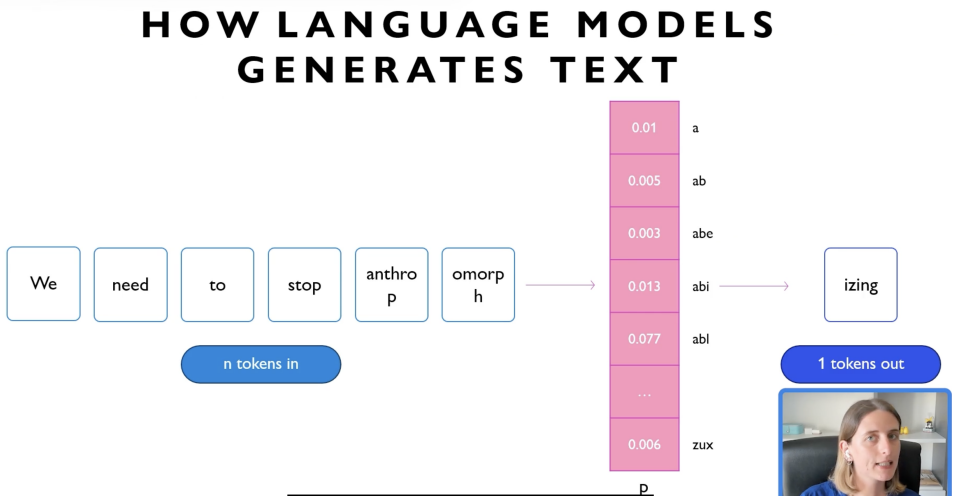
user prompts are tokenized and these token are then used to predict the next token.



Reason its tokenized is because model works better with numbers than raw text.

Tokenizer, text to numbers: Large Language Models receive a text as input and generate a text as output. However, being statistical models, they work much better with numbers than text sequences. That's why every input to the model is processed by a tokenizer, before being used by the core model. A token is a chunk of text – consisting of a variable number of characters, so the tokenizer's main task is splitting the input into an array of tokens. Then, each token is mapped with a token index, which is the integer encoding of the original text chunk.

Each token generated in output is then used as input for the next token while generating the model.



In the above example, the model chooses the output based on the probability of its occurance in the current sequence. this is calculated using its training data.
Model doesn't always select the same token and this is due to the usage of randomness that has been introduced in the selection process(using temperature as an example)

GPT takes in the prompt, convert it into tokens, process the prompt and convert the output token into words

Also check out Parameters vs Token
Create tokenization in this youtube video