Tokenizer

Learn about language model tokenization

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.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

It's important to note that the exact tokenization process varies between models. Newer models like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and will produce different tokens for the same input text.

GPT-3.5 & GPT-4 GPT-3 (Legacy)

```
Cliente2: Backoffice 40% amarillo, Caja 40% azul, Comercial 20%
rojo. Perteneciente a Multi-canal.
Cliente3: Backoffice 20% amarillo, Caja 60% azul, Comercial 20%
rojo. Perteneciente a Caja.
Cliente4: Backoffice 20% amarillo, Caja 10% azul, Comercial 70%
rojo. Perteneciente a Comercial.
Cliente5: Backoffice 27% amarillo, Caja 27% azul, Comercial 45%
rojo. Perteneciente a Comercial.
```

Clear Show example

Tokens
191

Characters
534

```
Los prompts para cada cliente son los siguientes: Cliente1: Backoffice
42% amarillo, Caja 25% azul, Comercial 33% rojo. Perteneciente a Multi
-canal.
Cliente2: Backoffice 40% amarillo, Caja 40% azul, Comercial 20% rojo. P
erteneciente a Multi-canal.
```

```
Cliente4: Backoffice 20% amarillo, Caja 10% azul, Comercial 70% rojo. P
erteneciente a Comercial.
Cliente5: Backoffice 27% amarillo, Caja 27% azul, Comercial 45% rojo. P
erteneciente a Comercial.
```

Text Token IDs

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).

If you need a programmatic interface for tokenizing text, check out our tiktoken package for Python. For JavaScript, the community-supported @dbdq/tiktoken package works with most GPT models.