

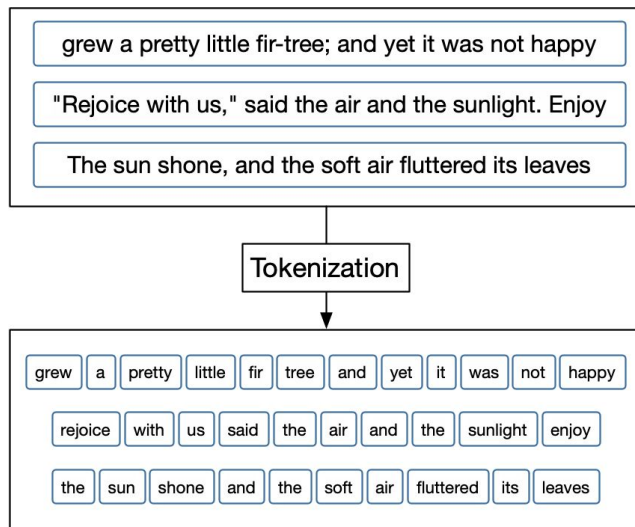
What are Tokens?

Examine token limits what are and how to gracefully handle limits.

What Is Tokenization?

— — —

Sentences, words and characters are transformed into lots of *individual tokens*.



Token Limits

— — —

- *Restricted by the models* that we choose to use.
- Each model has an input token limit and output token limit.

Exercise: View OpenAI Pricing

— — —

Navigate to [OpenAI's pricing](#) by Google searching
“OpenAI pricing”

Explore the different models and their pricing for:

- Input tokens
- Cached input tokens
- Output tokens

How to get the token limit for ChatGPT and GPT-X

— — —

<https://platform.openai.com/tokenizer>

Tokenizer

The GPT family of models process text using **tokens**, which are common sequences of characters found in text. The models 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 would be tokenized by the API, and the total count of tokens in that piece of text.

GPT-3 Codex

This is some text

Clear

Show example

Tokens

5

Characters

18

This is some text

How to get the token limit - Automatically

— — —

```
import tiktoken
enc = tiktoken.get_encoding("o200k_base")
assert enc.decode(enc.encode("hello world")) == "hello world"

# To get the tokeniser corresponding to a specific model in the OpenAI API:
enc = tiktoken.encoding_for_model("gpt-4o")
```

Approaches To Avoid Hitting Token Limits

— — —

- Use different models
- Shortening your prompt/context
- Chunking
- Windowed chunks
- Summarisation