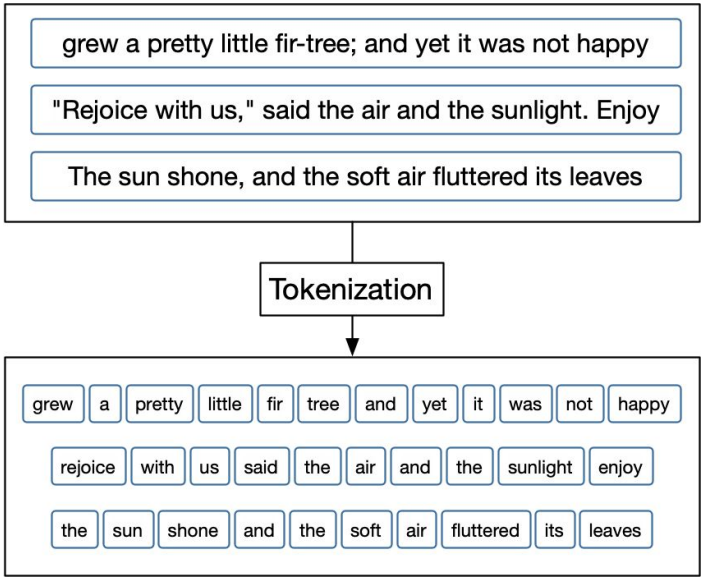
What are Tokens?

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

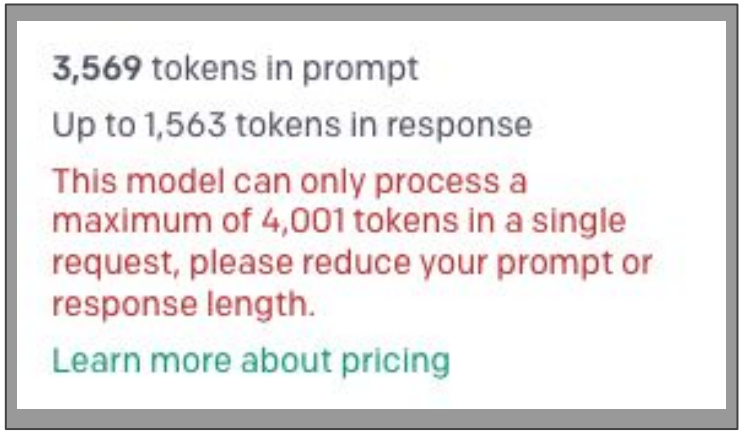
What Is Tokenization?

- Sentences are transformed into lots of individual tokens.



Token Limits

- Restricted by the models that we choose to use.
- Each model has a unique token limit.



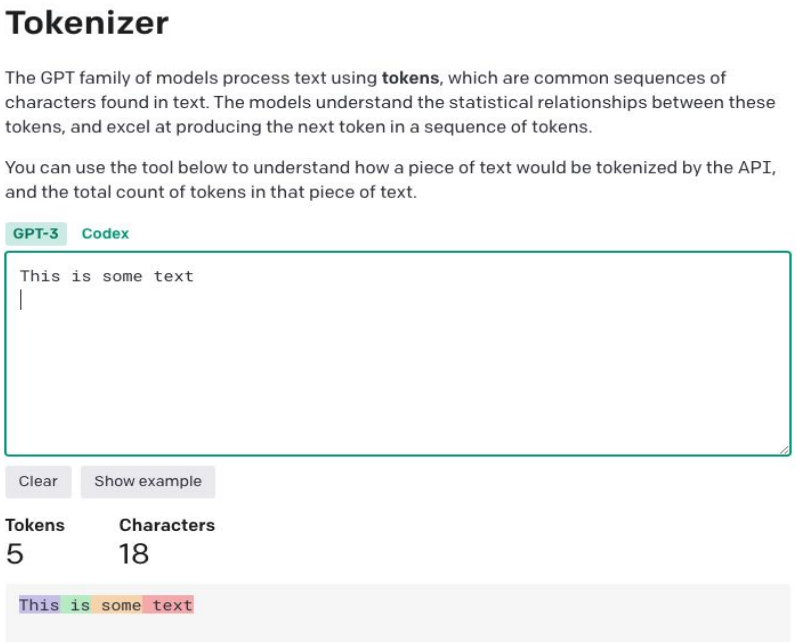
Standard Token Limits

LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
text-davinci-003	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality, longer output and better instruction-following. Also supports inserting completions within text.	4,000 tokens	Up to Jun 2021
text-curie-001	Very capable, but faster and lower cost than Davinci.	2,048 tokens	Up to Oct 2019
text-babbage-001	Capable of straightforward tasks, very fast, and lower cost.	2,048 tokens	Up to Oct 2019
text-ada-001	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,048 tokens	Up to Oct 2019

How to get the token limit for ChatGPT and GPT-3

---

https://platform.openai.com/tokenizer



How to get the token limit - Automatically

```
import tiktoken
enc = tiktoken.get_encoding("gpt2")
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("text-davinci-003")
```

Approaches To Avoid Hitting Token Limits

- Shortening the prompt
- Chunking
- Windowed chunks
- Summarisation