Tokenization in Practice - Real Examples

See How Different Texts Are Tokenized

Example 1: Simple Text

Input:

```
"The quick brown fox"
```

GPT-3.5 Tokens:

`"The"` `" quick"` `" brown"` `" fox"`

4 tokens

Example 2: Technical Terms

Input:

```
"API authentication token"
```

GPT-3.5 Tokens:

`"API"` `" authentication"` `" token"`

3 tokens

Example 3: Numbers & Special Chars

Input:

```
"Price: $29.99 (USD)"
```

GPT-3.5 Tokens:

`"Price"` `":"` `" $"` `"29"` `"."` `"99"` `" ("` `"USD"` `")"`

9 tokens

Example 4: Code Snippet

Input:

```
"def hello(): print('Hi!')"
```

GPT-3.5 Tokens:

`"def"` `" hello"` `"():"` `" print"` `"('"` `"Hi"` `"!')"` `"\n"`

8 tokens

Language Comparison

English: "Hello world" → 2 tokens

Spanish: "Hola mundo" → 3 tokens

Chinese: "你好世界" → 4 tokens

Code: `print("hello")` → 4 tokens

🔍 Observations

- Spaces are often part of tokens (e.g., `" quick"`).
- Punctuation usually gets separate tokens.
- Familiar words are more efficiently tokenized.
- Code and technical terms may use more tokens.