토큰(Token)

Fast campus

토큰(Token)은 자연어 처리(NLP)에서 텍스트를 작은 단위로 나누어 처리하기 위해 사용되는 기본 단위입니다. 단어, 부분 단어, 문자 등이 토큰이 될 수 있습니다.

LLM(대형 언어 모델)에서 토큰은 텍스트 데이터를 모델이 이해하고 처리하기 위해 분할된 기본 단위입니다. **텍스트를 토큰으로 나누는 과정을 '토큰화'** 라고 합니다.

토큰화의 방법

토큰화 방법에는 여러 가지가 있으며, 사용하는 방법에 따라 토큰의 정의가 달라질 수 있습니다.

- **문자 기반 토큰화**: 텍스트를 문자 단위로 나누는 방법입니다.
  - 예시: "Hello" → ["H", "e", "l", "l", "o"]
- **단어 기반 토큰화**: 텍스트를 단어 단위로 나누는 방법입니다.
  - 예시: "Hello, world!" → ["Hello", ",", "world", "!"]
- **서브워드 기반 토큰화**: 단어를 더 작은 단위(서브워드)로 나누는 방법입니다. **자주 사용되는 서브워드** 를 기준으로 분할합니다.
  - 예시: "unhappiness" → ["un", "happiness"]
  - 📌 [참고] BPE(Byte Pair Encoding): 자주 등장하는 문자 쌍을 합쳐가며 서브워드를 생성하는 알고리즘입니다.

토큰의 중요성

토큰은 모델이 텍스트를 이해하고 처리하는데 핵심적인 역할을 합니다. 토큰화의 결과에 따라 모델의 성능이 크게 영향을 받을 수 있습니다. 잘 정의된 토큰화 방법을 사용하면 모델이 텍스트의 의미를 더 정확하게 파악할 수 있습니다.

- **문맥 이해**: 모델이 문맥을 이해하고 적절하게 응답할 수 있게 도와줍니다.
- **효율성**: 적절한 크기의 토큰을 사용함으로써 연산 자원을 효율적으로 사용할 수 있습니다.

예시

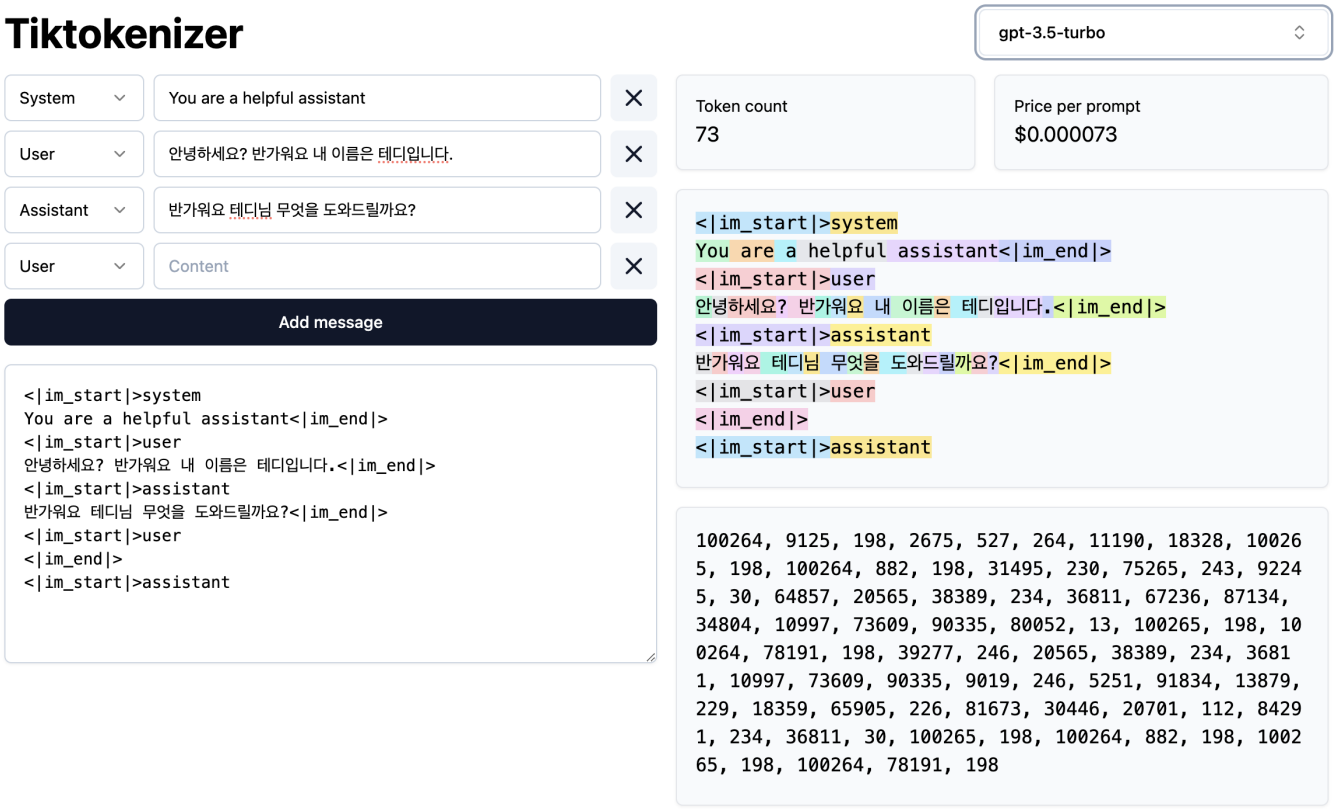
- 단어 기반 토큰화의 예시:
  - 텍스트: "Machine learning is fun."
  - 토큰화 결과: ["Machine", "learning", "is", "fun", "."]
- 서브워드 기반 토큰화의 예시(BPE 사용):
  - 텍스트: "Machine learning"
  - 토큰화 결과: ["Ma", "chine", "learn", "ing"]

토큰화는 자연어 처리의 기본적인 단계이므로, LLM을 이해하고 사용하는데 필수적인 개념입니다. 다양한 토큰화 방법을 잘 이해하면, 모델의 성능을 최적화하는데 큰 도움이 될 것입니다.

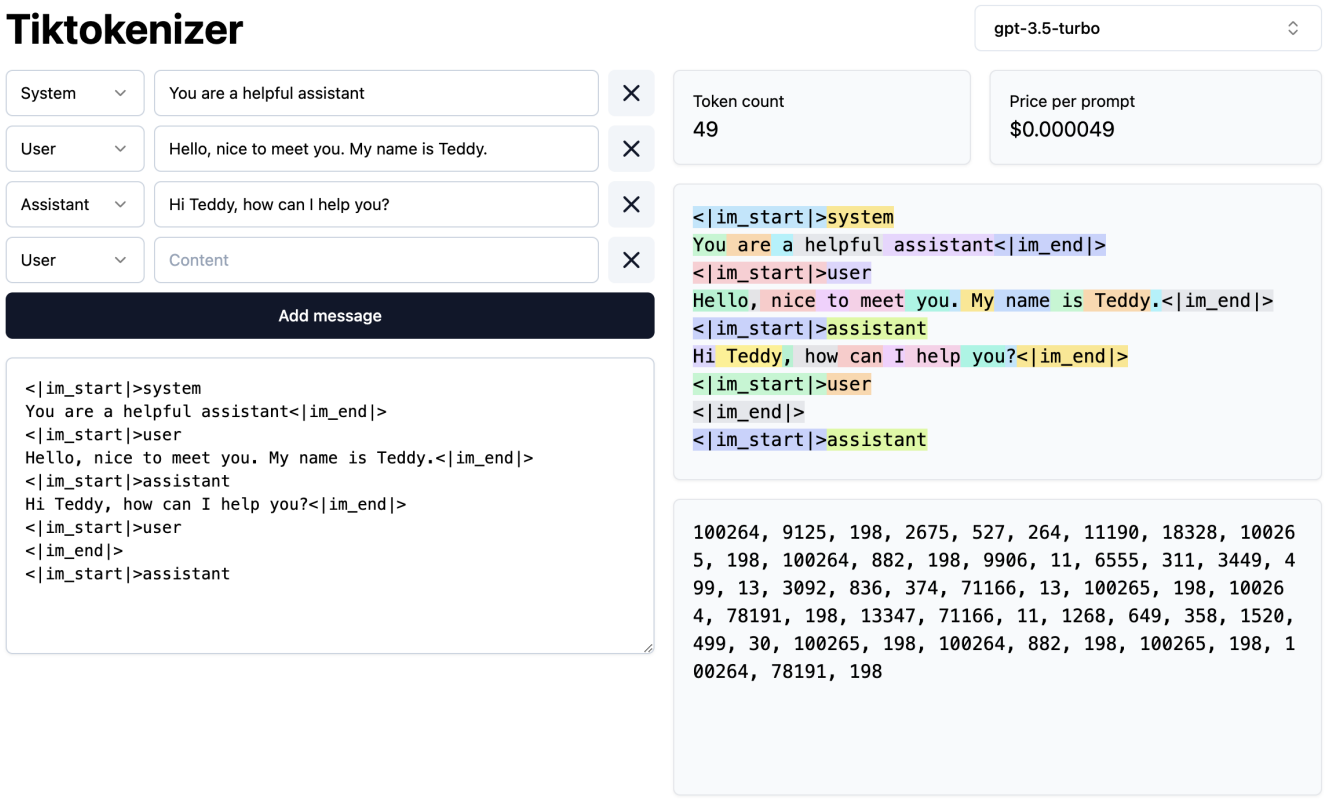
토큰 사용량 = 돈$

이처럼 LLM은 입력 텍스트를 다양한 방법으로 토큰화할 수 있으며, 어떤 방식이 사용되느냐에 따라 모델의 성능과 효율성이 달라질 수 있습니다. 토큰화 방법을 선택할 때는 모델의 목적과 텍스트 데이터의 특성을 고려해야 합니다.

`gpt-3.5-turbo` 모델의 **한글** 토큰 사용량과 요금



`gpt-3.5-turbo` 모델의 **영문** 토큰 사용량과 요금



`gpt-4-turbo` 모델의 **한글** 토큰 사용량과 요금

Tiktokenizer

gpt-4-1106-preview

System	You are a helpful assistant	
User	안녕하세요? 반가워요 내 이름은 테디입니다.	
Assistant	반가워요 테디님 무엇을 도와드릴까요?	
User	Content	

Add message

```
<|im_start|>system<|im_sep|>You are a helpful
assistant<|im_end|><|im_start|>user<|im_sep|>안녕하세요? 반가워요 내
이름은 테디입니다.<|im_end|><|im_start|>assistant<|im_sep|>반가워요
테디님 무엇을 도와드릴까요?<|im_end|><|im_start|>user<|im_sep|>
<|im_end|><|im_start|>assistant<|im_sep|>
```

Token count
132

Price per prompt
$0.00132

<|im_start|>system<|im_sep|>You are a helpful assistant<|im_end|><|im_start|>user<|im_sep|>안녕하세요? 반가워요 내 이름은 테디입니다.<|im_end|><|im_start|>assistant<|im_sep|>반가워요 테디님 무엇을 도와드릴까요?<|im_end|><|im_start|>user<|im_sep|><|im_end|><|im_start|>assistant<|im_sep|>

```
27, 91, 318, 5011, 91, 29, 9125, 27, 91, 318, 55875, 9
1, 29, 2675, 527, 264, 11190, 18328, 27, 91, 318, 634
5, 91, 1822, 91, 318, 5011, 91, 29, 882, 27, 91, 318,
55875, 91, 29, 31495, 230, 75265, 243, 92245, 30, 6485
7, 20565, 38389, 234, 36811, 67236, 87134, 34804, 1099
7, 73609, 90335, 80052, 16134, 91, 318, 6345, 91, 182
2, 91, 318, 5011, 91, 29, 78191, 27, 91, 318, 55875, 9
1, 29, 39277, 246, 20565, 38389, 234, 36811, 10997, 73
609, 90335, 9019, 246, 5251, 91834, 13879, 229, 18359,
65905, 226, 81673, 30446, 20701, 112, 84291, 234, 3681
1, 76514, 91, 318, 6345, 91, 1822, 91, 318, 5011, 91,
29, 882, 27, 91, 318, 55875, 91, 1822, 91, 318, 6345,
91, 1822, 91, 318, 5011, 91, 29, 78191, 27, 91, 318, 5
5875, 91, 29
```

`gpt-4-turbo` 모델의 **영문** 토큰 사용량과 요금

Tiktokenizer

gpt-4-1106-preview

System	You are a helpful assistant
User	Hello, nice to meet you. My name is Teddy.
Assistant	Hi Teddy, how can I help you?
User	Content

Add message

```
<|im_start|>system<|im_sep|>You are a helpful
assistant<|im_end|><|im_start|>user<|im_sep|>Hello, nice to
meet you. My name is Teddy.<|im_end|>
<|im_start|>assistant<|im_sep|>Hi Teddy, how can I help you?
<|im_end|><|im_start|>user<|im_sep|><|im_end|>
<|im_start|>assistant<|im_sep|>
```

Token count
108

Price per prompt
$0.00108

```
<|im_start|>system<|im_sep|>You are a helpful assistan
t<|im_end|><|im_start|>user<|im_sep|>Hello, nice to me
et you. My name is Teddy.<|im_end|><|im_start|>assista
nt<|im_sep|>Hi Teddy, how can I help you?<|im_end|><|i
m_start|>user<|im_sep|><|im_end|><|im_start|>assistant
<|im_sep|>
```

```
27, 91, 318, 5011, 91, 29, 9125, 27, 91, 318, 55875, 9
1, 29, 2675, 527, 264, 11190, 18328, 27, 91, 318, 634
5, 91, 1822, 91, 318, 5011, 91, 29, 882, 27, 91, 318,
55875, 91, 29, 9906, 11, 6555, 311, 3449, 499, 13, 309
2, 836, 374, 71166, 16134, 91, 318, 6345, 91, 1822, 9
1, 318, 5011, 91, 29, 78191, 27, 91, 318, 55875, 91, 2
9, 13347, 71166, 11, 1268, 649, 358, 1520, 499, 76514,
91, 318, 6345, 91, 1822, 91, 318, 5011, 91, 29, 882, 2
7, 91, 318, 55875, 91, 1822, 91, 318, 6345, 91, 1822,
91, 318, 5011, 91, 29, 78191, 27, 91, 318, 55875, 91,
29
```

`gpt-4o` 모델의 **한글** 토큰 사용량

Tiktokenizer

gpt-4o

System	You are a helpful assistant
User	안녕하세요? 반가워요 내 이름은 테디입니다.
Assistant	반가워요 테디님 무엇을 도와드릴까요?
User	Content

Add message

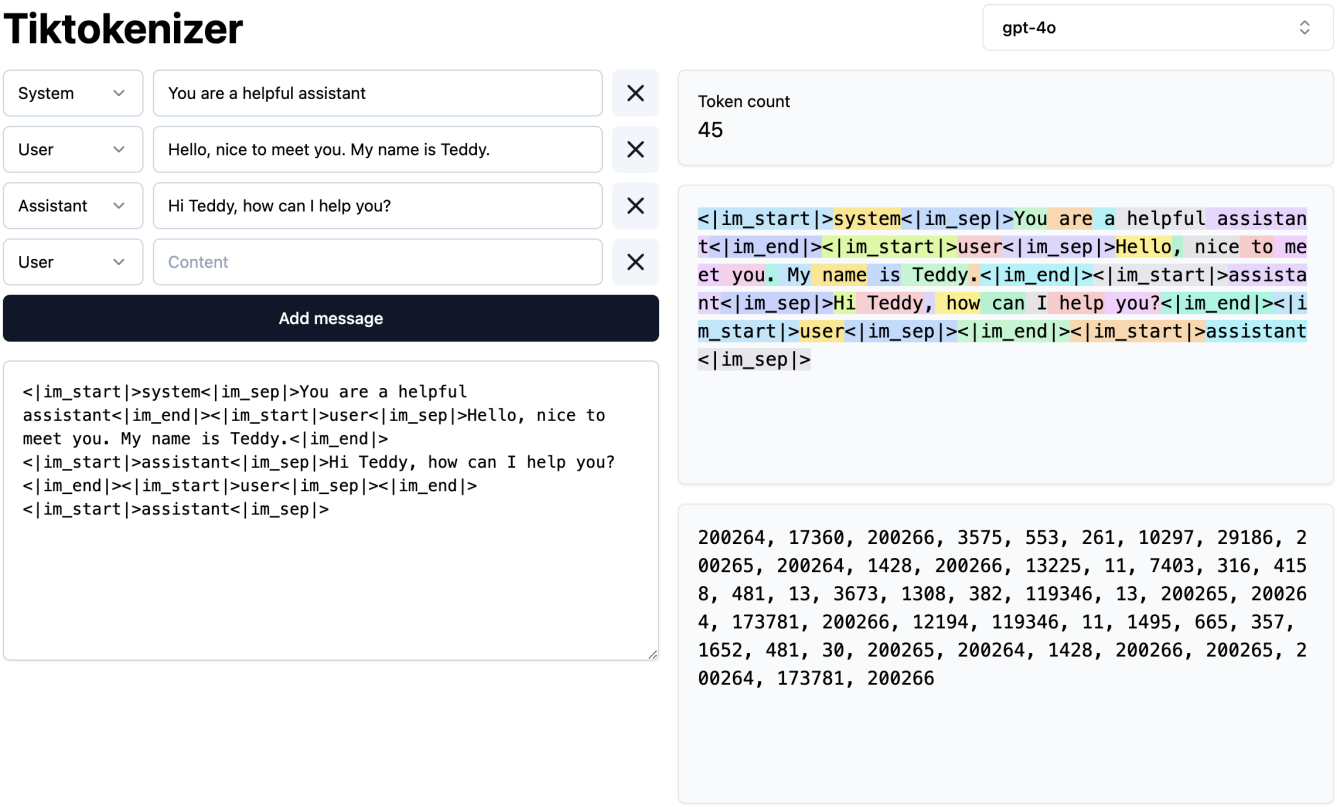
```
<|im_start|>system<|im_sep|>You are a helpful
assistant<|im_end|><|im_start|>user<|im_sep|>안녕하세요? 반가워요 내
이름은 테디입니다.<|im_end|><|im_start|>assistant<|im_sep|>반가워요
테디님 무엇을 도와드릴까요?<|im_end|><|im_start|>user<|im_sep|>
<|im_end|><|im_start|>assistant<|im_sep|>
```

Token count
53

```
<|im_start|>system<|im_sep|>You are a helpful assistan
t<|im_end|><|im_start|>user<|im_sep|>안녕하세요? 반가워요 내
이름은 테디입니다.<|im_end|><|im_start|>assistant<|im_sep|>
반가워요 테디님 무엇을 도와드릴까요?<|im_end|><|im_start|>user<
|im_sep|><|im_end|><|im_start|>assistant<|im_sep|>
```

```
200264, 17360, 200266, 3575, 553, 261, 10297, 29186, 2
00265, 200264, 1428, 200266, 14307, 171731, 30, 35007,
4081, 33771, 7952, 21566, 78825, 4740, 74754, 21198, 2
7001, 13, 200265, 200264, 173781, 200266, 23099, 4081,
33771, 7952, 74754, 21198, 28012, 103740, 3281, 27433,
12753, 9389, 70984, 157244, 30, 200265, 200264, 1428,
200266, 200265, 200264, 173781, 200266
```

`gpt-4o` 모델의 **영문** 토큰 사용량



참고

- 모델별 토큰 계산
- 토큰 계산기