Tokens in Generative AI

Estimated Time: 30 minutes

Objectives

After completing this reading, you will be able to:

- Define tokens, token limits, and exploring tokens
- Identify and explain token tools
- Describe the process of counting tokens
- Explain the OpenAI Guidelines for Tokens Count
- Summarize how to estimate AI costs
- List the current pricing of AI models

What are Tokens in AI APIs? :

Tokens are crucial fragments of ChatGPT API. Tokens are fragments or segments of words. Before processing prompts, the API breaks down the input into these individual tokens.

These tokens do not necessarily align exactly with the start or end of words — they can include trailing spaces and even sub-words.

Token limits: Usually, requests can employ up to 4097 tokens shared between the prompt and completion, depending on the model. For instance, if your prompt consists of 4000 tokens, your completion can contain a maximum of 97 tokens. This limit is currently a technical constraint, but there are often creative ways to work within it, such as condensing your prompt or breaking the text into smaller chunks.

Token pricing: The API provides various model types at different price points. Each model possesses a range of capabilities, with "gpt-3.5-turbo" being the most capable. Requests made to these different models have different prices. Detailed information regarding token pricing is available on the product API page.

Exploring tokens: The API treats words based on their context in the corpus data. GPT-3 takes the prompt, converts the input into a list of tokens, processes the prompt, and then converts the predicted tokens back into the words as a response. Note that what may appear as two identical words are generated as different tokens depending on their structure within the text. For example, consider how the API generates token values for the word "red" based on its context.

Token Tools

The most popular of all tools is the OpenAI interactive tokenizer tool. This tool helps calculate the number of tokens and observe how text is broken down into tokens.

Alternatively, if there is a need to tokenize text programmatically, use Tiktoken, a fast BPE tokenizer specifically designed for OpenAI models.

Other libraries include the transformers package for Python or the gpt-3-encoder package for Node.js.

How to count tokens?

To count tokens for calling an OpenAI, follow these steps:

- Identify the API endpoint: Determine which endpoint of the OpenAI you plan to call.
- Each endpoint represents a specific function or resource that the API provides.
- Review the API documentation: Access the API documentation provided by the API provider. The documentation will outline the structure of the API requests and responses, including any required headers, parameters, or authentication methods.
- Check if API needs token-based authentication: Token-based authentication involves including an access token in the request header to authorize the API call. The token is typically obtained by registering an application with the API provider and following their authentication process.
- Obtain an access token: If token-based authentication is required, obtain an access token by following the authentication process specified by the API provider. This action may involve registering an application, providing credentials, and receiving a token.
- Count the tokens for each API: Once you have an access token, count it as one token for each API call you make. Each request to the API endpoint, including the access token in the header, counts as a single token.
- Track token usage: Keep track of the number of tokens you have used to ensure you stay within any usage limits or quotas the API provider sets. Some APIs may restrict the number of tokens you can use within a certain period.

OpenAI guidelines for tokens count

As per the official document of OpenAI, below are the estimates for the token count:

- 1 token ~= 4 chars in English
- 1 token ~= ¾ words
- 100 tokens ~= 75 words
- 1-2 sentences ~= 30 tokens
- 1 paragraph ~= 100 tokens
- 1,500 words ~= 2048 tokens

Example: Counting tokens for API calls

For example, you have an OpenAI endpoint that requires a GET request to retrieve a user's information. The endpoint has the following details:

- Request Method: GET
- Endpoint URL: https://api.example.com/users/{userId}
- Path Parameter: userId (value: "123")
- Query Parameter: includeAddress (value: true)

Counting the tokens for this request:

- The request method 'GET' typically requires 1 token.
- The endpoint URL doesn't consume additional tokens.
- The path parameter 'userId\' with value '123' doesn't consume additional tokens.
- The query parameter 'includeAddress' with the value 'true' consumes 2 tokens (1 token for the parameter name and 1 token for the parameter value).
- Summing up the tokens from steps 1 to 4, the total token count for this API request would be 3 tokens.

Note: The exact method of counting tokens may vary depending on the specific API implementation or API provider. Consult the API documentation for accurate information on token usage and any associated costs or limitations.

Estimating AI Costs

Example: Calculating the cost for an input prompt and output using a sample cost calculation formula.

Input prompt: Prompt: “Can you provide a brief overview of the benefits of exercise?”

- Step 1: Determine the number of words in the prompt.
  - In this case, the prompt has 9 words.
- Step 2: Calculate the cost for the prompt based on the number of tokens. Assume the cost per 1000 tokens is $0.03 (as mentioned in the pricing chart).
  - Since there are 9 words in the prompt and the average conversion rate is 750 words per 1000 tokens, the cost for the prompt = (9 / 750) * (0.03) = $0.00036
- Step 3: If the AI model generates an output of approximately 500 words, use the same formula to calculate the cost of generating the output. Output cost = (500 / 750) * (0.06) = $0.04
- Step 4: Calculate the total cost. Total estimated price = Prompt cost + Output cost = $0.00036 + $0.04 = $0.04036

Therefore, the estimated price for generating an output of approximately 500 words using the given prompt is **$0.04036**.

Similarly, If the application calls the API 1000 times a day, the cost is shown as follows:

- Cost per day = $40.36
- Cost per month = $1,210.80

Note: The actual cost may vary depending on the exact number of tokens used in generating the content and the type of model used.

Current Pricing of Models

Disclaimer: The pricing chart provided is accurate as of the time of writing. However, it is subject to change in the future. We recommend consulting the documentation for the most up-to-date and accurate pricing information.

Under consideration: Multiple models, each with different capabilities and price points.

Prices are per 1,000 tokens. Since tokens are pieces of words in this model, 1,000 tokens is about 750 words. [This paragraph is 35 tokens].

Pricing chart based on the model:

Model	Name	Input/Usage	Output/Usage
GPT-4 Turbo	gpt-4-1106-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
	gpt-4-1106-vision-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
GPT-4	gpt-4	$0.03 / 1K tokens	$0.06 / 1K tokens
	gpt-4-32k	$0.06 / 1K tokens	$0.12 / 1K tokens
GPT-3.5-Turbo	gpt-3.5-turbo-1106	$0.0010 / 1K tokens	$0.0020 / 1K tokens
	gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens
Code Interpreter	Code Interpreter	$0.03 / session	
DALL·E 3	DALL·E 3 Standard	1024×1024 - $0.040 / image	
DALL·E 3	DALL·E 3 HD	1024×1024 $0.080 / image	

Summary

In this reading, you learned that:

- Before processing prompts, the API breaks down the input into these individual tokens. Tokens are fragments or segments of words.
- OpenAI interactive and Tiktoken are common tokenization tools.
- To count tokens, identify the API endpoint, review the API documentation, check for API authentication, obtain an access token, count the token for each API, and track its usage.

Author(s)

Zehra Afzal

Skills Network