

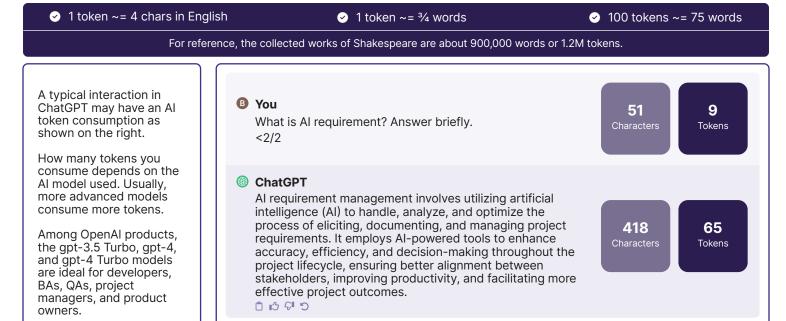


## **Tokens in Copilot4DevOps**

## What are Al Tokens?

Al Tokens are like pieces of words in natural language processing. As a user, you consume tokens based on the length of your input and output. One token may include characters, spaces, and punctuation marks.

In English, input and output tokens may break down the following ways:



## Al Token Limits and Costs

The maximum number of tokens you can consume is called your **context window**. Here are some examples:

Model	Context Window	Token Consumption Example
ChatGPT-3.5 Turbo	16k tokens	Input: 2000 tokens Output: 1500 tokens Consumption = 3500 tokens
ChatGPT-4 Turbo	128k tokens	Input: 4000 tokens Output: 4192 tokens Consumption = 8192 tokens

The number of tokens is usually mentioned in the following format: 250,000 4k tokens. This refers to the limits of what a model can handle. 250,000 4k means:

- √ 250,000 is the total number of tokens you can consume
- √ 4k (or 4000) tokens is the maximum number of tokens you can consume per interaction

## Token Usage in Copilot4DevOps

Copilot4DevOps is available in several models, with a special type of token called a CP4DO Token¹. The conversion between CP4DO Tokens and OpenAl tokens is as follows:<sup>2,3</sup>

1 CP4DO Token = 1 gpt-3.5 Turbo Token 16k context

60 CP4DO Token = 1 gpt-4 Turbo Tokens 128k context

- For most use cases, gpt 3.5 is ideal as it is the most cost-effective option.
  However, gpt 4 Turbo often gives the best results, as it can process larger input/output sizes.
- 2. Token equivalency is subject to change.
- . As new gpt models become available, Modern Requirements may add them as an option.

