Tokens and Cost: Explained Simply

This guide explains how AI apps count words, and how much they cost. You do not need any background. Every new word is explained the first time you see it.

> **The big idea in one line:** an AI reads and writes in small pieces called **tokens**. You pay for the pieces it reads and the pieces it writes.

1. What is a token?

A token is a small piece of text. It can be a whole short word, part of a long word, or a space or a full stop.

A simple rule: **one token is about four letters.** So a 100-letter message is about 25 tokens.

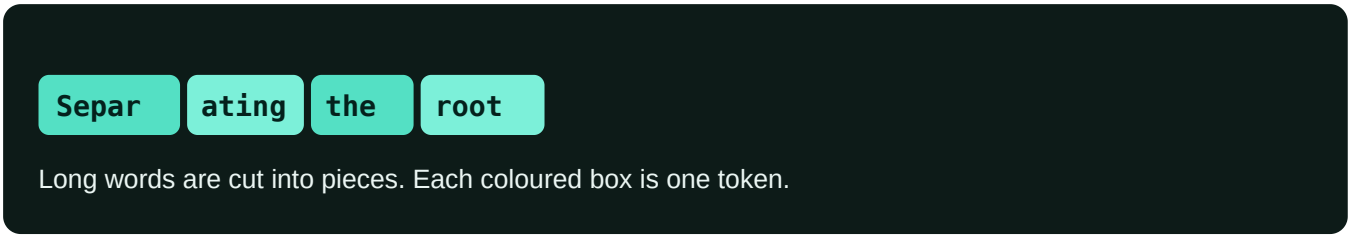


Figure 1. A long word can be more than one token.

2. Input and output

- **Input** is everything you send to the AI. This includes your instructions and the message.
- **Output** is everything the AI writes back.
- **Total** is input plus output. This is the number that matters for cost.

Example: a food app. A seller types a dish name, and the app asks the AI for a short description.

Part	Tokens (example)
Instructions (sent every time)	about 55
Dish name, type and price	about 15
Input total	**about 70**
Description written back	about 65
Total for one request	**about 135**

The instructions are sent again every time. That is why a long instruction costs a lot over many requests.

3. How much does one request cost?

AI companies charge a price for each **million** tokens. The example below uses old class prices. **These are not today's prices.** Always check the price page on the day you decide.

Easy steps:
1. Input cost = input tokens ÷ 1,000,000 × input price
2. Output cost = output tokens ÷ 1,000,000 × output price
3. Add them together.

Step	Working	Result
Input	70 ÷ 1,000,000 × $0.80	$0.000056
Output	65 ÷ 1,000,000 × $4.00	$0.000260
One request	add them	**about $0.0003**

One request is tiny. But 5,000 requests a month is about $1.58 at these example prices. The cost grows with the number of requests, so you must plan for volume.

4. Three quick tips to use fewer tokens

1. **Write shorter instructions.** Remove words like "please" and "very" when they add nothing.
2. **Set a limit on the answer.** Ask for "under 50 words" and set a maximum length.
3. **Send only what is needed.** Do not send a whole web page when a title is enough.

5. The context window: the AI's desk

The **context window** is how much the AI can hold at one time. Think of it as a desk. Your messages, its answers and any files all go on the desk. When the desk is full, older things must be removed.

Sizes depend on the model. Check the model's page for the current size.

Tip: start a new chat when you change topic. A fresh desk costs less.

6. How the AI picks a word

The AI looks back at earlier words and decides which ones matter most for the next word. For example, in "Riya poured tea into the cup and added some", the words *tea* and *cup* matter most for the word *some*.

Then it picks the next word. Two settings control this:

Setting	What it means in plain words	Use it when
Temperature	Low = safe and the same each time. High = more surprising.	Low for data formats. Medium for writing.
Max tokens	A hard limit on how long the answer can be.	Always. It stops long, costly answers.

7. Which AI model should I pick?

Pick the smallest model that does the job well. Test it on about 20 examples first.

Job	Start with
Sort things into groups	A small, fast model
Write a short description	A small, fast model
Compare two answers and choose	A bigger model

8. Word bank

Word	Meaning
Token	A small piece of text. About four letters.
Input	What you send to the AI.
Output	What the AI writes back.
Context window	How much the AI can hold at one time.
Temperature	How safe or surprising the answer is.
Max tokens	The longest the answer can be.

9. Check yourself

1. About how many letters is one token?
2. Which costs more per token, input or output?
3. Why is a long instruction costly?
4. Which setting limits the length of an answer?

Answers: 1. About four. 2. Output. 3. It is sent again with every request. 4. Max tokens.