interweb

Generated on 15/11/2026 00:12

KNOWLEDGE BASE · PUBLIC DOCUMENT

Tokens explained: how AI model APIs are counted

How AI Works

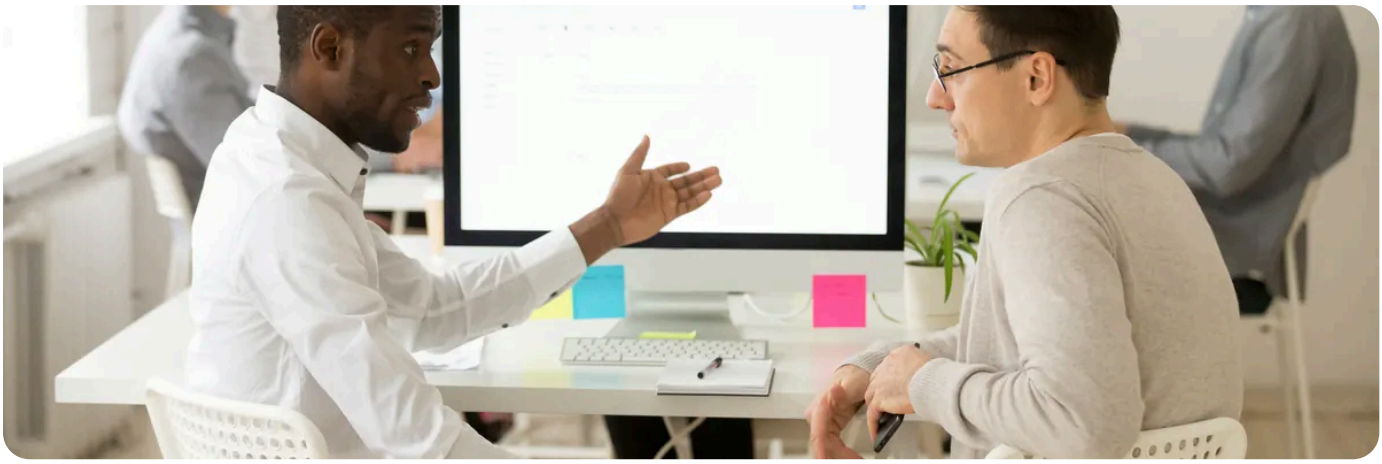
CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-936	v2026.10.10	Angola · English

Interweb Angola
RLX Angola, Lda. · Group RLX Holdings
Tax ID (NIF) 500 260 5724 · Commercial registration no. 26464-25/250724

https://www.meu.interweb.ao/knowledgebase/936

Tokens explained: how AI model APIs are counted

How AI Works Created on **October 10, 2026** 3 min read



A **token** is a piece of text: a short word, part of a long word, a punctuation mark or a space. AI APIs count tokens, not words or characters, to measure **what the model read and what it wrote**, and tokens are what they charge for. Understanding this is understanding why a long conversation costs more than it looked.

We do not write the price of any model here, because it changes. The price per token is on the **provider's pricing page**, and what each model used shows in the provider's console.

What goes into the count

What counts	Example	Note
What you send (input)	Your question, the instructions, the documents you attach	Includes the system instructions an application adds on its own.
What the model writes (output)	The reply	As a rule, providers price output differently from input. Check the pricing page.
The conversation history	The earlier messages	It is usually resent with every request, so the cost of a conversation grows.
Tool results	A page read, a search, a database queried	Everything the model "reads" to answer counts as input.

The three facts that explain almost everything

1. **Each model counts in its own way.** The same text can give different token numbers on models from different providers. Use the provider's counting tool if there is one, or the usage field the API response returns.

2 **Other languages usually cost more.** In general, Portuguese uses more tokens than English to say the same thing. The difference depends on the model, so measure with your own text.

3 **There is a context ceiling.** Each model only reads a certain volume of tokens at once (the "context window"). If the history, the documents and the reply do not fit, the model cuts or fails. The value is in the model's documentation.

How to spend less without losing quality

1 **Start a fresh conversation** when the subject changes, instead of dragging a history that no longer helps.

2 **Send only what matters.** Do not attach a whole document to ask about one paragraph.

3 **Ask for short answers** when they are enough. Output is paid for too.

4 **Use the right model.** Simple tasks do not need the biggest one. See **which tool suits which task**.

5 **Measure before you automate.** Run the task by hand a few times, check the usage in the console, and only then schedule it.

An agent working on its own spends tokens even when you are not looking.
Set a spending limit at the provider before you schedule tasks. See **what an agent costs**.

A subscription to a chat app and an API key are different things, with different accounts. Paying for one does not put credit on the other.

NEED A HAND?

Have a server or a site where you want to plug in AI and do not know where to start? Tell us what you have in mind.

Open a support ticket →

SEE ALSO

Getting an API key for an AI model, and keeping it safe →

What OpenClaw costs: the API keys, not the server

ChatGPT, Claude or Gemini: which AI tool suits which task

interweb

Interweb Angola

RLX Angola, Lda. · Group RLX Holdings

TAX ID (NIF)	500 260 5724
COMMERCIAL REGISTRATION NO.	26464-25/250724
REGISTERED OFFICE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACT	info@interweb.ao · +244 942 011 191 · www.interweb.ao
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

RLX-AO-KB-936 · v2026.10.10 · Generated on 15/11/2026 00:12

Online version of this document