interweb

Generated on 27/11/2026 09:55

KNOWLEDGE BASE · PUBLIC DOCUMENT

What a chat request costs through the API: the reservation, max_tokens and the final adjustment

All in One AI

CODE	VERSION	COUNTRY · LANGUAGE
RLX-MZ-KB-2878	v2026.11.26	Mozambique · English

Interweb Mozambique
RLX Holdings, S.A. · Group RLX Holdings
Tax ID (NUIT) 401 191 607 · Legal entity no. (NUEL) 105 045 974 · Economic activity code (CAE) 62022

https://www.meu.interweb.co.mz/knowledgebase/2878

What a chat request costs through the API: the reservation, max_tokens and the final adjustment

All in One AI Created on **November 26, 2026** 2 min read



In short: before calling the model, we reserve credits for the worst case. At the end we charge for the tokens actually used and return the rest to your balance. What weighs most in the reservation is `max_tokens`, so always set it.

The three steps of every request

1. **The reservation.** We count the input (about one token for every 3 characters) plus `max_tokens` of output. Without `max_tokens`, the reservation counts 4096 output tokens.

2. **Your balance is checked.** If it does not cover the reservation, the answer is `402` with `insufficient_credits` and nothing is charged. If it does, the model replies.

3. **The adjustment.** We charge for the real tokens, reported in `usage`, and the rest of the reservation goes back to your balance.

An example with numbers

With a model at 56 credits per million input tokens and 112 per million output tokens, a request with 1000 input tokens and 500 output tokens costs 0.056 + 0.056 = 0.112 credits. It is rounded up to 0.12. The minimum per request is 0.01 credits. Each model's real prices are on the Models tab.

When you get a 402

The error includes `details.required`, what the reservation needed, and `details.balance`, what you have. There are two ways out: **buy credits** or lower `max_tokens`. The reply often fits in far less than 4096.

When streaming, ask for the count.

With `stream: true`, the token count comes in the last event. If the model does not send it, the whole reservation is charged. Always send `stream_options` with `include_usage`. See **showing the reply as it is written**.

Where to see what you spent

On the History tab of All in One AI, each request shows the model, the key, the input, output and cached tokens, and the credits spent. Through the API, your balance comes from `GET /v1/ai/balance`. See also **tokens explained**.

To spend less:

Keep conversations and instructions short, and use the cheapest model that does the job. Before a big batch, check your balance and the model's price.

RECOMMENDED

Want to see what each request cost? Open your All in One AI History.

See your History →

SEE ALSO

- **Choosing a text model: price, context window and what each one is for**
- **AI credits: how to buy them, the volume tiers and what each request uses**
- **Showing the AI reply as it is written in your website or app (streaming)**
- **Tokens explained: how AI model APIs are counted**
- **Chat and text in All in One AI**

interweb

Interweb Mozambique

RLX Holdings, S.A. · Group RLX Holdings

TAX ID (NUIT)	401 191 607
LEGAL ENTITY NO. (NUEL)	105 045 974
ECONOMIC ACTIVITY CODE (CAE)	62022
REGISTERED OFFICE	Bairro 7 de Setembro Vilankulo 1304 · CEP 0411-02, Inhambane, Mozambique
CONTACT	info@interweb.co.mz · +258 84 635 1111 · +258 86 635 1111 · www.interweb.co.mz
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-MZ-KB-2878 · v2026.11.26 · Generated on 27/11/2026 09:55