interweb

Gerado a 27/11/2026 08:52

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Quanto custa uma conversa pela API: a reserva, o max_tokens e o acerto no fim

All in One AI

CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-AO-KB-2878	v2026.11.26	Angola · Português

Interweb Angola
RLX Angola, Lda. · Grupo RLX Holdings
NIF 500 260 5724 · Matrícula 26464-25/250724

https://www.meu.interweb.ao/knowledgebase/2878

Quanto custa uma conversa pela API: a reserva, o max_tokens e o acerto no fim

All in One AI Criado em **26 de novembro de 2026** 2 min de leitura



A resposta curta: antes de chamar o modelo, reservamos créditos para o pior caso. No fim, cobramos os tokens que foram mesmo usados e devolvemos o resto ao saldo. O que mais pesa na reserva é o `max_tokens`. Por isso defina-o sempre.

Os três passos de cada pedido

1. **A reserva.** Contamos a entrada (cerca de um token por cada 3 caracteres) e o `max_tokens` de saída. Sem `max_tokens`, a reserva conta 4096 tokens de saída.

2. **O saldo é conferido.** Se não chegar para a reserva, a resposta é `402` com `insufficient_credits` e nada é cobrado. Se chegar, o modelo responde.

3. **O acerto.** Cobramos pelos tokens reais, que vêm em `usage`, e o resto da reserva volta ao saldo.

Um exemplo com números

Com um modelo a 56 créditos por milhão de tokens de entrada e 112 por milhão de saída, um pedido com 1000 tokens de entrada e 500 de saída custa 0,056 + 0,056 = 0,112 créditos. O valor é arredondado para cima: 0,12. O mínimo por pedido é 0,01 créditos. Os preços reais de cada modelo estão no separador Modelos.

Quando aparece o 402

O erro traz `details.required`, o que a reserva precisava, e `details.balance`, o que tem. Há duas saídas: **comprar créditos** ou baixar o `max_tokens`. Muitas vezes a resposta cabe num valor bem menor do que 4096.

No streaming, peça a contagem.

Com `stream: true`, a contagem chega no último evento. Se o modelo não a enviar, cobramos a reserva inteira. Envie sempre `stream_options` com `include_usage`. Veja **como mostrar a resposta aos poucos**.

Onde ver o que gastou

No separador Histórico do All in One AI, cada pedido mostra o modelo, a chave, os tokens de entrada, de saída e em cache, e os créditos gastos. Pela API, o saldo sai em `GET /v1/ai/balance`. Veja também **os tokens explicados**.

Para gastar menos:

Conversas curtas, instruções curtas e o modelo mais barato que faça o trabalho. Antes de um lote grande, confirme o saldo e o preço do modelo.

RECOMENDADO

Quer ver o que cada pedido gastou? Abra o Histórico do All in One AI.

Ver o Histórico →

VEJA TAMBÉM

- **Escolher um modelo de texto: preço, contexto e para que serve cada um** →
- **Créditos de IA: como se compram, os escalões de preço e o que cada pedido gasta** →
- **Mostrar a resposta da IA aos poucos no seu site ou aplicação (streaming)** →
- **Tokens explicados: como se contam as APIs de modelos de IA** →
- **Chat e texto no All in One AI** →

interweb

Interweb Angola

RLX Angola, Lda. · Grupo RLX Holdings

NIF	500 260 5724
MATRÍCULA	26464-25/250724
SEDE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACTOS	info@interweb.ao · +244 942 011 191 · www.interweb.ao
HORÁRIO	2ª. à 6ª. Feira - 08 à 16 Horas Sábado & Domingo - Fechado

Versão online deste documento

RLX-AO-KB-2878 · v2026.11.26 · Gerado a 27/11/2026 08:52