interweb

Gerado a 28/11/2026 21:58

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Antes de um lote grande de pedidos de IA: ver primeiro o saldo e o preço

Créditos de IA e custos

CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-MZ-KB-2980	v2026.11.28	Moçambique · Português

Interweb Moçambique
RLX Holdings, S.A. · Grupo RLX Holdings
NUIT 401 191 607 · NUEL 105 045 974 · CAE 62022

https://www.meu.interweb.co.mz/knowledgebase/2980

Antes de um lote grande de pedidos de IA: ver primeiro o saldo e o preço

Créditos de IA e custos Criado em **28 de novembro de 2026** 2 min de leitura



Mandar centenas ou milhares de pedidos de uma vez é onde um erro sai caro. Uns minutos de contas antes poupam créditos e evitam um lote parado a meio por falta de saldo.

1. Ver o preço do modelo

Na aba **Modelos** ou pela API, por exemplo à procura de um modelo com «flash» no nome:

```
curl "https://api.interweb.co.mz/v1/models?type=text&q=flash" -H "Authorization: Bearer $INTERWEB_API_KEY"
```

2. Fazer a conta

Por pedido: (tokens de entrada × preço de entrada + `max_tokens` × preço de saída) / 1.000.000. Para estimar a entrada, conte cerca de um token por cada 3 caracteres. Depois multiplique pelo número de pedidos. Lembre-se de que cada pedido custa pelo menos 0,01 créditos.

Um exemplo com os preços que a documentação dá para `deepseek-ai/deepseek-v4-flash` (56 de entrada e 112 de saída): 2.000 pedidos com 1.500 tokens de entrada e 300 de saída dão 0,084 + 0,0336 = 0,1176 por pedido, arredondado para 0,12. No total, cerca de 240 créditos.

3. Ver o saldo

```
curl "https://api.interweb.co.mz/v1/ai/balance" -H "Authorization: Bearer $INTERWEB_API_KEY"
```

Compare o `credits` com o total. Cada pedido reserva o `max_tokens` inteiro antes de responder, por isso deixe alguma folga.

4. Testar e controlar o ritmo

1. Escreva e teste o programa com uma chave `sandbox_`: as respostas vêm no formato certo e não custam nada.

2. Passe para a chave `prod_` com uma amostra pequena, por exemplo 20 pedidos, e veja o custo real no Histórico.

3. Respeite o limite de pedidos por minuto da chave: num `429`, espere os segundos de `Retry-After`. Veja **os limites de pedidos**.

4. Se quiser um tecto, use os Limites de gasto por dia e por mês, na aba Comprar créditos.

Ponha sempre `max_tokens`.
Limita o tamanho de cada resposta e a reserva, e evita `402` quando o saldo está baixo. Nas imagens e nos vídeos, mande o cabeçalho `Idempotency-Key`: se repetir um pedido depois de uma falha de rede, não paga duas gerações.

Muitas tarefas minúsculas? Juntar várias numa só mensagem evita pagar o mínimo de 0,01 em cada uma. Veja **como se arredonda o custo**.

RECOMENDADO

Antes do lote, veja o saldo e o preço no All in One AI.

Ver o All in One AI →

VEJA TAMBÉM

- **Ver o saldo de créditos de IA: a página, o GET /v1/ai/balance e o aio balance** →
- **Ler o preço de um modelo: créditos por milhão de tokens, por imagem, por segundo ou por geração** →
- **Limites de pedidos da API: por chave, por conta e por tarefa, os cabeçalhos e o que fazer num 429** →
- **O erro 402 insufficient_credits: o que querem dizer required e balance, e como resolver** →

All in One AI

interweb

Interweb Moçambique

RLX Holdings, S.A. · Grupo RLX Holdings

NUIT	401 191 607
NUEL	105 045 974
CAE	62022
SEDE	Bairro 7 de Setembro Vilankulo 1304 · CEP 0411-02, Inhambane, Moçambique
CONTACTOS	info@interweb.co.mz · +258 84 635 1111 · +258 86 635 1111 · www.interweb.co.mz
HORÁRIO	2ª. à 6ª. Feira - 08 à 16 Horas Sábado & Domingo - Fechado

Versão online deste documento

RLX-MZ-KB-2980 · v2026.11.28 · Gerado a 28/11/2026 21:58