interweb

Gerado a 28/11/2026 22:01

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

O limite de pedidos por chave: 429 rate_limited e o cabeçalho Retry-After

Usar a API de IA

CÓDIGO
RLX-MZ-KB-3010

VERSÃO
v2026.11.28

PAÍS · IDIOMA
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/3010

O limite de pedidos por chave: 429 rate_limited e o cabeçalho Retry-After

Usar a API de IA · Criado em **28 de novembro de 2026** · 3 min de leitura



A resposta curta: cada chave tem um limite de pedidos por minuto. Quando passa dele, a API responde `429` com o código `rate_limited` e o cabeçalho `Retry-After`, que diz quantos segundos esperar. O seu programa só tem de esperar esse tempo e repetir.

Os limites que contam

Quem pede	Limite
Cada chave	Um número de pedidos por minuto, que vem no cabeçalho `X-RateLimit-Limit`.
A conta inteira, somando as chaves	300 pedidos por minuto (1.000 nas contas de revenda, que também têm 300 por chave).

São os valores de hoje e podem mudar. No seu programa, leia os cabeçalhos em vez de escrever os números no código.

Ver o limite e o que resta

Cada resposta traz três cabeçalhos:

Cabeçalho	O que é
`X-RateLimit-Limit`	Pedidos permitidos por minuto.
`X-RateLimit-Remaining`	Pedidos que restam neste minuto.
`X-RateLimit-Reset`	Quando acaba o minuto, em segundos Unix.

Para os ver, junte `-i` ao curl. Listar os modelos não gasta créditos:

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

O `GET /v1/limits` mostra o mesmo em JSON, com o limite e os pedidos que restam no minuto:

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

Quando chega ao limite

```
{
  "error": {
    "code": "rate_limited",
    "message": "Too many requests. Please try again shortly.",
    "correlation_id": "0f8fad5b-d9cb-469f-a165-70867728950e",
    "details": {"retry_after": 42}
  }
}
```

1. Espere pelo menos os segundos de `Retry-After` (também vêm em `details.retry_after`).
2. Se o `429` voltar, espere mais tempo em cada tentativa.
3. Evite ciclos muito rápidos, como consultar uma geração várias vezes por segundo.

Um exemplo em Python, que espera o tempo pedido e dobra a espera se o erro voltar:

```
import os, time, requests

URL = "https://api.interweb.co.mz/v1/chat/completions"
HEADERS = {"Authorization": "Bearer " + os.environ["INTERWEB_API_KEY"]}

def send(body, attempts=5):
    wait = 1
    for _ in range(attempts):
        r = requests.post(URL, headers=HEADERS, json=body, timeout=60)
        if r.status_code != 429 or r.json()["error"]["code"] != "rate_limited":
            return r
        wait = max(int(r.headers.get("Retry-After", wait)), wait)
        time.sleep(wait)
        wait *= 2
    return r
```

O código confirma `rate_limited` porque há outro `429`: o `plan_limit_reached`, quando os pontos do Coding Plan acabam. Esse não passa em segundos: veja **os planos e os créditos**.

O que gasta pedidos sem dar por isso

Cada consulta do estado de uma geração conta como um pedido: consulte de 5 em 5 segundos ou mais. Também contam as ferramentas do servidor MCP, os comandos do `aio` e os pedidos dos agentes de programação, que fazem muitos seguidos. Os SDK da OpenAI repetem sozinhos um `429`, duas vezes por omissão. Se tem vários programas, dê uma chave a cada um: o limite é por chave, mas o da conta soma todas.

Os limites de outras operações da API, como domínios e SMS, estão em **limites de pedidos da API**.

RECOMENDADO

Quer saber o que cada chave gastou? Veja o Histórico do All in One AI.

Ver o All in One AI →

VEJA TAMBÉM

- **Limites de pedidos da API: por chave, por conta e por tarefa, os cabeçalhos e o que fazer num 429** →
- **Erro 429 Too Many Requests: quem está a pedir demais e o que fazer** →
- **Criar muitas imagens com IA de uma vez: o saldo, o preço e o limite de pedidos** →
- **Erros da API de IA: 402, 403, 404, 422, 429 e 502, e o que fazer em cada caso** →
- **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-3010 · v2026.11.28 · Gerado a 28/11/2026 22:01