Mozout

Generated on 29/11/2026 04:04

KNOWLEDGE BASE · PUBLIC DOCUMENT

# Before a large batch of AI requests: checking the balance and the price first

AI Credits & Costs

| CODE | VERSION | COUNTRY · LANGUAGE |
| --- | --- | --- |
| RLX-INT-KB-9024 | v2026.11.28 | International · English |

**MozOut International**
RLX Holdings (Pty) Ltd. · Group RLX Holdings
Tax no. (SARS) 943 907 6234 · Registration no. (CIPC) 2025/667953/07

https://www.my.mozout.com/knowledgebase/9024

# Before a large batch of AI requests: checking the balance and the price first

AI Credits & Costs Created on **November 28, 2026** 2 min read



Sending hundreds or thousands of requests at once is where a mistake gets expensive. A few minutes of sums beforehand save credits and stop a batch from halting halfway for lack of balance.

## 1. Check the model price

On the **Models** tab or through the API, for example looking for a model with "flash" in its name:

```
curl "https://api.mozout.com/v1/models?type=text&q=flash" -H "Authorization: Bearer $MOZOUT_API_KEY"
```

## 2. Do the sum

Per request: (input tokens × input price + `max_tokens` × output price) / 1,000,000. To estimate the input, count about one token per 3 characters. Then multiply by the number of requests. Remember that each request costs at least 0.01 credits.

An example with the prices the documentation gives for `deepseek-ai/deepseek-v4-flash` (56 input and 112 output): 2,000 requests with 1,500 input tokens and 300 output tokens come to 0.084 + 0.0336 = 0.1176 per request, rounded up to 0.12. In total, about 240 credits.

## 3. Check the balance

```
curl "https://api.mozout.com/v1/ai/balance" -H "Authorization: Bearer $MOZOUT_API_KEY"
```

Compare `credits` with the total. Each request reserves the full `max_tokens` before answering, so leave some room.

## 4. Test and control the pace

1. Write and test the program with a `sandbox_` key: the answers come in the right format and cost nothing.

2. Switch to the `prod_` key with a small sample, for example 20 requests, and check the real cost in History.

3. Respect the key's requests-per-minute limit: on a `429`, wait the seconds in `Retry-After`. See **the request limits**.

4. If you want a ceiling, use the daily and monthly Spending limits on the Buy credits tab.

**Always set `max_tokens`.**
It limits the size of each answer and the reservation, and avoids a `402` when the balance is low. For images and videos, send the `Idempotency-Key` header: if you repeat a request after a network failure, you do not pay for two generations.

Lots of tiny tasks? Putting several in one message avoids paying the 0.01 minimum on each. See **how costs are rounded up**.

RECOMMENDED

Before the batch, check your balance and the price in All in One AI.

See All in One AI →

SEE ALSO

- **Checking your AI credit balance: the page, GET /v1/ai/balance and aio balance**
- **Reading a model's price: credits per million tokens, per image, per second or per generation**
- **API request limits: per key, per account and per task, the headers and what to do after a 429**
- **The 402 insufficient_credits error: what required and balance mean, and how to fix it**
- **All in One AI**

Mozout

## MozOut International

**RLX Holdings (Pty) Ltd. · Group RLX Holdings**


| | |
|---|---|
| TAX NO. (SARS) | 943 907 6234 |
| REGISTRATION NO. (CIPC) | 2025/667953/07 |
| REGISTERED OFFICE | 1111 Olive Street<br>Saint Louis 63101, Missouri, United States |
| CONTACT | info@mozout.com · +27 78 823 3325 · www.mozout.com |
| OPENING HOURS | Monday to Friday - 08 AM to 04 PM<br>Saturday & Sunday - Closed |


Online version of this document

RLX-INT-KB-9024 · v2026.11.28 · Generated on 29/11/2026 04:04