interweb

Gerado a 28/11/2026 20:36

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

# Entrada em cache: porque é que o texto que o modelo já leu gasta menos créditos

Créditos de IA e custos

| CÓDIGO | VERSÃO | PAÍS · IDIOMA |
| --- | --- | --- |
| RLX-AO-KB-2972 | v2026.11.28 | Angola · Português |

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

https://www.meu.interweb.ao/knowledgebase/2972

# Entrada em cache: porque é que o texto que o modelo já leu gasta menos créditos

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



Os modelos de texto têm três preços: entrada, saída e entrada em cache. A entrada em cache é texto de entrada que o modelo já tinha lido e que agora é lido da cache. Esses tokens pagam-se a um preço mais baixo, o `cached_input`, em vez do preço normal de entrada.

## Onde está o preço da cache

A aba Modelos mostra a entrada e a saída. O preço da cache vem em `GET /v1/models`, no campo `pricing.cached_input`, em créditos por milhão de tokens. Dois exemplos da documentação:

| Modelo | Entrada | Entrada em cache |
|---|---|---|
| `deepseek-ai/deepseek-v4-flash` | 56 | 11,2 |
| `moonshotai/kimi-k2.6` | 380 | 64 |

Os valores são exemplos e podem mudar: os actuais estão sempre em `GET /v1/models`.

## Quanto se poupa

Com os preços do exemplo, 8.000 tokens de entrada no `deepseek-ai/deepseek-v4-flash` custam 8.000 × 56 / 1.000.000 = 0,448 créditos ao preço normal. Se vierem da cache, custam 8.000 × 11,2 / 1.000.000 = 0,0896.

## Onde vê a cache usada

| Onde | O campo |
|---|---|
| **Histórico do All in One AI** | «Tokens em cache», no detalhe de cada pedido. |
| **Chat Completions (SDK da OpenAI)** | `usage.prompt_tokens_details.cached_tokens` |

| Onde | O campo |
| --- | --- |
| **Formato Messages (Claude Code)** | `usage.cache_read_input_tokens` |
| **Formato Responses (Codex)** | `usage.input_tokens_details.cached_tokens` |

## Quando aparece

A cache não se pede num campo do pedido: quando há, vem na contagem da resposta. Costuma aparecer quando vários pedidos começam pelo mesmo texto: as mesmas instruções, o mesmo documento ou uma conversa que vai crescendo. Por isso ajuda manter esse começo igual de pedido para pedido. Com um Coding Plan, os tokens de cache também contam pontos, ao valor do modelo.

**Nem todos os pedidos têm cache, e não a podemos garantir.**
Faça as contas com o preço normal de entrada e veja a poupança real no Histórico.

RECOMENDADO

Quer ver os tokens em cache de cada pedido? Abra o Histórico.

Ver o All in One AI →

VEJA TAMBÉM

- **Quanto custa uma conversa pela API: a reserva, o max_tokens e o acerto no fim**
- **Ler o preço de um modelo: créditos por milhão de tokens, por imagem, por segundo ou por geração**
- **A aba Histórico: quanto gastou cada pedido de IA, com filtros e exportação para CSV**
- **Tokens explicados: como se contam as APIs de modelos de IA**
- **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<br>Belas, Luanda, Angola |
| **CONTACTOS** | info@interweb.ao · +244 942 011 191 · www.interweb.ao |
| **HORÁRIO** | 2ª. à 6ª. Feira - 08 à 16 Horas<br>Sábado & Domingo - Fechado |


Versão online deste documento

RLX-AO-KB-2972 · v2026.11.28 · Gerado a 28/11/2026 20:36