interweb

Generated on 28/11/2026 21:58

KNOWLEDGE BASE · PUBLIC DOCUMENT

Cached input: why text the model has already read costs fewer credits

AI Credits & Costs

CODE	VERSION	COUNTRY · LANGUAGE
RLX-MZ-KB-2972	v2026.11.28	Mozambique · English

Interweb Mozambique
RLX Holdings, S.A. · Group RLX Holdings
Tax ID (NUIT) 401 191 607 · Legal entity no. (NUEL) 105 045 974 · Economic activity code (CAE) 62022

https://www.meu.interweb.co.mz/knowledgebase/2972

Cached input: why text the model has already read costs fewer credits

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



Text models have three prices: input, output and cached input. Cached input is input text the model had already read and is now read from the cache. Those tokens are charged at a lower price, `cached_input`, instead of the normal input price.

Where the cache price is

The Models tab shows input and output. The cache price comes in `GET /v1/models`, in the `pricing.cached_input` field, in credits per million tokens. Two examples from the documentation:

Model	Input	Cached input
`deepseek-ai/deepseek-v4-flash`	56	11.2
`moonshotai/kimi-k2.6`	380	64

These values are examples and can change: the current ones are always in `GET /v1/models`.

How much you save

With the example prices, 8,000 input tokens on `deepseek-ai/deepseek-v4-flash` cost 8,000 × 56 / 1,000,000 = 0.448 credits at the normal price. If they come from the cache, they cost 8,000 × 11.2 / 1,000,000 = 0.0896.

Where you see the cache used

Where	The field
All in One AI History	"Cached tokens", in the details of each request.
Chat Completions (OpenAI SDK)	`usage.prompt_tokens_details.cached_tokens`

Where	The field
Messages format (Claude Code)	`usage.cache_read_input_tokens`
Responses format (Codex)	`usage.input_tokens_details.cached_tokens`

When it shows up

You do not ask for the cache in a request field: when there is some, it comes in the response count. It usually shows up when several requests start with the same text: the same instructions, the same document or a conversation that keeps growing. So it helps to keep that beginning the same from one request to the next. With a Coding Plan, cache tokens also count points, at the model's value.

Not every request gets the cache, and we cannot guarantee it.
Do your sums with the normal input price and check the real saving in History.

RECOMMENDED

Want to see the cached tokens of each request? Open History.

See All in One AI →

SEE ALSO

- What a chat request costs through the API: the reservation, max_tokens and the final adjustment →
- Reading a model's price: credits per million tokens, per image, per second or per generation →
- The History tab: what each AI request spent, with filters and CSV export →
- Tokens explained: how AI model APIs are counted →
- All in One AI →

interweb

Interweb Mozambique

RLX Holdings, S.A. · Group RLX Holdings

TAX ID (NUIT)	401 191 607
LEGAL ENTITY NO. (NUEL)	105 045 974
ECONOMIC ACTIVITY CODE (CAE)	62022
REGISTERED OFFICE	Bairro 7 de Setembro Vilankulo 1304 · CEP 0411-02, Inhambane, Mozambique
CONTACT	info@interweb.co.mz · +258 84 635 1111 · +258 86 635 1111 · www.interweb.co.mz
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-MZ-KB-2972 · v2026.11.28 · Generated on 28/11/2026 21:58