Mozout

Generated on 27/11/2026 08:06

KNOWLEDGE BASE · PUBLIC DOCUMENT

Choosing a text model: price, context window and what each one is for

All in One AI

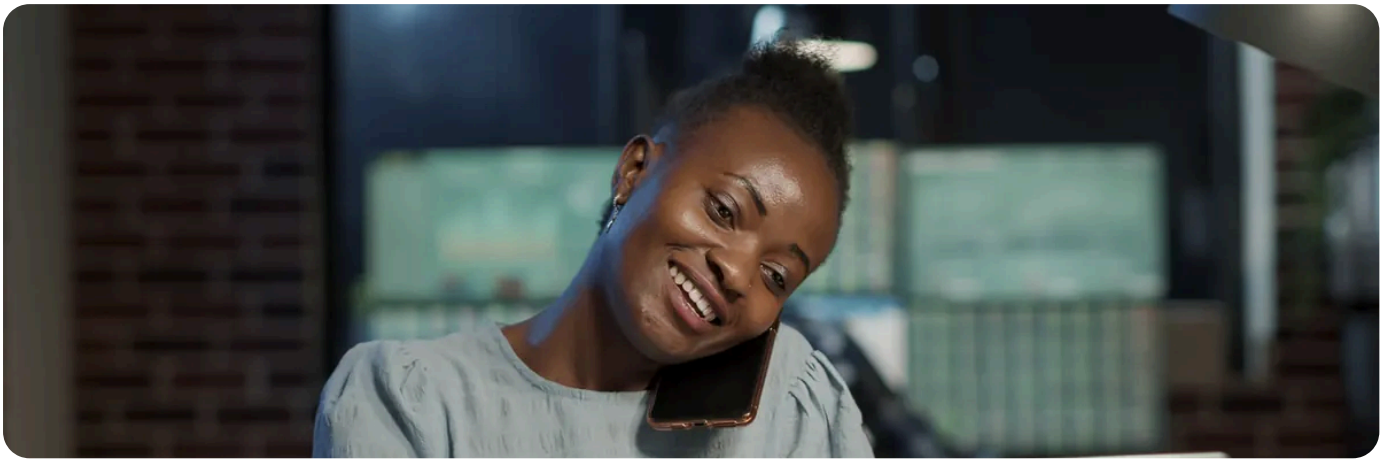
CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-8712	v2026.11.26	International · English

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/8712

Choosing a text model: price, context window and what each one is for

All in One AI Created on **November 26, 2026** 2 min read



In short: open the **Models** tab and choose **Text**. Each model shows its context window and its price in credits per million input and output tokens. Choose according to the job: a fast, cheap model for simple tasks, a stronger one for reasoning and difficult writing.

What to look at for each model

Field	What it means
Id	The name you send in `model` in the API. It always has a slash, like `deepseek-ai/deepseek-v4-flash`. Copy it with **Copy the id**.
Context	How many tokens the model reads at once: your instructions, the conversation and any text you give it. A large document needs a large context.
Input price	Credits per million tokens you send.
Output price	Credits per million tokens the model writes. It is usually higher than the input price.

If you are not sure what a token is, see **tokens explained**. Text in Portuguese often uses more than one token per word.

Which model for what

No model is best at everything. The documentation gives a few starting points: `deepseek-ai/deepseek-v4-flash` is fast and cheap, good for short questions and background tasks. `deepseek-ai/deepseek-v4-pro` reasons better, for hard problems. `moonshotai/kimi-k2.7-code` is built for coding. The full list is on the Models tab and in the **documentation**.

1. Start with the cheapest model that does the job.
2. Try the same prompt on two or three models in the Playground, using **Try it**, and compare the replies and the real cost.
3. Keep the one that answers well for the lowest cost. Only move to a pricier model if the difference shows.

Through the API

The same list comes from `GET /v1/models?type=text`. Search with `q`, for example `q=kimi`. Each model includes `context_length`, `max_output_tokens` and its price in `pricing`, with `input`, `output` and `cached_input`. Input read from the cache costs less.

Prices change over time. Before a big job, check the model's price on the Models tab or with `GET /v1/models`.

RECOMMENDED

Compare the text models and their prices on the Models tab.

See the models →

SEE ALSO

- **Chatting with a text model in the Playground before you use the API** →
- **What a chat request costs through the API: the reservation, max_tokens and the final adjustment** →
- **Tokens explained: how AI model APIs are counted** →
- **Why the AI forgets what you said earlier: context and long conversations** →
- **Chat and text in 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 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 Saturday & Sunday - Closed

Online version of this document

RLX-INT-KB-8712 · v2026.11.26 · Generated on 27/11/2026 08:06