interweb

Gerado a 27/11/2026 09:55

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Escolher um modelo de texto: preço, contexto e para que serve cada um

All in One AI

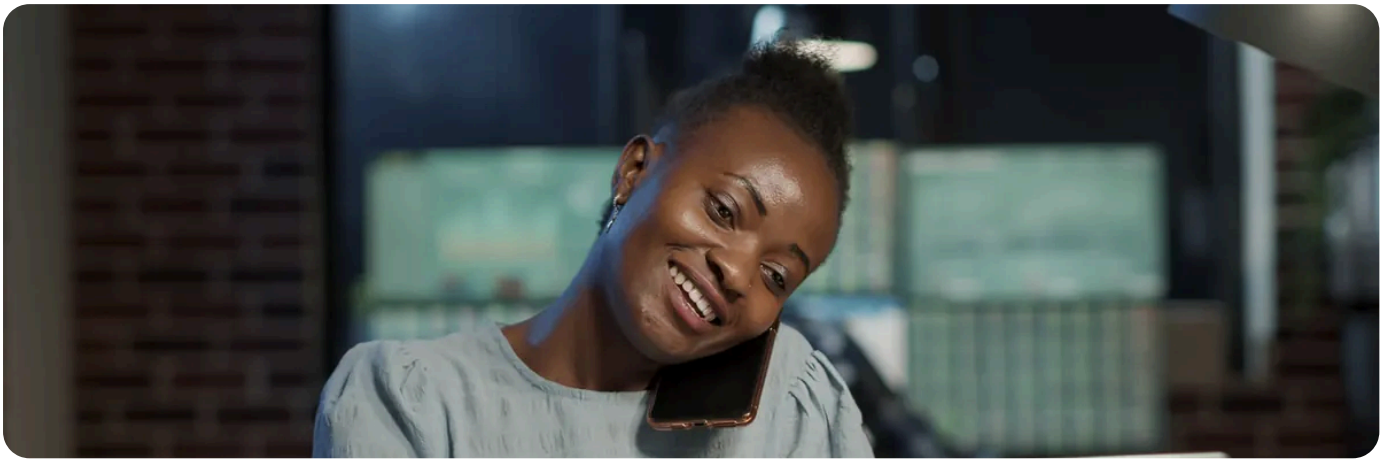
CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-MZ-KB-2876	v2026.11.26	Moçambique · Português

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/2876

Escolher um modelo de texto: preço, contexto e para que serve cada um

All in One AI Criado em **26 de novembro de 2026** 2 min de leitura



A resposta curta: abra o separador **Modelos** e escolha **Texto**. Cada modelo mostra o contexto e o preço em créditos por 1 milhão de tokens de entrada e de saída. Escolha conforme o trabalho: um modelo rápido e barato para tarefas simples, um mais forte para raciocínio e textos difíceis.

O que olhar em cada modelo

Campo	O que quer dizer
Id	O nome a enviar em `model` na API. Tem sempre uma barra, como `deepseek-ai/deepseek-v4-flash`. Copie-o com **Copiar o id**.
Contexto	Quantos tokens o modelo lê de uma vez: as instruções, a conversa e o texto que lhe dá. Um documento grande precisa de um contexto grande.
Preço de entrada	Créditos por 1 milhão de tokens que envia.
Preço de saída	Créditos por 1 milhão de tokens que o modelo escreve. Costuma ser mais caro do que a entrada.

Se não sabe o que é um token, veja **os tokens explicados**. Em português, uma palavra gasta muitas vezes mais do que um token.

Que modelo para quê

Não há um modelo melhor para tudo. Na documentação damos alguns pontos de partida: o `deepseek-ai/deepseek-v4-flash` é rápido e barato, bom para perguntas curtas e tarefas de fundo. O `deepseek-ai/deepseek-v4-pro` raciocina melhor, para problemas difíceis. O `moonshotai/kimi-k2.7-code` foi feito para programar. A lista completa está no separador Modelos e na **documentação**.

1. Comece pelo modelo mais barato que faça o trabalho.

2. Teste o mesmo pedido em dois ou três modelos no Playground, com **Experimentar**, e compare a resposta e o custo real.

3. Fique com o que responde bem pelo menor custo. Só suba para um modelo mais caro se a diferença se notar.

Pela API

A mesma lista sai em `GET /v1/models?type=text`. Procure com `q`, por exemplo `q=kimi`. Cada modelo traz `context_length`, `max_output_tokens` e o preço em `pricing`, com `input`, `output` e `cached_input`. A entrada lida da cache sai mais barata.

Os preços mudam com o tempo. Antes de um trabalho grande, confirme o preço do modelo no separador Modelos ou em `GET /v1/models`.

RECOMENDADO

Compare os modelos de texto e os preços no separador Modelos.

Ver os modelos →

VEJA TAMBÉM

- **Conversar com um modelo de texto no Playground, antes de usar a API**
- **Quanto custa uma conversa pela API: a reserva, o max_tokens e o acerto no fim**
- **Tokens explicados: como se contam as APIs de modelos de IA**
- **Porque é que a IA se esquece do que lhe disse: contexto e conversas longas**
- **Chat e texto no 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-2876 · v2026.11.26 · Gerado a 27/11/2026 09:55