interweb

Gerado a 29/11/2026 11:26

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Mudar de modelo no gateway de IA sem mexer no código: o id do modelo numa definição

Playground e modelos de texto

CÓDIGO
RLX-MZ-KB-3282

VERSÃO
v2026.11.29

PAÍS · IDIOMA
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/3282

Mudar de modelo no gateway de IA sem mexer no código: o id do modelo numa definição

Playground e modelos de texto · Criado em **29 de novembro de 2026** · 2 min de leitura



Num gateway de IA, o modelo é só um campo do pedido. Se guardar o id do modelo numa definição, e não no código, muda de modelo sem tocar numa linha do programa: altera a definição e reinicia.

O modelo numa variável de ambiente

```
import os
from openai import OpenAI

client                    =                    OpenAI(base_url="api.interweb.co.mz/v1",
api_key=os.environ["INTERWEB_API_KEY"])
modelo = os.environ.get("AI_MODEL", "deepseek-ai/deepseek-v4-flash")

r = client.chat.completions.create(
    model=modelo,
    messages=[{"role": "user", "content": "Resuma este texto em duas frases: ..."}],
    max_tokens=300,
)
print(r.choices[0].message.content)
```

Para passar ao `deepseek-ai/deepseek-v4-pro`, por exemplo, basta mudar `AI_MODEL`. O pedido, o formato da resposta e a chave ficam iguais.

Antes de mudar, confirme três coisas

O quê	Onde ver
O preço	Na aba Modelos ou em `GET /v1/models`, no campo `pricing`. Um modelo mais forte pode custar várias vezes mais.

O quê	Onde ver
O contexto e a saída máxima	`context_length` e `max_output_tokens`, no mesmo sítio. Um texto longo pode não caber num modelo com contexto menor. Veja **o que quer dizer cada um**.
As ferramentas	Se o seu programa envia `tools`, confirme que o modelo novo as aceita, com um pedido pequeno antes de mudar em produção.

Um modelo de reserva, se o principal falhar

Com o modelo numa definição, também pode ter um segundo para quando o primeiro responde `502 upstream_error`. Se o primeiro falhar com 502, o programa passa ao segundo:

```
import openai

modelos = [os.environ.get("AI_MODEL", "deepseek-ai/deepseek-v4-flash"),
           os.environ.get("AI_MODEL_RESERVA", "zai-org/glm-5.2")]
for m in modelos:
    try:
        r = client.chat.completions.create(model=m, messages=mensagens, max_tokens=300)
        break
    except openai.APIStatusError as e:
        if e.status_code != 502:
            raise
```

Veja **quando tentar outra vez e quando escolher outro modelo**.

Antes de escolher o modelo principal, compare: **o mesmo pedido em vários modelos**.

RECOMENDADO

Quer escolher o próximo modelo? Abra a aba Modelos.

Ver o All in One AI →

VEJA TAMBÉM

- **Porque é que cada id de modelo tem uma barra** →
- **Contexto e saída máxima** →

O erro 502 upstream_error →

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-3282 · v2026.11.29 · Gerado a 29/11/2026 11:26