Mozout

Generated on 29/11/2026 11:27

KNOWLEDGE BASE · PUBLIC DOCUMENT

Changing models in the AI gateway without touching your code: keep the model id in a setting

Playground & Text Models

CODE	VERSION	COUNTRY · LANGUAGE
RLX-MZ-KB-3281	v2026.11.29	Mozambique · English

MozOut 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.mozout.co.mz/knowledgebase/3281

Changing models in the AI gateway without touching your code: keep the model id in a setting

Playground & Text Models Created on **November 29, 2026** 2 min read



In an AI gateway, the model is just one field of the request. If you keep the model id in a setting, not in the code, you change models without touching a line of the program: you change the setting and restart.

The model in an environment variable

```
import os
from openai import OpenAI

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

r = client.chat.completions.create(
    model=model,
      messages=[{"role": "user", "content": "Summarise this text in two sentences:
..."}],
    max_tokens=300,
)
print(r.choices[0].message.content)
```

To move to `deepseek-ai/deepseek-v4-pro`, for example, you only change `AI_MODEL`. The request, the response format and the key stay the same.

Before you switch, check three things

What	Where to see it
The price	On the Models tab or in `GET /v1/models`, in the `pricing` field. A stronger model can cost several times more.

What	Where to see it
The context and maximum output	`context_length` and `max_output_tokens`, in the same place. A long text may not fit in a model with a smaller context. See **what each one means**.
Tools	If your program sends `tools`, check that the new model accepts them, with a small request before you switch in production.

A backup model, if the main one fails

With the model in a setting, you can also keep a second one for when the first answers `502 upstream_error`. If the first fails with 502, the program moves on to the second:

```
import openai

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

See **when to try again and when to choose another model**.

Before you pick the main model, compare: **the same request on several models**.

RECOMMENDED

Want to pick your next model? Open the Models tab.

See All in One AI →

SEE ALSO

- **Why every model id has a slash** →
- **Context window and maximum output** →

The 502 upstream_error

All in One AI

MozOut 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@mozout.co.mz · +258 84 635 1111 · +258 86 635 1111 · www.mozout.co.mz
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-MZ-KB-3281 · v2026.11.29 · Generated on 29/11/2026 11:27