Mozout

Generated on 29/11/2026 09:29

KNOWLEDGE BASE · PUBLIC DOCUMENT

Trying the same request on several models in the AI gateway before you choose one

Playground & Text Models

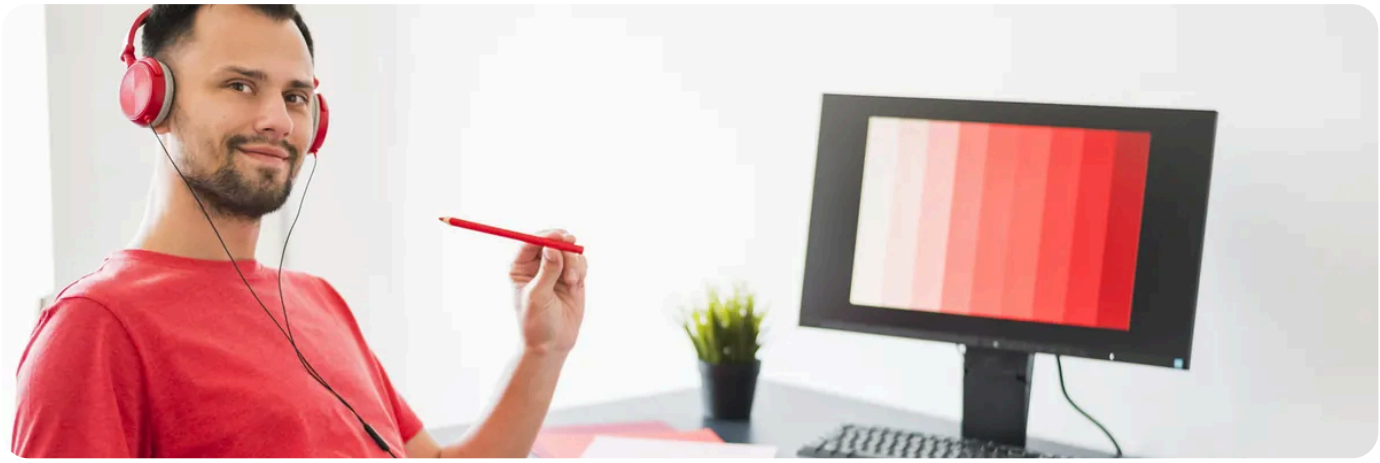
CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-9936	v2026.11.29	International · English

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/9936

Trying the same request on several models in the AI gateway before you choose one

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



One of the advantages of an AI gateway is putting several models to the test with the same request, the same key and the same code. Before you choose the model for your program, it is worth seeing how three or four answer the same text and what each one spends.

In the Playground, without writing code

1. Open **All in One AI** and go to the Playground.
2. Type the request your program will make, with a real text.
3. Pick a model, see what it will cost before you send it, and send it. See **how to see the cost before sending**.
4. Switch model and send the same request. Compare the answers and the cost of each one.

Through the API, with a loop

```
import os
from openai import OpenAI

client = OpenAI(base_url="api.mozout.com/v1", api_key=os.environ["MOZOUT_API_KEY"])
request = [{"role": "user", "content": "Write a short reply to a customer asking about our Saturday opening hours."}]

for model in ["deepseek-ai/deepseek-v4-flash", "deepseek-ai/deepseek-v4-pro", "zai-org/glm-5.2"]:
```

```
    r = client.chat.completions.create(model=model, messages=request, max_tokens=300)
    print("==", model, r.usage.prompt_tokens, "+", r.usage.completion_tokens, "tokens")
    print(r.choices[0].message.content)
```

The cost of each request is on the History tab, line by line, with the model used. See **the History tab**.

The Sandbox is no good for comparing.
With a `sandbox_` key no model is called: they all answer "Resposta de teste (sandbox).". To compare real answers you need the `prod_` key and credits. Use short texts and a low `max_tokens` to spend little.

What to compare

What	Why
The quality of the answer	It is what your customer will read. Use real texts, not made-up examples.
Tokens and cost	Two models can give similar answers at very different prices.
Response time	Measure it in your program: it depends on the model and the length of the answer.
Length	Some models write more than you need, and you pay for the output too.

RECOMMENDED

Want to compare models? Open the All in One AI Playground.

See All in One AI →

SEE ALSO

- **Chatting in the Playground** →
- **Cheap or strong text model** →
- **Changing models without touching your code** →
- **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-9936 · v2026.11.29 · Generated on 29/11/2026 09:29