Mozout

Generated on 27/11/2026 08:09

KNOWLEDGE BASE · PUBLIC DOCUMENT

Using your AI agent through the API: agent/ in the model field

All in One AI

CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-8754	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/8754

Using your AI agent through the API: agent/ in the model field

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



You call an agent through the chat endpoint you already know, `POST /v1/chat/completions`, with the same key and the `ai:write` permission. In `model`, send `agent/` followed by the agent id. Do not send the instructions or the knowledge: the agent adds them itself, before the messages.

Where to find the id

In the Agents tab of **All in One AI**, under Use through the API. It also shows up in the model list: `GET /v1/models?type=agent` returns your agents with the id `agent/<id>`, the model they use in `base_model`, their tools in `tools` and their state in `status`. Each account only sees its own agents.

An example in Python

With the OpenAI SDK only the address and the key change, as in **using the OpenAI SDK**. The key lives in the `MOZOUT_API_KEY` environment variable:

```
import os
from openai import OpenAI

client = OpenAI(base_url="https://api.mozout.com/v1", api_key=os.environ["MOZOUT_API_KEY"])

completion = client.chat.completions.create(
model="agent/7c1e9a42-b35d8f06-a2c4e719-5d0b3f8e",
messages=[{"role": "user", "content": "What are your opening hours on Saturday?"}],
max_tokens=400,
)
print(completion.choices[0].message.content)
```

The answer has the chat format: the text is in `choices[0].message.content` and the token count in `usage`. It also works with `stream: true`.

What the request can change

The model is always the agent's. The request's `max_tokens` and `temperature` can lower the agent's values but never raise them: a higher value is treated as the agent's.

When the answer is an error

Answer	When	What to do
`409 agent_paused`	The agent is paused.	Turn it back on in the Agents tab.
`404 model_not_found`	The id is wrong, or the agent belongs to another account.	Check the id with `GET /v1/models?type=agent`.
`403 insufficient_scope`	The key does not have `ai:write`.	Create a key with the right permission.
`402 insufficient_credits`	There are not enough credits for the reservation. Nothing was charged.	Buy credits or lower `max_tokens`.

The key stays on your server.
To put the agent in a chat on your website, your server calls the API, never the browser or a mobile app. The ways to set that up are in **adding an AI chat assistant to your website**. If the key leaks, see **what to do, in the right order**.

Write your program with a `sandbox_` key.
The agent answers in the same format without calling the model and without charging anything. To see real answers, use Try it or a `prod_` key. More in **testing in the Sandbox at no cost**.

RECOMMENDED

Want an agent working inside your program? Visit the AI agents page.

See AI agents →

SEE ALSO

Creating an AI agent in All in One AI, step by step →

- What an AI agent costs, and how to test it without spending
- Using the OpenAI SDK with All in One AI: only the address and the key change
- Testing All in One AI in the Sandbox at no cost: what each answer looks like
- AI agents in All in One AI

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-8754 · v2026.11.26 · Generated on 27/11/2026 08:09