interweb

Generated on 27/11/2026 09:56

KNOWLEDGE BASE · PUBLIC DOCUMENT

What an AI agent costs, and how to test it without spending

All in One AI

CODE	VERSION	COUNTRY · LANGUAGE
RLX-MZ-KB-2892	v2026.11.26	Mozambique · English

Interweb 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.interweb.co.mz/knowledgebase/2892

What an AI agent costs, and how to test it without spending

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



An agent has no price of its own. You pay for the model it uses, in AI credits, at the price shown in `GET /v1/models`. Each round trip to the model spends credits like a chat request.

What goes into each answer

Part	How it counts
Instructions and knowledge	Input tokens, on every request
The conversation messages	Input tokens
The answer	Output tokens
Each use of a tool	One more call to the model, which also carries the text of the ones before

So an answer that uses two tools makes three calls to the model and costs more than one without tools. To understand tokens, see **tokens explained**.

Reservation and settlement

1. Before calling the model, we reserve credits for the input and for the `max_tokens` of output. Without `max_tokens`, the reservation counts 4096 output tokens.

2. If your balance does not cover the reservation, the answer is `402 insufficient_credits` and nothing is charged.

3. At the end, we charge the real tokens and return the rest of the reservation to your balance. The minimum per request is 0.01 credits.

Testing for free, and testing for real

Way	Spends credits?	What it is for
A `sandbox_` key	No	Writing and testing your program. The agent answers in the same format but does not call the model or the tools: the text is always "Resposta de teste (sandbox)."
The Try it button	Yes	Seeing the agent's real answers, with the cost of each one. The requests show up in the history as the All in One AI Playground.

How to keep the bill down

1. Keep the instructions and the knowledge short. They go with every answer. See **what to add to the knowledge**.

2. Set `max_tokens`. It caps the answer length and lowers the reservation, which avoids a `402` when your balance is low.

3. Turn on only the tools the agent needs.

4. For simple questions, choose a cheaper model. Each model's price is in `GET /v1/models`, in the `pricing` field.

Before you open the agent to many people,
check your balance with `GET /v1/ai/balance` or in **All in One AI**, and ask half a dozen questions in Try it to see what a normal conversation costs.

RECOMMENDED

Want an agent whose costs you control? Visit the AI agents page.

See AI agents →

SEE ALSO

- Creating an AI agent in All in One AI, step by step
- Your AI agent's knowledge: what to add and how to keep each answer cheap
- AI agent tools: reading web pages and knowing today's date
- Tokens explained: how AI model APIs are counted
- AI agents in All in One AI

interweb

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

Online version of this document

RLX-MZ-KB-2892 · v2026.11.26 · Generated on 27/11/2026 09:56