interweb

Generated on 28/11/2026 22:08

KNOWLEDGE BASE · PUBLIC DOCUMENT

Agent limits: 5 tool rounds, and max_tokens and temperature that can only go down

Agents Through the API

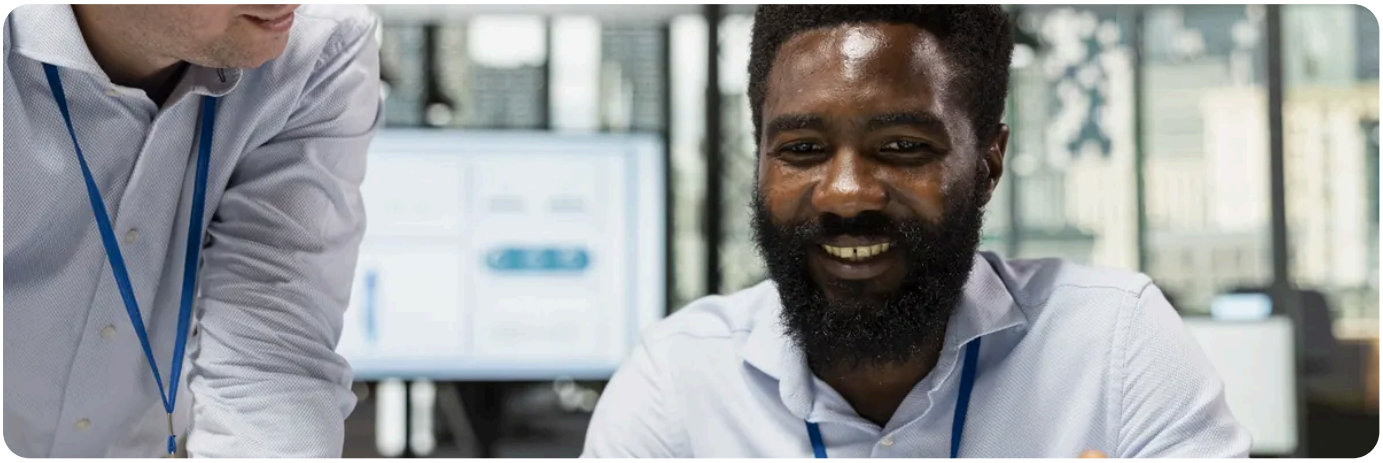
CODE	VERSION	COUNTRY · LANGUAGE
RLX-MZ-KB-3056	v2026.11.28	Mozambique · English

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/3056

Agent limits: 5 tool rounds, and max_tokens and temperature that can only go down

Agents Through the API Created on **November 28, 2026** 2 min read



An agent brings its own settings to every request. Your program can lower some of them, but cannot go above them. These are the limits, and what to do when one gets in the way.

The limits of each request

Limit	Value
Tool rounds per request	At most 5
Web searches per answer	At most 3, at 0.5 credits each
A page read with `web_fetch`	Up to 20,000 characters of text; it waits at most 10 seconds and reads up to 1 MB
Model	Always the agent's
`max_tokens` and `temperature`	The request can lower the agent's values, never raise them

The limits of the agent

What	Limit
Agents per account	50
Instructions	20,000 characters
Knowledge	200 KB per agent; each file up to 10 MB

What	Limit
Temperature	From 0 to 2
Maximum tokens in the reply	From 16 to 32,768

max_tokens and temperature only go down

Picture an agent with a maximum of 1024 tokens and a temperature of 0.7, the values the Agents tab gives a new agent:

The request sends	The agent uses
`"max_tokens": 300`	300
`"max_tokens": 4000`	1024, the agent's
`"temperature": 0.2`	0.2
`"temperature": 1.5`	0.7, the agent's
Neither	The agent's

For longer answers, raise the maximum in the agent's Settings, in **All in One AI**. A higher maximum also raises the credit reservation of each request. The reservation is explained in **what a chat request costs**.

The 5 rounds

Each tool use is a round to the model, and in each request the agent makes at most 5. For questions that would need many pages, split the work into smaller requests. Each round also adds to the cost and the response time: see **an agent's answer with streaming**.

On top of this, each key has a limit of requests per minute, as in the rest of the API. It is covered in **API request limits**.

RECOMMENDED

Want to know how far your agent can go? See All in One AI.

See All in One AI →

SEE ALSO

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

What a chat request costs through the API: the reservation, max_tokens and the final adjustment

API request limits: per key, per account and per task, the headers and what to do after a 429

An agent's answer with streaming: why the first piece can take longer

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-3056 · v2026.11.28 · Generated on 28/11/2026 22:08