interweb

Generated on 28/11/2026 20:37

KNOWLEDGE BASE · PUBLIC DOCUMENT

temperature and max_tokens: what they change in a text model's answer

Playground & Text Models

CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-2996	v2026.11.28	Angola · English

Interweb Angola
RLX Angola, Lda. · Group RLX Holdings
Tax ID (NIF) 500 260 5724 · Commercial registration no. 26464-25/250724

https://www.meu.interweb.ao/knowledgebase/2996

temperature and max_tokens: what they change in a text model's answer

Playground & Text Models Created on **November 28, 2026** 2 min read



These are the two fields that change an answer the most. `max_tokens` limits the length of the answer and the credit reservation. `temperature` changes how much the answer varies: lower is more predictable; higher, more varied.

max_tokens

It is the most tokens the answer may have. Without it, we reserve 4,096. A lower value reserves fewer credits. It never goes above the model's `max_output_tokens`.

If the answer reaches the limit, it is cut off. The API tells you like this:

Format	Sign of a cut-off answer
Chat Completions	`finish_reason` : `length`
Messages (Claude Code)	`stop_reason` : `max_tokens`
Responses (Codex)	`status` : `incomplete`

The Playground shows "(The reply was cut at the token limit. Raise the maximum to see all of it.)".

temperature

It goes from 0 to 2. Lower gives more predictable, precise answers; higher, more creative ones. For summaries, extracting data or sorting, use a low value. For ideas, names or slogans, a higher value gives more variety.

Where you change them

Where	max_tokens	temperature
API	`max_tokens`	`temperature`

Where	max_tokens	temperature
Playground	"Maximum tokens in the reply"	Not available.
`aio chat` **command**	`--max-tokens`	Not available.
Agents	In the agent settings. A request can lower it, never raise it.	In the agent settings. A request can lower it, never raise it.

An example through the API

Test first with a `sandbox_` key: it answers in the same format and spends no credits. On Linux and macOS:

```
curl "https://api.interweb.ao/v1/chat/completions" \
-H "Authorization: Bearer $INTERWEB_API_KEY" \
-H "Content-Type: application/json" \
-d '{"model": "deepseek-ai/deepseek-v4-flash", "messages": [{"role": "user", "content": "Write a slogan for a bakery"}], "max_tokens": 200, "temperature": 0.3}'
```

A very low `max_tokens` cuts the answer; a very high one raises the reservation
and can cause a `402` with a low balance. Choose a value that fits what you expect to get back. See **how each request is charged**.

RECOMMENDED

Try the maximum tokens setting in the All in One AI Playground.

See All in One AI →

SEE ALSO

- **What a chat request costs through the API: the reservation, max_tokens and the final adjustment**
- **Context window and maximum output: what context_length and max_output_tokens mean**
- **The 402 insufficient_credits error: what required and balance mean, and how to fix it**
- **Creating an AI agent in All in One AI, step by step**

All in One AI

interweb

Interweb Angola

RLX Angola, Lda. · Group RLX Holdings

TAX ID (NIF)	500 260 5724
COMMERCIAL REGISTRATION NO.	26464-25/250724
REGISTERED OFFICE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACT	info@interweb.ao · +244 942 011 191 · www.interweb.ao
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-AO-KB-2996 · v2026.11.28 · Generated on 28/11/2026 20:37