interweb

Gerado a 28/11/2026 20:38

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

temperature e max_tokens: o que mudam na resposta de um modelo de texto

Playground e modelos de texto

CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-AO-KB-2996	v2026.11.28	Angola · Português

Interweb Angola
RLX Angola, Lda. · Grupo RLX Holdings
NIF 500 260 5724 · Matrícula 26464-25/250724

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

temperature e max_tokens: o que mudam na resposta de um modelo de texto

Playground e modelos de texto Criado em **28 de novembro de 2026** 2 min de leitura



São os dois campos que mais mexem numa resposta. O `max_tokens` limita o tamanho da resposta e a reserva de créditos. A `temperature` muda o quanto a resposta varia: mais baixa, mais previsível; mais alta, mais variada.

max_tokens

É o máximo de tokens que a resposta pode ter. Sem ele, reservamos 4.096. Um valor mais baixo reserva menos créditos. Nunca passa do `max_output_tokens` do modelo.

Se a resposta chegar ao limite, fica cortada. A API avisa assim:

Formato	Sinal de resposta cortada
Chat Completions	`finish_reason` : `length`
Messages (Claude Code)	`stop_reason` : `max_tokens`
Responses (Codex)	`status` : `incomplete`

No Playground aparece «(A resposta foi cortada no limite de tokens. Aumente o máximo para a ver inteira.)».

temperature

Vai de 0 a 2. Mais baixa, respostas mais previsíveis e certas; mais alta, mais criativas. Para resumos, extrair dados ou classificar, use um valor baixo. Para ideias, nomes ou slogans, um valor mais alto dá mais variedade.

Onde se mudam

Onde	max_tokens	temperature
API	`max_tokens`	`temperature`

Onde	max_tokens	temperature
Playground	«Máximo de tokens na resposta»	Não tem.
Comando `aio chat`	`--max-tokens`	Não tem.
Agentes	Nas definições do agente. O pedido pode baixar, nunca subir.	Nas definições do agente. O pedido pode baixar, nunca subir.

Um exemplo pela API

Teste primeiro com uma chave `sandbox_`: responde no mesmo formato e não gasta créditos. No Linux e no 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}'
```

Um `max_tokens` muito baixo corta a resposta; muito alto aumenta a reserva
e pode dar `402` com pouco saldo. Escolha um valor à medida do que espera receber. Veja **como se cobra cada pedido**.

RECOMENDADO

Experimente o máximo de tokens no Playground do All in One AI.

Ver o All in One AI →

VEJA TAMBÉM

- **Quanto custa uma conversa pela API: a reserva, o max_tokens e o acerto no fim**
- **Contexto e resposta máxima: o que querem dizer context_length e max_output_tokens**
- **O erro 402 insufficient_credits: o que querem dizer required e balance, e como resolver**
- **Criar um agente de IA no All in One AI, passo a passo**

All in One AI

interweb

Interweb Angola

RLX Angola, Lda. · Grupo RLX Holdings

NIF	500 260 5724
MATRÍCULA	26464-25/250724
SEDE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACTOS	info@interweb.ao · +244 942 011 191 · www.interweb.ao
HORÁRIO	2ª. à 6ª. Feira - 08 à 16 Horas Sábado & Domingo - Fechado

RLX-AO-KB-2996 · v2026.11.28 · Gerado a 28/11/2026 20:38

Versão online deste documento