interweb

Gerado a 28/11/2026 20:37

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Contexto e resposta máxima: o que querem dizer context_length e max_output_tokens

Playground e modelos de texto

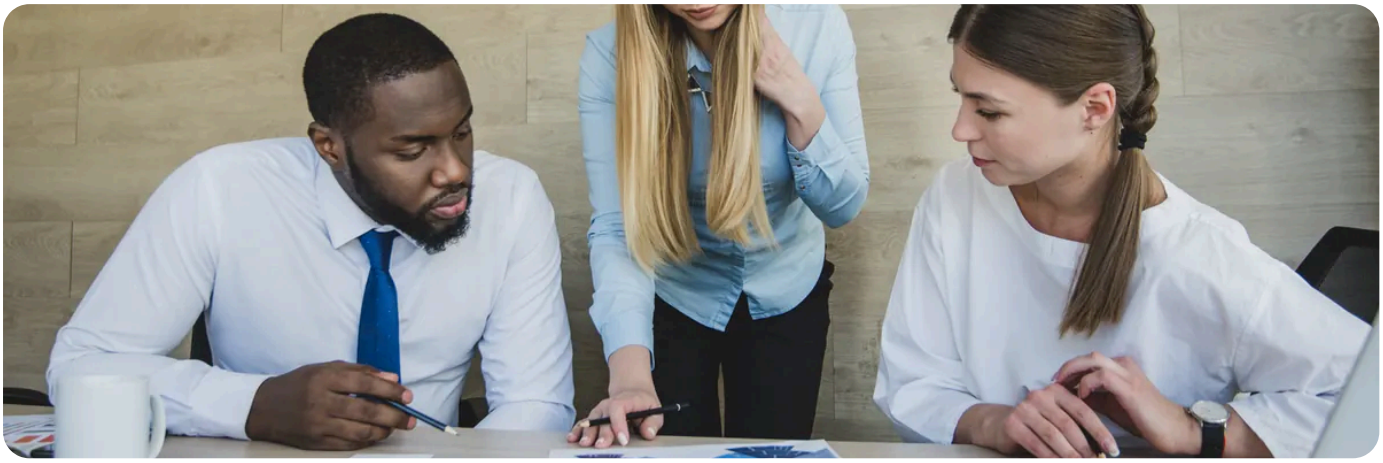
CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-AO-KB-2986	v2026.11.28	Angola · Português

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

https://www.meu.interweb.ao/knowledgebase/2986

Contexto e resposta máxima: o que querem dizer context_length e max_output_tokens

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



Cada modelo de texto tem dois limites: quanto lê de uma vez, o `context_length`, e quanto escreve numa resposta, o `max_output_tokens`. Os dois contam-se em tokens.

Os dois números

Campo	O que é
`context_length`	O máximo de tokens que o modelo lê num pedido, com a entrada e a saída juntas.
`max_output_tokens`	O máximo de tokens que o modelo escreve numa resposta.

Dois exemplos da documentação: `deepseek-ai/deepseek-v4-flash` lê até 1.048.576 tokens e escreve até 393.216; `moonshotai/kimi-k2.6` lê até 262.144 e escreve até 262.144.

Onde os vê

A aba **Modelos** mostra o contexto de cada modelo de texto («Contexto: N tokens»). Os dois campos vêm em `GET /v1/models` e no comando `aio model`, seguido do id.

O que conta para o contexto

Tudo o que vai no pedido: as instruções, a conversa inteira, os textos que colou e a própria resposta. Numa conversa, cada mensagem nova leva as anteriores. Nos agentes, as instruções e o conhecimento vão em todos os pedidos. Por isso uma conversa muito longa acaba por bater no limite. No Playground aparece «A conversa ficou muito longa. Comece uma nova.» Veja **porque é que a IA se esquece do que lhe disse**.

O max_output_tokens e o seu max_tokens

O `max_output_tokens` é o tecto do modelo. O `max_tokens` é o limite que põe em cada pedido, abaixo desse tecto. A reserva de créditos conta o seu `max_tokens`, ou 4.096 se não o mandar. No Playground, o campo «Máximo de tokens na resposta» aceita de 16 a 32.768, e a estimativa nunca conta mais do que o máximo de saída do modelo.

Nas ferramentas de programação

Quando uma ferramenta pede o tamanho do contexto, use o `context_length`. Para a saída máxima, use um valor menor. Reservamos créditos pelo máximo de saída que a ferramenta pede em cada resposta. Por isso os exemplos pedem 16 000 tokens: chega para programar e precisa de menos saldo do que um valor maior.

RECOMENDADO

Veja o contexto de cada modelo na aba Modelos.

Ver o All in One AI →

VEJA TAMBÉM

- **Porque é que a IA se esquece do que lhe disse: contexto e conversas longas**
- **Escolher um modelo de texto: preço, contexto e para que serve cada um**
- **A aba Modelos: procurar um modelo e copiar o id**
- **Tokens explicados: como se contam as APIs de modelos de IA**
- **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

Versão online deste documento

RLX-AO-KB-2986 · v2026.11.28 · Gerado a 28/11/2026 20:37