interweb

Gerado a 28/11/2026 20:44

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Contexto e resposta no Claude Code: CLAUDE_CODE_MAX_CONTEXT_TO e CLAUDE_CODE_MAX_OUTPUT_TOK

Claude Code

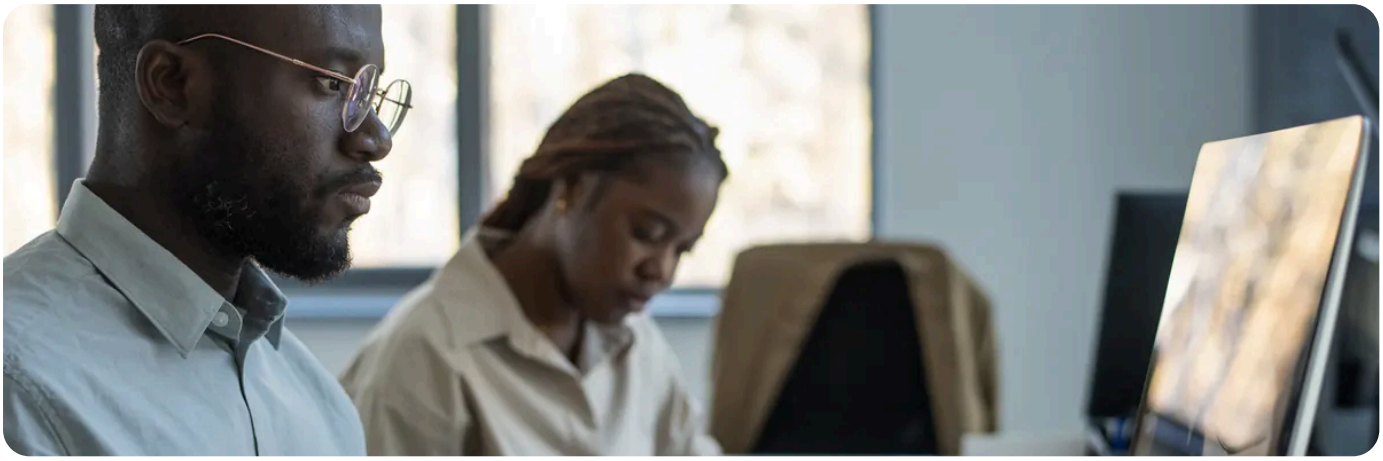
CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-AO-KB-3116	v2026.11.28	Angola · Português

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

https://www.meu.interweb.ao/knowledgebase/3116

Contexto e resposta no Claude Code: CLAUDE_CODE_MAX_CONTEXT_TOKENS e CLAUDE_CODE_MAX_OUTPUT_TOKENS

Claude Code Criado em **28 de novembro de 2026** 2 min de leitura



Estas duas variáveis dizem ao Claude Code o tamanho do modelo: quanto texto cabe numa conversa e quanto pode ter cada resposta. O Claude Code não conhece os ids do catálogo, por isso precisa que lhe diga.

CLAUDE_CODE_MAX_CONTEXT_TOKENS: quando resumir a conversa

O Claude Code manda a conversa inteira em cada pedido e resume-a quando ela chega ao limite do contexto. Com um id que não conhece, avisa que não reconhece o modelo e assume um tamanho. Nas versões recentes, assume 1 milhão de tokens. Com o `moonshotai/kimi-k2.7-code`, que aceita 262 144, a conversa passava do limite do modelo antes de ser resumida, e o pedido falhava.

Ponha na variável o `context_length` do modelo, que está na **aba «Modelos»** e em `GET /v1/models`. Segundo a documentação do Claude Code, para um id que ele não reconhece e que não começa por `claude-`, como `moonshotai/kimi-k2.7-code`, a variável vale sozinha.

CLAUDE_CODE_MAX_OUTPUT_TOKENS: o máximo de cada resposta

Sem esta variável, o Claude Code pede até 32 000 tokens por resposta a um modelo que não conhece. Antes de cada pedido, reservamos créditos pela entrada e pelo máximo de saída, e devolvemos o que sobrar no fim. Com 16 000, cada pedido precisa de menos saldo livre. Se aparecer o erro `402 insufficient_credits`, baixe este valor ou compre créditos.

Não ponha mais do que o `max_output_tokens` do modelo, em `GET /v1/models`. E não desça demasiado: uma resposta que chega ao máximo fica cortada.

Valores para os modelos recomendados

Modelo	CLAUDE_CODE_MAX_CONTEXT_TOKENS	CLAUDE_CODE_MAX_OUTPUT_TOKENS
`moonshotai/kimi-k2.7-code`	262144	16000

Modelo	CLAUDE_CODE_MAX_CONTEXT_TOKENS	CLAUDE_CODE_MAX_OUTPUT_TOKENS
deepseek-ai/deepseek-v4-pro	1048576	16000
zai-org/glm-5.2	1048576	16000
qwen/qwen3.8-max	1000000	16000
deepseek-ai/deepseek-v4-flash	1048576	16000
zai-org/glm-5.3-flash	1048576	16000

Pôr as duas variáveis

```
export CLAUDE_CODE_MAX_CONTEXT_TOKENS=262144
export CLAUDE_CODE_MAX_OUTPUT_TOKENS=16000
claude
```

No `settings.json`, vão na chave `env`, entre aspas: `"CLAUDE_CODE_MAX_CONTEXT_TOKENS": "262144"`. O Claude Code lê estas variáveis ao arrancar: depois de as mudar, feche-o e abra-o outra vez.

O contexto é de cada modelo.
Se trocar com `/model` para um modelo com menos contexto, o valor antigo fica. Use o contexto do modelo mais pequeno que costuma usar, ou feche e abra o Claude Code com o valor certo.

A reserva e o acerto no fim estão explicados em **quanto custa uma conversa pela API**.

RECOMENDADO

Quer programar com IA no Claude Code? Veja a página Programar com IA.

Ver Programar com IA →

VEJA TAMBÉM

- Usar o Claude Code com o All in One AI: as variáveis, o modelo e os erros comuns
- Mudar de modelo a meio de uma sessão do Claude Code com /model
- Quanto custa uma conversa pela API: a reserva, o max_tokens e o acerto no fim
- Escolher um modelo de texto: preço, contexto e para que serve cada um
- Programar com IA no 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-3116 · v2026.11.28 · Gerado a 28/11/2026 20:44