interweb

Generated on 28/11/2026 20:44

KNOWLEDGE BASE · PUBLIC DOCUMENT

Context and output in Claude Code: CLAUDE_CODE_MAX_CONTEXT_TO and CLAUDE_CODE_MAX_OUTPUT_TOK

Claude Code

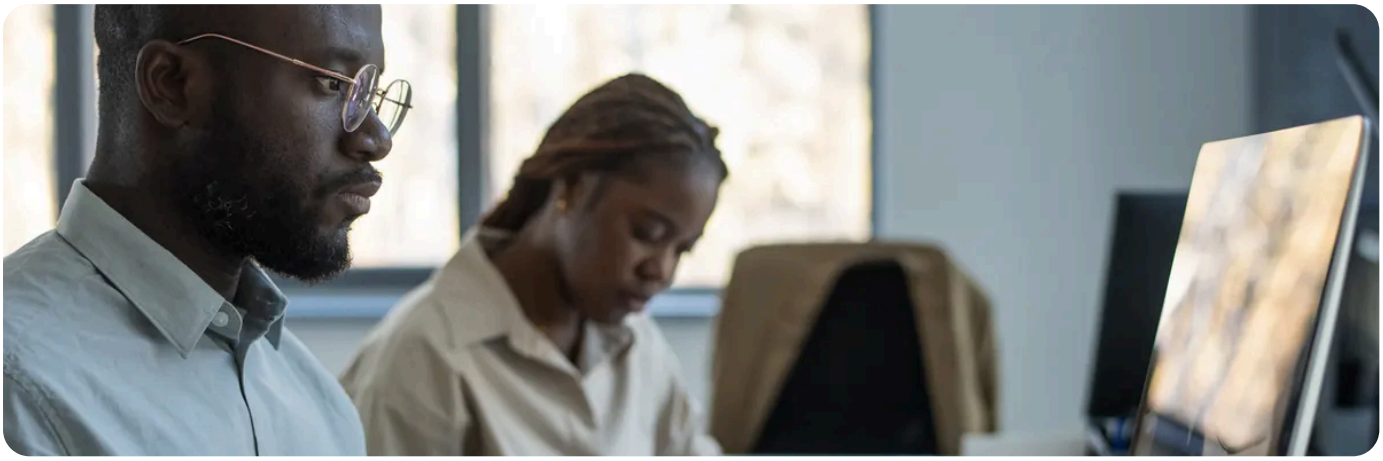
CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-3116	v2026.11.28	Angola · English

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/3116

Context and output in Claude Code: CLAUDE_CODE_MAX_CONTEXT_TOKENS and CLAUDE_CODE_MAX_OUTPUT_TOKENS

Claude Code Created on **November 28, 2026** 2 min read



These two variables tell Claude Code the size of the model: how much text fits in a conversation and how long each answer can be. Claude Code does not know the catalogue ids, so it needs you to tell it.

CLAUDE_CODE_MAX_CONTEXT_TOKENS: when to summarise the conversation

Claude Code sends the whole conversation with every request and summarises it when it reaches the context limit. With an id it does not know, it warns that it does not recognise the model and assumes a size. Recent versions assume 1 million tokens. With `moonshotai/kimi-k2.7-code`, which takes 262,144, the conversation would go past the model limit before being summarised, and the request would fail.

Set the variable to the model's `context_length`, shown in the **Models tab** and in `GET /v1/models`. According to the Claude Code documentation, for an id it does not recognise that does not start with `claude-`, such as `moonshotai/kimi-k2.7-code`, the variable works on its own.

CLAUDE_CODE_MAX_OUTPUT_TOKENS: the maximum per answer

Without this variable, Claude Code asks for up to 32,000 tokens per answer from a model it does not know. Before each request, we reserve credits for the input and for the maximum output, and return what is left at the end. With 16,000, each request needs less free balance. If you get the `402 insufficient_credits` error, lower this value or buy credits.

Do not go above the model's `max_output_tokens`, in `GET /v1/models`. And do not go too low: an answer that hits the maximum is cut short.

Values for the recommended models

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

Model	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

Setting both variables

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

In `settings.json`, they go under the `env` key, in quotes: `"CLAUDE_CODE_MAX_CONTEXT_TOKENS": "262144"`. Claude Code reads these variables at start: after changing them, close it and open it again.

The context belongs to each model.
If you switch with `/model` to a model with less context, the old value stays. Use the context of the smallest model you usually work with, or close and reopen Claude Code with the right value.

The reservation and the final adjustment are explained in **what a chat request costs through the API**.

RECOMMENDED
Want to code with AI in Claude Code? Visit the Coding with AI page.
See Coding with AI →

SEE ALSO

- Using Claude Code with All in One AI: the variables, the model and common errors
- Changing the model during a Claude Code session with /model
- What a chat request costs through the API: the reservation, max_tokens and the final adjustment
- Choosing a text model: price, context window and what each one is for
- Coding with AI in 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-3116 · v2026.11.28 · Generated on 28/11/2026 20:44