interweb

Generated on 28/11/2026 20:46

KNOWLEDGE BASE · PUBLIC DOCUMENT

Model Configuration in Cline: the context window and the maximum output

Cline & Roo Code

CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-3164	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/3164

Model Configuration in Cline: the context window and the maximum output

Cline & Roo Code Created on **November 28, 2026** 2 min read



With the **OpenAI Compatible** provider, Cline knows nothing about the model you typed in **Model ID**. The **Model Configuration** section tells it how much text the model can take and how much it can write in each answer. With the right values, you avoid errors and spending too much.

The fields

According to **Cline's documentation**, Model Configuration has the context window, the maximum output, image support, Computer Use and the input and output prices. The ones that matter here:

Field	What to enter
Context Window	The model's context, the `context_length` in the model list. For `moonshotai/kimi-k2.7-code`, `262144`.
Max Output Tokens	The maximum for each answer. We recommend `16000`.
Input Price and Output Price	They belong to Cline: Interweb does not receive them and they do not change what you pay. What each request spent is in History.
The other fields	Leave them as they are.

Where to find the context

Model	Context	What for
`moonshotai/kimi-k2.7-code`	262,144 tokens	Built for coding: the starting point for everyday work.
`deepseek-ai/deepseek-v4-pro`	1,048,576 tokens	Strong reasoning, for big changes and hard bugs.

Model	Context	What for
`zai-org/glm-5.2`	1,048,576 tokens	Made for agents: handles tools well in long tasks.
`deepseek-ai/deepseek-v4-flash`	1,048,576 tokens	Fast and cheap, for background tasks and short questions.

The ids are in the Models tab of **All in One AI** or in the API list. On Linux and macOS:

```
curl "https://api.interweb.ao/v1/models?type=text" -H "Authorization: Bearer $INTERWEB_API_KEY"
```

In Windows PowerShell:

```
curl.exe "https://api.interweb.ao/v1/models?type=text" -H "Authorization: Bearer $env:INTERWEB_API_KEY"
```

Each model comes with its `id` (always with the slash), its `context_length` and its price in `pricing`. Listing the models spends no credits.

Why 16,000 for the maximum output

We reserve credits for the maximum output the tool asks for in each answer. That is why the examples ask for 16,000 tokens: it is enough for coding and needs a smaller balance than a higher value. Do not go above the model's `max_output_tokens`, which comes in the list. And do not go too low: the answer stops at the limit, and a big change to a file gets cut off halfway.

Why the right context matters

Context Window tells Cline how far the task's conversation can grow. With a value above the model's, the conversation can go past what the model accepts and requests can fail. Use the exact `context_length` value.

When you change the **Model ID**, change the **Context Window** too: each model has its own. See **how to choose a model for Cline**.

RECOMMENDED

Want our models in Cline? Visit the Coding with AI page.

See Coding with AI →

SEE ALSO

- **Cline with All in One AI: the OpenAI Compatible provider, field by field**
- **Lowering the maximum output in Cline to avoid 402 insufficient_credits**
- **Choosing a model for Cline: models that accept tools**
- **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-3164 · v2026.11.28 · Generated on 28/11/2026 20:46