interweb

Generated on 28/11/2026 20:47

KNOWLEDGE BASE · PUBLIC DOCUMENT

Lowering the maximum output in Cline to avoid 402 insufficient_credits

Cline & Roo Code

CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-3176	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/3176

Lowering the maximum output in Cline to avoid 402 insufficient_credits

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



`402 insufficient_credits` shows up when your balance does not cover the request's reservation, even if the real answer would cost little. In Cline, what weighs most in the reservation is **Max Output Tokens**. Lowering it often fixes the error without buying credits.

How the reservation works

1. **The reservation.** Before calling the model, we count the input (about one token for every 3 characters) plus the maximum output asked for. With no maximum, the reservation counts 4096 output tokens.
2. **Your balance is checked.** If it does not cover the reservation, the answer is `402` and nothing is charged.
3. **The adjustment.** At the end, we charge the real tokens and the rest of the reservation goes back to your balance.

What to change in Cline

1. Open the Cline settings and the **Model Configuration** section.
2. In **Max Output Tokens**, enter `16000`. With a low balance, a smaller value also works for small tasks.
3. Send the request again.

Do not go too low: the answer stops at the limit, and a big change to a file gets cut off halfway.

The input counts too

In a long task, each request carries more text, because it carries the conversation so far, and the reservation grows with it. If the 402 shows up in the middle of a long task, start a new task: the next request carries less text.

What the error tells you

The API answer includes `details.required`, what the reservation needed, and `details.balance`, what you have. If the gap is large, buy credits on the **Buy credits tab**. You can also see the balance in `GET /v1/ai/balance`.

With the Coding Plan, requests to the plan's models spend points instead of credits. See **which Cline requests spend points**.

RECOMMENDED

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

See Coding with AI →

SEE ALSO

- **What a chat request costs through the API: the reservation, max_tokens and the final adjustment**
- **Model Configuration in Cline: the context window and the maximum output**
- **Cline with All in One AI: the OpenAI Compatible provider, field by field**
- **Errors when setting up a tool with All in One AI: 401, model not found, no credits, /v1 and rejected config files**
- **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

RLX-AO-KB-3176 · v2026.11.28 · Generated on 28/11/2026 20:47

Online version of this document